UNIVERZITA KARLOVA

Přírodovědecká fakulta

Katedra aplikované geoinformatiky a kartografie

Studijní program: Geografie (navazující magisterské studium)

Studijní obor: Kartografie a geoinformatika

Bc. Barbora Chytilová

OPTIMALIZACE TVORBY TRÉNOVACÍHO A VALIDAČNÍHO DATASETU PRO ZVÝŠENÍ PŘESNOSTI KLASIFIKACE V DÁLKOVÉM PRŮZKUMU ZEMĚ

TRAINING AND VALIDATION DATASET OPTIMALIZATION FOR EARTH OBSERVATION CLASSIFICATION ACCURACY IMPROVEMENT

Diplomová práce

Vedoucí diplomové práce: RNDr. Lucie Kupková, Ph.D.

Konzultant diplomové práce: RNDr. Jakub Lysák, Ph.D.

Praha 2019

Vysoká škola: Univerzita Karlova **Fakulta:** Přírodovědecká

Katedra: Aplikované geoinformatiky a kartografie **Akademický rok:** 2018/2019

Zadání diplomové práce

pro Bc. Barboru Chytilovou

obor Kartografie a geoinformatika

Název tématu: Optimalizace tvorby trénovacího a validačního datasetu pro zvýšení přesnosti klasifikace v DPZ

Zásady pro vypracování

Optimalizace trénovacího a validačního datasetu pro řízenou klasifikaci dat v DPZ může přispět ke zvýšení přesnosti klasifikace a kvalitnímu vyhodnocení její přesnosti. V literatuře je popsáno mnoho přístupů, jak navrhnout trénovací/validační dataset a způsobů dělení terénních dat na trénovací a validační část. Přístupy jsou různé i v závislosti na zvoleném algoritmu klasifikace.

V rámci řešení práce budou v modelovém území (lesně-luční krajina v Podkrkonoší) prováděny pro dva vybrané klasifikační algoritmy experimenty s trénovacími a validačními daty. Bude měněn podíl/množství trénovacích a validačních dat s cílem vyhodnotit vliv těchto parametrů na přesnost klasifikace multispektrálních dat senzoru Sentinel-2A a navrhnout optimalizovaný trénovací a validační dataset, jehož parametry umožní zlepšit přesnost klasifikace a kvalitu hodnocení přesnosti.
Analýza může být doplněna vytvořením skriptu, který umožní automatické experimentování s nastavením trénovacích/validačních dat.

Dalším cílem práce bude porovnat dva zvolené klasifikační algoritmy a ověřit, zda je pro ně optimální stejné nebo rozdílné nastavení trénovacího/validačního datasetu.

Rozsah grafických prací: dle potřeby

Rozsah průvodní zprávy: cca 50 stran

Seznam odborné literatury:

Foody, G. M. (2004) Thematic map comparison: evaluating the statistical signifikance of differences in classification accuracy, Photogrammetric Engineering and Remote Sensing, 70, 627-633.

Foody, G. M. (2008) Harshness in image classification accuracy assessment, International Journal of Remote Sensing, 29, 3137-3158.

Foody, G. M. (2009) Sample size determination for image classification accuracy assessment and comparison, International Journal of Remote Sensing.

Foody, G. M. and Mathur, A. (2004) Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification, Remote Sensing of Environment, 93, 107

Foody, G.M., Mathur, A., 2006. The use of small training sets containing mixed pixels for accurate hard image classification: training on mixed spectral responses for classification by a SVM. Remote Sens. Environ. 103 (2), 179–189.

Radoux, J., Lamarche, C., Van Bogaert, E., Bontemps, S., Brockmann, C., Defourny, P., 2014. Automated training sample extraction for global land cover mapping. Remote Sens. 6 (5), 3965–3987.

Vedoucí diplomové práce: RNDr. Lucie Kupková, Ph.D.
Konzultant diplomové práce: RNDr. Jakub Lysák, Ph.D.

Datum zadání diplomové práce: 8.12.2016

Termín odevzdání diplomové práce: červenec 2019

…… .....……

Vedoucí diplomové práce Garant studijního oboru

V Praze dne 24.7.2019

Prohlášení

Prohlašuji, že jsem závěrečnou práci zpracovala samostatně a že jsem uvedla všechny použité informační zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

V Praze dne 24.7.2019 .....……

Bc. Barbora Chytilová

Poděkování

Ráda bych na tomto místě poděkovala RNDr. Lucii Kupkové, Ph.D. za odborné vedené mé diplomové práce, věnovaný čas, cenné rady a připomínky. Dále bych ráda poděkovala RNDr. Jakubu Lysákovi, Ph.D. za věnovaný čas a pomoc při tvorbě skriptů.

Také bych ráda poděkovala mému příteli a rodině za podporu během celého studia.

Abstrakt

Diplomová práce se zabývá optimalizací trénovacího a validačního datasetu pro řízenou klasifikaci dat v DPZ. V rámci řešení práce jsou v území lesně-luční krajiny v Podkrkonoší prováděny pro dva klasifikační algoritmy (Maximum Likelihood – MLC a Support Vector Machine – SVM) experimenty s trénovacími a validačními daty. Práce vychází z předpokladu, že pro dosažení maximální přesnosti klasifikace je ideální podíl 1/3 trénovacích a 2/3 validačních dat (Foody, 2009). Další hypotézou práce byl předpoklad, že v případě klasifikace pomocí algoritmu SVM je pro dosažení stejné/podobné přesnosti klasifikace potřeba nižší počet trénovacích bodů než v případě klasifikačního algoritmu Maximum Likelihood (Foody, 2004). Cílem práce bylo testovat vliv podílu/množství trénovacích a validačních dat na přesnost klasifikace multispektrálních dat senzoru Sentinel-2A s využitím algoritmu Maximum Likelihood. Nejvyšší celkové přesnosti při využití klasifikačního algoritmu Maximum Likelihood bylo dosaženo pro podíl 375 trénovacích a 625 validačních bodů. Celková přesnost pro tento podíl byla 72,88 %. Teorie Foodyho (2009), že pro dosažení nejvyšší přesnosti klasifikace je ideální podíl 1/3 trénovacích a 2/3 validačních dat potvrzují výsledky hodnocení celkové přesnosti a Kappa koeficientu pro Maximum Likelihood. Avšak výsledné uživatelské a zpracovatelské přesnosti pro jednotlivé třídy nedosáhly v případě tohoto podílu nejvyšších hodnot. Také se ukázalo, že změna velikosti validačního datasetu při zachování stabilní velikosti trénovacího datasetu má vliv na stabilitu výsledku hodnocení přesnosti klasifikace metodou Maximum Likelihood. Výsledek hodnocení celkové přesnosti pro algoritmus SVM pro tento podíl 375 trénovacích a 625 validačních bodů byl 79,09 %. V případě, že se počet trénovacích bodů snížil na 50 byla celková přesnost klasifikace 76,49 %. Předpoklad, že pomocí algoritmu SVM je pro dosažení stejné/podobné přesnosti klasifikace potřeba nižší počet trénovacích bodů než v případě klasifikačního algoritmu Maximum Likelihood, byl pro tento dataset potvrzen.

Klíčová slova: trénovací dataset, validační dataset, hodnocení přesnosti klasifikace, Maximum Likelihood, Support Vector Machine

Abstract

This thesis deals with training dataset and validation dataset for Earth observation classification accuracy improvement. Experiments with training data and validation data for two classification algorithms (Maximum Likelihood – MLC and Support Vector Machine – SVM) are carried out from the forest-meadow landscape located in the foothill of the Giant Mountains (Podkrkonoší). The thesis is base on the assumption that 1/3 of training data and 2/3 of validation data is an ideal ratio to achieve maximal classification accuracy (Foody, 2009). Another hypothesis was that in a case of SVM classification, a lower number of training point is required to achieve the same or similar accuracy of classification, as in the case of the MLC algorithm (Foody, 2004). The main goal of the thesis was to test the influence of proportion / amount of training and validation data on the classification accuracy of Sentinel – 2A multispectral data using the MLC algorithm. The highest overal accuracy using the MLC classification algorithm was achieved for 375 training and 625 validation points. The overal accuracy for this ratio was 72,88 %. The theory of Foody (2009) that 1/3 of training data and 2/3 of validation data is an ideal ratio to achieve the highest classification accuracy, was confirmed by the overal accuracy and Kappa coefficient results for MLC. It should be noted, that resulting Producer´s and User´s Accuracies for particular classes did not reach the highest values for this ratio. While size of the training dataset is sustained, further test showed that the change in the size of the validation dataset has an effect on the stability of MLC classsification accuracy assesment result. Result of overall accuracy evaluationin the case of SVM algorithm for the ratio 375 training points and 625 validation points was 79,09 %. In the case of 50 validation points the overal accuracy had reached 76,49 %. The assumption, that SVM algorithm needs lower number of training points to achieve similar classification accuracy as MLC algorithm was confirmed for this dataset.

Keywords: training dataset, validation dataset, classification accuracy improvement, Maximum Likelihood, Support Vector Machine

OBSAH

PŘEHLED POUŽITÝCH ZKRATEK

cLHS	conditional latin hypercube sampling
ČR	Česká republika
ČÚZK	Český úřad zeměměřický a katastrální
ESA	European Spase Agency
ISODATA	Iterative Self-Organizing Data Analysis Technique
LHS	latin hypercube sampling
MLC	Maximum Likelihood Classification
MSI	Multispectral imager
NDVI	Normalized Difference Vegetation Index
OBIA	Object-based Image Analysis
RGB	Red Green Blue color model
scLHS	stratifie conditional latin hypercube sampling
SGS	sekvenční Gaussova simulace
SNAP	Sentinel Application Platform
SVM	Support Vector Machine
UTM	Universal Transverse Mercator
VQT	variance quadtree technique
WGS-84	World Geodetic System 1984
WMS	Web Map Service

SEZNAM TABULEK, OBRÁZKŮ A GRAFŮ

ÚVOD

Využití snímků, které jsou získávány z družic na zemské orbitě nachází v posledních letech stále širší uplatnění. Produkce tematických map, které zobrazují land cover Země, pomocí klasifikace obrazu je jednou z nejběžnějších aplikací dálkového průzkumu Země. Tyto mapy odvozené z klasifikací mohou být značně nepřesné, protože pokryv Země je složen z různorodých heterogenních částí. Zavádí se proto pojem přesnost klasifikace, kdy je porovnáván klasifikovaný obraz s realitou. Zásadní roli při klasifikaci hrají trénovací data a při posuzování přesnosti klasifikace pak validační data.

Existuje mnoho přístupů pro hodnocení a vykazování přesnosti klasifikace. Přesnost klasifikace ovlivňuje mnoho faktorů a výsledky posouzení přesnosti mohou být také hodnoceny různými způsoby. Posouzení přesnosti se v zásadě provádí ve studiích dálkového průzkumu Země s cílem poskytnout objektivní základ pro hodnocení kvality tematické mapy reprezentované výstupem klasifikace (Foody, 2009). Práce vychází z předpokladu, že pro dosažení maximální přesnosti v případě klasifikačního algoritmu je ideální podíl 1/3 trénovacích a 2/3 validačních dat (Foody, 2009). Další hypotézou práce byl předpoklad, že v případě klasifikace pomocí algoritmu SVM je pro dosažení stejné/podobné přesnosti klasifikace potřeba nižší počet trénovacích bodů než v případě klasifikačního algoritmu Maximum Likelihood (Foody and Mathur, 2004).

Tato diplomová práce je zaměřena na optimalizaci trénovacích a validačních dat pro zvýšení přesnosti klasifikace. Literatura věnovaná strategiím výběru trénovacích datasetů se zaměřuje na tři aspekty, které mají největší vliv na následnou přesnost klasifikace. Hlavními aspekty jsou distribuce neboli podíl a rozdělení trénovacích/validačních dat do tříd, množství vstupních dat a jejich rozmístění.

Hlavním cílem této diplomové práce je testovat vliv podílu/množství trénovacích dat a validačních dat na přesnost klasifikace multispektrálních dat senzoru Sentinel-2A s využitím algoritmu Maximum Likelihood.

Dalším cílem práce je pro podíl trénovacích a validačních dat, který přinese nejlepší výsledek klasifikace v případě algoritmu Maximum Likelihood, měnit množství validačních dat (při zachování stabilní velikosti trénovacího datasetu) a sledovat vliv změny velikosti validačního datasetu na stabilitu výsledku hodnocení přesnosti klasifikace.

Práce chce dále ověřit, zda pro klasifikační algoritmy Maximum Likelihood a Support Vector Machine je optimální podíl trénovacích a validačních dat stejný a zda v případě metody Support Vector Machine stačí k dosažení určité přesnosti klasifikace méně trénovacích dat než v případě metody Maximum Likelihood. Přesnost klasifikace je hodnocena na základě chybové matice. Je hodnocena celková přesnost, přesnost zpracovatelská a uživatelská a kappa koeficient. Validační a trénovací data jsou vybírána náhodně pomocí náhodného stratifikovaného výběru tak, aby splňovala předem stanovená kritéria daného experimentu.

Automatizované experimenty s nastavením trénovacího/validačního datasetu jsou umožněny díky skriptu, který byl v práci navržen.

1 LITERÁRNÍ REŠERŠE A ÚVOD DO PROBLEMATIKY

1.1 Teoretický základ

Velká část zemské pevniny je pokryta vegetací. Vegetační složka krajiny je tak výraznou dominantou v dálkovém průzkumu zemského povrchu – kromě pouštních a polárních oblastí je vegetace obsažena ve všech datových souborech pořízených z letadlových a družicových nosičů (Kolář, 1990). Snímky, které jsou získávány z družic, nachází v posledních letech stále větší uplatnění. Dálkový průzkum Země se stává stále častějším nástrojem pro analýzu struktur zemského povrchu a hodnocení změn na něm probíhajících. S vývojem nových a dokonalejších zařízení pro snímání zemského povrchu se zvyšuje kvalita získaných snímků a s ní spojená přesnost algoritmů, které jsou určeny pro analýzu a vyhodnocení těchto dat.

Hlavním principem Dálkového průzkumu Země je měření elektromagnetického záření odraženého nebo vyzařovaného samotným zemským povrchem. Mezi základní parametry zařízení, které je neseno družicí patří prostorové a spektrální rozlišení.

Spektrální projev porostu určitého rostlinného druhu je výslednicí odrazivých a emisních vlastností různých částí rostliny i jejího pozadí. Dominující jsou však příspěvky od listů, a tak se nejčastěji využívá jejich odrazových vlastností k charakteristice spektrálního chování vegetačního porostu. Odrazové vlastnosti pokryvu jsou dle Dobrovolného (1998), formovány především následujícími faktory: vnější uspořádání vegetačního pokryvu, vnitřní struktura jednotlivých částí rostlin, vodní obsah, zdravotní stav a vlastnosti půdního substrátu. Charakterizovat spektrální odrazivost vegetace lze pomocí rozdělení do tří hlavních oblastí, které odpovídají třem faktorům určujících velikost spektrální odrazivosti. Jak uvádí Dobrovolný (1998), spektrální křivka odrazivosti vegetace se dělí do tří hlavních částí, které odpovídají faktorům určujícím velikost spektrální odrazivosti. Jedná se o oblast pigmentační absorpce (0,4 – 0,7 μm), oblast buněčné struktury (0,7 – 1,3 μm) a oblast vodní absorpce (1,3 – 3,0 μm).

Pro pásmo pigmentační absorpce je průběh spektrální křivky formován pigmentační látkou, kterou je u většiny rostlin chlorofyl (příčina zelené barvy rostlin ve vegetačním období). Pro toto pásmo jsou tak typické absorpční pásy, které vznikají, jelikož v těchto intervalech spektra je zeleným barvivem pohlcováno 70 % až 90 % dopadajícího záření (Campbell, 1996).

Oblast buněčné struktury lze nalézt v pásmu blízkého infračerveného záření, které je ovlivněno především morfologickou strukturou listu či jehlice. Základní stavební látku zde tvoří celulóza, která se vyznačuje nízkou pohltivostí záření. Odrazivost v oblasti buněčné struktury se často používá k charakterizování míry hustoty vegetačního krytu prostřednictvím tzv. indexu listové pokryvnosti (Leaf Area Index – LAI) jedná se o bezrozměrné číslo, které udává, kolikrát je plocha všech listů větší než jednotková plocha sloupce, ve kterém se listy nacházejí. Rozdílné odrazivosti vegetace ve viditelné červené (cca 0,6 až 0,7 μm) a blízké infračervené části spektra (kolem 0,8 μm) se využívá k výpočtu vegetačních indexů. Ty mohou být ukazatelem míry přítomnosti zelené hmoty nebo jejího zdravotního stavu.

V oblasti vodní absorpce je spektrální odrazivost formována absorpčními pásy vody – odrazivost v této části spektra je nepřímo úměrná obsahu vody v listu. Právě v těchto vlnových délkách se nejvíce projeví změny ve vodním obsahu vegetace, např. vodní stres rostlin. Odrazivost v této oblasti spektra závisí kromě vodního obsahu také například na tloušťce listu (Dobrovolný, 1998).

1.2 Klasifikační metody

Klasifikace je metoda, při které dochází k rozřazování jednotlivých obrazových prvků do informačních tříd. Základním předpokladem pro provedení klasifikace je použití klasifikátoru. Klasifikátorem je rozhodovací algoritmus či pravidlo, podle nějž jsou prvky zařazovány do určitých tříd (Kolář, Halounová, Pavelka, 1997). V současnosti existuje mnoho klasifikačních metod, které můžeme volit podle toho jakou analýzu jakých dat chceme provádět (Magiera et al., 2013, Sha et al., 2009, Dobrowski et al., 2008, Schmidtlein and Sassin, 2004). Jevy v obraze lze klasifikovat na základě spektrálních, prostorových či časových příznaků, které tvoří příznakový prostor. Rovnoměrnost tohoto prostoru se odvíjí od počtu spektrálních pásem obrazu.

Existují základní metody klasifikace, které spoléhají pouze na spektrální informace z obrazu bez trénovacích dat. Tyto metody jsou méně úspěšné, zejména při použití multispektrálních snímků anebo při klasifikaci horských a výrazně heterogenních oblastí (Magiera et al., 2013). Můžeme rozlišit několik druhů klasifikací, přičemž základní rozdělení odděluje klasifikaci pixelovou a objektově orientovanou.

Objektová klasifikace (OBIA) nezkoumá pixely, ale pracuje se shluky pixelů. Základním principem objektově orientované klasifikace tedy je, že informace získané při klasifikaci nejsou reprezentovány jednotlivými pixely, nýbrž objekty a jejich vzájemnými vztahy (Baatz, Schäpe, 2000). Na rozdíl od pixelové klasifikace bere v potaz vztah mezi jednotlivými pixely a věrohodněji tak vykresluje vegetační celky (Dobrovolný, 1998). Objektová klasifikace začíná vždy tzv. segmentací obrazu, kde pomocí předem definovaných parametrů jsou obrazová data rozdělena do určitých segmentů (objektů). Následně jsou nastavena klasifikační pravidla. Pro některé úlohy může být pixelová klasifikace založená pouze na spektrálních vlastnostech pixelu nedostačující a nepřesná. Proto se využívá i objektově-orientovaná klasifikace, která kromě spektrálních příznaků využívá také příznaky kontextuální, tvarové či geometrické (Yu et al., 2006). Podle Dobrowskiho et al. (2008) je možné tuto metodu použít, když se prostorové rozlišení obrazu stává jemnější a obrazová struktura se stává korelována s vegetačními strukturami (Dobrowski et al., 2008).

Klasifikátory založené na spektrálních vlastnostech jevů a objektů se nazývají per-pixel klasifikátory, jelikož k zařazení obrazových prvků do jednotlivých tříd nepoužívají vlastností a příznaků okolních pixelů, ale pouze pixelu klasifikovaného (Dobrovolný, 1998). Při pixelové klasifikaci je využíváno rozdílných vlastností pixelů, přičemž pixely s podobnými vlastnostmi jsou přiřazeny do jedné třídy příznakového prostoru na základě klasifikačního pravidla (Halounová, Pavelka, 2005). Obecně lze klasifikaci dělit podle toho, ve kterém okamžiku a jakým způsobem uživatel zasahuje do procesu klasifikace, na klasifikaci neřízenou a řízenou. Řízená klasifikace vyžaduje trénovací plochy, které představují známý povrch a identifikují se na základě terénního měření (Dobrovolný, 1998). Řízená klasifikace je náročnější na zpracování, ale výsledky bývají přesnější než u přístupu neřízené klasifikace.

1.3 Neřízená klasifikace

Neřízená klasifikace vychází z předpokladu, že pixely, které patří do jedné třídy, jsou ve vícerozměrném prostoru blízko sebe, a naopak pixely odlišných skupin, které představují povrchy lišící se spektrálním chováním, jsou dobře separované. Pomocí využití vícerozměrných statistických metod tzv. shlukových analýz se poté v multispektrálním příznakovém prostoru vymezí odlišné shluky. Výsledkem definování přibližného počtu výsledných shluků (maximální a minimální počet) nejsou ještě třídy informační, ale třídy spektrální. Těm je v interpretační fázi klasifikace dán určitý tematický obsah, jsou srovnány s jinými třídami, podpůrnými a referenčními daty, je jim přiřazena informační hodnota a stávají se tak třídami informačními. Běžně využívanými algoritmy neřízené klasifikace jsou např. algoritmus K-Means nebo ISODATA. Jako rozhodovací pravidlo je většinou použita modifikovaná metoda nejbližšího souseda, kdy vzdálenost mezi středy jednotlivých shluků je hodnocena různými měrami vzdálenosti (Dobrovolný, 1998).

1.4 Řízená klasifikace

Na počátku řízené klasifikace zájmového území je nutné sestavit klasifikační schéma, které je ovlivněno účelem klasifikace a informacemi, které chceme získat. Celý proces řízené klasifikace je založený na definování trénovacích ploch. Tyto trénovací plochy jsou položkami legendy výsledné tematické mapy. Pro jejich vymezení je nutné shromáždit validační data tedy informace o zpracovaném území z terénního průzkumu či jiných zdrojů.

Celý proces řízené klasifikace zahrnuje následující fáze (Dobrovolný, 1998):

1. Definování tzv. trénovacích ploch.
2. Výpočet statistických charakteristik (spektrálních příznaků) pro trénovací plochy charakterizující jednotlivé třídy, jejich editace a výběr vhodných pásem pro vlastní klasifikaci.
3. Volba vhodného klasifikátoru pro zařazení všech prvků obrazu do jednotlivých tříd.
4. Zatřídění všech obrazových prvků do vymezených tříd.
5. Úprava, hodnocení a prezentace výsledků klasifikace.

Klasifikace obrazu nemusí být závěrečným stádiem zpracování. Může sloužit například pouze k vymezení jedné určité třídy povrchů, která bude dále analyzována (Dobrovolný, 1998).

Definování trénovacích ploch je nejdůležitější fází, protože rozhoduje o úspěšnosti klasifikace. Trénovací data musí být především kompletní a reprezentativní to znamená, že musí být charakterizovány všechny třídy, které mají být klasifikovány (Dobrovolný, 1998). Jejich výběr závisí zcela na zpracovateli a kvalita výběru značně ovlivňuje přesnost výsledků klasifikace, proto by neměly zahrnovat několik spektrálních tříd.

Trénovací plocha vytvořena z vybraných pixelů tvoří masku, kterou lze v druhé fázi řízené klasifikace využít ke kontrole správnosti vytvořených trénovacích ploch pomocí grafických a statistických charakteristik. Tyto charakteristiky definují tzv. spektrální příznaky pro každou hledanou třídu a jsou charakterizovány například směrodatnou odchylkou, kovariační maticí nebo průměrovým vektorem. Díky těmto charakteristikám lze posoudit, zda vybrané trénovací plochy vhodně charakterizují jednotlivé třídy a zda se trénované třídy ve zvoleném multispektrálním prostoru dostatečně odlišují svým spektrálním chováním. Vykreslením histogramů pro jednotlivé třídy a pásma lze zjistit, zda pixely vybraných trénovacích ploch mají normální rozdělení. Pokud se nepotvrdí normální rozdělení dat, znamená to, že je třída tvořena dvěma spektrálními podtřídami nebo je definování této třídy neúplné.

Výsledkem prvních dvou fází řízené klasifikace je statistický popis hledaných tříd získaný jen z malé části zpracovaného obrazu – trénovacích ploch. V klasifikační fázi řízené klasifikace jsou pomocí vhodného klasifikátoru jednotlivé prvky obrazu postupně zařazovány do jedné z tříd. Klasifikátory jsou založeny na předpokladu, že obrazové prvky patřící do jedné třídy se budou shlukovat ve stejné části vícerozměrného příznakového prostoru.

Standartními klasifikačními metodami řízené klasifikace jsou tzv. tvrdé klasifikátory, kde je každý pixel považován za „čistý“, obsahuje tedy pouze jednu třídu (Jones and Vaughan, 2010). Nicméně ve vegetačních komunitách obzvláště v takových rozmanitých komunitách, jako jsou louky, je téměř stoprocentně pravděpodobné, že ne všechny pixely budou čisté, ale budou naopak smíšené (Sha et al., 2009). Tyto smíšené pixely se vyskytují buď na hranicích biotopů, nebo pokud je v pixelu obsažena jiná třída.

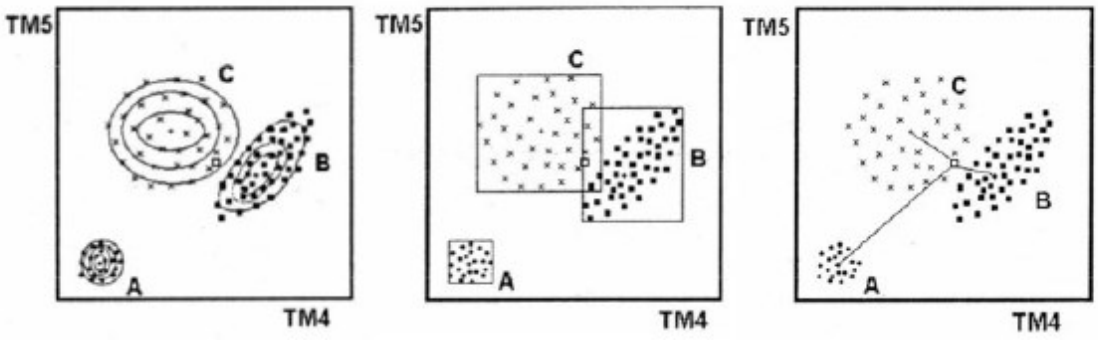
Dle Magiera et al. (2013) smíšené pixely negativně ovlivňují výsledek klasifikace, protože jsou většinou chybně zařazeny do jedné z okolních tříd. Metody měkkého klasifikátoru mohou být při klasifikaci vegetace podstatně lepší. Zatímco dálkový průzkum země obvykle rozděluje vegetaci na skupiny založené na jejich spektrální rozlišitelnosti, ekologové a botanici rozdělují vegetaci na ekologicky nebo botanicky smysluplné skupiny, které však nemusí být spektrálně odlišitelné (Magiera et al., 2013; Verrelst et al., 2009, Schmidtlein and Sassin, 2004).

Proto se v současné době testují metody, jak shromažďovat rostliny do komunit takovým způsobem, aby tyto komunity udržovaly ekologický nebo botanický význam a zároveň bylo možné je rozlišit pomocí dálkového průzkumu Země (Schmidtlein and Sassin, 2004). K často požívaným klasifikátorům řízené klasifikace patří například klasifikátor Maximum Likelihood (maximální pravděpodobnosti) nebo Support Vector Machine.

1.4.1 Klasifikátor Maximální pravděpodobnosti

Klasifikátor maximální pravděpodobnosti (Maximum Likelihood) je široce užívaným algoritmem, který je řazen do kategorie klasifikátorů per-pixel. Klasifikátor tak kvantitativně hodnotí při zatřiďování pixelu rozptyl hodnot, kovarianci a korelaci každé třídy. Předpokladem je normální rozdělení pixelů v trénovacích datech (Lillesand et al., 2014). Klasifikátor je řízen dle předem stanovených trénovacích množin.

Jejich přesná volba je základem pro kvalitní klasifikaci snímku, jelikož i jeden či dva špatně kategorizované pixely v dané trénovací množině mohou mít zásadní vliv na výsledek klasifikace. Klasifikovaný pixel je zařazen do té třídy, do které s největší pravděpodobností spadá (Bolstad, Lillesand, 1991). Hodnoty stejných pravděpodobností tvoří v n-rozměrném příznakovém prostoru tzv. izolinie, které mají tvar elipsy (viz obrázek 1) a pomocí ní, tvoří shluky pixelů (Dobrovolný, 1998).



Obrázek 1: Schéma přiřazování pixelů dle klasifikátorů, zprava: klasifikátor maximální pravděpodobnosti, klasifikátor pravoúhelníků, klasifikátor minimální vzdálenosti, zdroj: Dobrovolný, 1998

Za tohoto předpokladu lze pro každý shluk z tzv. průměrového vektoru a z kovariační matice sestavit frekvenční funkci normálního rozdělení. Průměrový vektor v tomto shluku popisuje míru úrovně. Kovariační matice pak popisuje míru variability. Z teorie normálního rozdělení plyne, že tato frekvenční funkce omezuje spolu s horizontální osou plochu o velikosti 1 (100 %) a pro každou hodnotu na horizontální ose lze vypočítat pravděpodobnost jejího výskytu (Dobrovolný, 1998).

Klasifikátor maximální pravděpodobnosti je výpočetně nejnáročnější a také daleko citlivější na případné nedostatky v trénovacích datech. Nicméně pokud jsou trénovací data kvalitní, dává skvělé výsledky. (Campbell 1996; Lillesand a Kiefer, 1994).

1.4.2 Klasifikátor Support Vector Machine

Klasifikační algoritmus Support Vector Machine (SVM), je zástupcem řízených klasifikací neparametrických metod odvozených ze statistických teorií, který nabízí často dobré klasifikační výsledky pro komplexní data (ENVI, SVM background, 2017). Klasifikace pomocí metody SVM, tedy podpůrných vektorů, patří mezi moderní metody klasifikace obrazových záznamů. Základem metody je rozdělení prostoru podpůrným vektorem tak, aby si obě klasifikační třídy zachovaly maximální vzdálenost od daného vektoru (QC-Expert, SVM Background, 2013).

Tato metoda využívá efektivní algoritmy, pomocí kterých hledá optimální nadrovinu oddělující data trénovacích množin. Optimální nadrovina je taková, od níž mají datové body jednotlivých tříd největší vzdálenost. V případě lineárně neoddělitelných tříd provádí transformaci do prostoru s vyšší dimenzí pomocí jádrových funkcí. Jádrová transformace (kernel type) a chybový parametr C (C parameter) paří mezi základní parametry algoritmu SVM. Volba funkce jádrové transformace je klíčovým faktorem, který má přímý vliv na výsledky klasifikace. Nastavení samotné jádrové funkce může vést např. k přílišnému vyhlazení klasifikovaných tříd ve výsledcích nebo také k jejich přesahu (Mountrakis a kol., 2011).

Klasifikátor SVM byl původně vytvořen pouze jako binární klasifikátor. Následně byl rozšířen pro klasifikaci více tříd. V porovnání s alternativními metodami jako např. Neutral Network může dosáhnout SVM porovnatelných hodnot klasifikačních přesností s menším počtem trénovacích dat (Mountrakis a kol., 2011).

1.5 Posouzení přesnosti klasifikace

Dle Lillesanda and Kiefera (1994), klasifikace není ukončena, dokud není zhodnocena její přesnost. Za chybu v klasifikaci je považován případ, kdy danému pixelu je přiřazen význam jiné třídy, než má ve skutečnosti. Přičemž jak uvádí Campbell (2011), pro chyby v klasifikaci obrazu platí, že chybně klasifikované pixely se ve výsledném obrazu nevyskytují náhodně, ale mají určité prostorové uspořádání. Jsou více méně asociovány pouze s určitými třídami a většinou se nevyskytují izolovaně, ale ve skupinách a jsou svým výskytem vázány na typické části klasifikovaných ploch.

Přesnost výsledků je nutné hodnotit vždy s ohledem na polohu. Celková výměra nalezených tříd může být stejná pro referenční data i klasifikovaný snímek, jednotlivé třídy se však mohou značně lišit svou polohou.

Jak uvádí Dobrovolný (1998), jedním z nejvíce používaných přístupů k hodnocení přesnosti klasifikace je výpočet klasifikační chybové matice. Chybová matice je vždy čtvercová, počet řádků a sloupců odpovídá počtu hodnocených tříd. Pro objektivní testování klasifikace je nutné, aby se pixely, náhodně vygenerované pro srovnání s referenčními daty, nacházely mimo trénovací plochy. Řádky chybové matice představují klasifikovaná data neboli počet vygenerovaných pixelů, sloupce naopak představují data referenční (pixely, kterým byla přiřazena skutečná hodnota).

Jedním z přístupů hodnocení přesnosti je výpočet pomocí chybové matice (viz obrázek 2). Chybová matice, někdy nazývaná také jako matice klasifikace, obsahuje několik měřítek přesnosti. Prvním nejjednodušším opatřením je celková přesnost. Celková přesnost se udává pro celou klasifikace a je vypočítána jako podíl celkové množství správně klasifikovaných pixelů (suma hodnot na hlavní diagonále) ku celkovým množstvím klasifikovaných pixelů (Lillesand et al., 2008).

	Referenční data						
	třída	Voda	Les	Pole	TTP	Půda	SUMA	PU [%]
klasifikovaná data	Voda	**480**	0	5	0	0	485	99
	Les	0	**52**	0	20	0	72	72
	Pole	0	0	**313**	40	0	353	89
	TTP	0	16	0	**126**	0	142	89
	Půda	0	0	0	38	**342**	380	90
	SUMA	480	68	318	224	342	1432	
	CHO [%]	0	23	1	44	0		
	CHZ [%]	1	29	13	7	11		
	PZ [%]	100	76	98	56	100		

Průměrná přesnost: (480 + 52 + 313 + 126 + 342) / 1432 = 92 %
CHU - chyba z opomenutí
CHZ - chyba z nesprávného zařazení
PU - přesnost z hlediska uživatele
PZ - přesnost z hlediska zpracovatele

Obrázek 2: Hodnocení pomocí chybové matice, Zdroj: Dobrovolný 1998

Kromě výpočtu celkové přesnosti lze od chybové matice odvodit i měřítko přesnosti pro každou třídu. V závislosti na tom, co nás zajímá, můžeme počítat počet správně klasifikovaných pixelů určité třídy buď jako zlomek „skutečného“ počtu pixelů této třídy, nebo jako zlomek počtu pixelů zařazených do této třídy (Hromádková, 2015). Z chybové matice lze tedy odvodit další dva parametry jako jsou zpracovatelská a uživatelská přesnost.

Zpracovatelská přesnost, která se na rozdíl od celkové přesnosti počítá pro každou třídu zvlášť, je poměr mezi správně klasifikovanými pixely a pixely použitými pro testování dané třídy. Uživatelská přesnost pak udává, s jakou pravděpodobností pixel zařazený do určité třídy tuto třídu doopravdy představuje. Stejně jako přesnost zpracovatelská se také počítá pro každou třídu zvlášť, a to jako podíl správně klasifikovaných pixelů ku počtu pixelů, které do této kategorie byly zařazeny (Lillesand et al., 2008).

Dalším parametrem hodnocení přesnosti klasifikace, který nalezneme v chybové matici je Kappa koeficient. Kappa koeficient je také široce používaným parametrem posuzování přesnosti klasifikace, který na rozdíl od měření celkové přesnosti zahrnuje normalizaci dat a přihlédnutí k chybě z opomenutí. Kappa koeficient porovnává přesnost provedené klasifikace (určené z chybové matice) s přesností dosažitelnou čistě náhodným zařazením pixelů do jednotlivých tříd. Kappa koeficient se proto používá k určení, zda je klasifikace výrazně lepší než náhoda. Kappa koeficient vyšší než 0,75 znamená, že klasifikátor funguje dobře. Hodnota jedna by znamenala, že při dané klasifikaci bychom se vyhnuli 100 % chyb, které vznikly při čistě náhodném zařazování pixelů do jednotlivých tříd. Na druhou stranu kappa koeficient nižší než 0,40 naznačuje špatný výkon klasifikátoru (Jones Vaughan, 2010).

Základním předpokladem pro hodnocení přesnosti klasifikace a vytvoření chybové matice je mít v rámci klasifikace validační data, se kterými lze výslednou klasifikaci srovnávat. Tato data lze získat buď to terénním výzkumem v daném území a změřením vybraných ploch pro jednotlivé třídy legendy. Pokud však tato data nemáme lze je vytvořit náhodným vygenerováním a přiřadit jim atributy tříd legendy. Tato data jsou nejlépe vybírána pomocí stratifikovaného náhodného výběru (Stratified Random). Pro každou třídu legendy je vybrán počet bodů odpovídající jejímu relativnímu podílu v zájmovém území.

2 VÝBĚR TRÉNOVACÍCH A VALIDAČNÍCH DAT

Klasifikace je jednou z nejčastěji využívaných analýz při práci s daty dálkového průzkumu Země. Je zřejmé, že kvalita použitých trénovacích dat má zásadní význam pro klasifikaci a je rozhodujícím faktorem přesnosti klasifikace. Jelikož je nepraktické vyhodnocovat přesnost celé mapové oblasti, je hodnocení přesnosti klasifikace obvykle založeno na vzorku případů (např. pixelů), které jsou z ní vybrány. Tento vzorek je sadou trénovacích dat. Mnoho studií prokázalo, že se přesnost klasifikace mění v závislosti na rozsahu vlastností trénovacího datasetu (Foody and Mathur, 2006). Aby bylo možné z tohoto vzorku učinit věrohodné zobecnění celé mapy, je důležité, aby trénovací sada dat byla získána po vhodném výběru vzorků. Vzhledem k tomu, že pro přesnost klasifikace je potřeba vybrat trénovací pixely objektivně tedy bez zkreslení způsobeného lidským faktorem je v následujících kapitolách rozebraná metodika, která se v současnosti objevuje v literatuře. Jak již bylo zmíněno v úvodu, literatura věnovaná strategiím výběru trénovacích dat se zaměřuje na tři základní aspekty, které mají pravděpodobně největší vliv na následnou klasifikaci. Těmito aspekty jsou podíl a rozdělení dat do tříd, množství vstupních dat a jejich rozmístění.

2.1 Rozdělení trénovacích dat do tříd

Nejčastější používané rozdělení dat do tříd je proporcionální, kde je množství trénovacích dat rozděleno úměrně dle plochy jednotlivých klasifikovaných kategorií (Jin et al., 2014). Existují však dva další způsoby výběru vzorků. Jedním z nich je equal sample rate, ve které se vybírá pevný procentuální podíl pixelů z každé třídy. Druhý se nazývá equal sample size, ve které se odebírá stejný počet pixelů z každé třídy (Huang et al., 2002).

Při určování trénovacích datových souborů se využívá několik strategií prostorového odběru trénovacích dat. Obecně statistické testy předpokládají náhodné prostorové vybírání trénovacích dat a často jej označují za nejvhodnější typ výběru. Nicméně náhodně přidělené body jsou často nepřístupné z důvodu obtížného terénu nebo jiným překážkám, navíc malé kategorie mohou být podhodnoceny nebo zanedbány.

Dalším typem prostorového odběru trénovacích dat je shlukování. Takovéto vybírání trénovacích dat se v tomto případě využívá, pokud je omezený čas a prostředky pro výzkum v terénu (Jones Vaughan, 2010).

Náhodný výběr trénovacích dat lze často zlepšit zavedením tzv. systematic sampling a stratifikací, což jsou obvyklé způsoby modifikace náhodného výběru trénovacích dat (Gallego, 2005). Latin hypercube sampling (LHS), stratifikovaný náhodný postup, je účinným prostředkem vzorkování proměnných (McKay et al., 1979 , Minasny a McBratney, 2006). Metoda LHS je numerická simulační metoda typu Monte Carlo, která je vhodná pro realizaci pravděpodobnostních analýz. Metoda Latin hypercube sampling je obecně použitelná v libovolném oboru, kde je nutno sledovat charakteristiky náhodného výstupu v závislosti na popsaném náhodném vstupu. Autoři Svoboda a Hillar (2013) zkoumali možnost použití zlepšené metody LHS ke generaci náhodných procesů metodou ortogonální transformace korelační matice. Výhoda metody plyne ze způsobu výběru realizací, kdy celý rozsah náhodné proměnné je pokryt rovnoměrně vzhledem k distribuční funkci. Žádná reálná hodnota není předem vyloučena. Současně metoda zachovává zjištění odhadnuté funkce hustoty pravděpodobnosti pro jednotlivé náhodné veličiny a stanovené korelační koeficienty mezi nimi (Svoboda a Hilar, 2013).

Yu-Pin (2011) ve své práci zkoumala vliv výběru metody výběru trénovacích dat na výslednou přesnost klasifikace. V její studii je celá zájmová oblast rozdělena na bloky pro tento výběr metodou LHS. Množství vybraných trénovacích dat potvrzuje, že tato data vybraná pomocí LHS, poskytují dostatečné pokrytí zájmové oblasti.

Analýzy také ukazují, že přístup LHS může být použit k výběru trénovacích dat a zachycení prostorových struktur s větší přesností z obrazů NDVI. Tato studie navrhuje nový přístup k výběru trénovacích dat nazvaný stratifie conditional latin hypercube sampling (scLHS).

Tento přístup integruje přístup variance quadtree technique (VQT) a conditional latin hypercube sampling (cLHS) k vzorovému multiple, vzdáleně snímanému obrazu a usnadňuje efektivní sledování změn krajiny.

Rozbor více obrazů NDVI v různých časových krocích je nezbytný pro charakteristiku a kvantifikaci prostorové variability, struktury a heterogenity změn krajiny. Výsledky scLHS variaografie ukazují, že prostorová variabilita, struktura a heterogenita více obrazů NDVI ve studované oblasti může být zachycena dostatečným počtem trénovacích dat scLHS.

Yu-Pin (2009) metodu LHS kombinuje se sekvenční Gaussovou simulací (SGS). Metoda je aplikovaná pro generování více výsledků, včetně chybové složky, která chybí v klasických interpolačních přístupech (Lin, 2008). Tato studie představuje nový a efektivní přístup, který integruje cLHS, variogramy, kriging a SGS v dálkově snímaných snímcích pro efektivní sledování, odběr trénovacích dat a mapování dopadů změn krajiny na prostorové struktury.

Abychom zabránili nižšímu počtu trénovacích dat malých tříd, můžeme zvolit stratifikované verze trénovacích strategií (Belousov et al., 2002). To znamená, že studovaná oblast je nejprve rozdělena na polygony podle tříd a poté jsou odebírána trénovací data podle zvoleného vzorkování z každého polygonu. Dle Jonese Vauhana (2010) je tento přístup často neúčinnější strategií výběru.

Millard (2015) ve své studii zkoumal vliv výběru trénovacích dat náhodně v každé třídě klasifikace. Z plné sady 500 bodů byla náhodně oddělena sada 100 bodů pro validaci a množina 400 náhodných bodů trénovacích dat. Dále byly vytvořeny sady údajů trénovacích dat, aby se zajistil vliv podílu trénovacích dat v různých třídách, tak aby byl vždy zachován zadaný celkový počet trénovacích dat. Pro každou třídu provedl zvolený poměr 25krát s náhodným dílčím vzorkem trénovacích dat a validačními daty pro nezávislé ověření. Autoři studie zjistili, že když trénovací data byla vytvořena proporcionálně v každé třídě, tak výsledná klasifikace prokázala značnou přesnost. Celkově výsledky této studie ukazují, že klasifikace snímků je vysoce citlivá na charakteristiku trénovacích dat, včetně velikosti trénovacího datasetu, poměru tříd a prostorové autokorelace.

Vytvořením skupin trénovacích pixelů se zabývala i Hromádková (2015) ve své práci. Hromádková (2015) nejprve definovala oblast, odkud mohla být vybrána centra shluku. Tato oblast obsahovala každý pixel, který byl vzdálen více než 1 nebo 2 metry od hranice polygonu. Za druhé, centra shluků náhodně definovala podle metody, kterou využila již pro náhodný výběr pixelů, které byly určeny pomocí funkce random.sample. Tyto pixely byly následně vybrány pomocí funkce Select Layer Attribute management a uloženy jako nová vektorová vrstva. Výběr byl potom přepnut a zbylé pixely byly použity jako validační. Následně autorka definovala sousedství kolem těchto center. 8-N okolí bylo definováno jako všechny pixely ve vzdálenosti 1 metru od center.

Nakonec bylo z definovaných čtvrtí náhodně vybráno určité množství pixelů s využitím techniky, která byla využita pro náhodný výběr pixelů, tyto pixely pak byly uloženy jako trénovací a zbylé jako validační.

Foody and Marthur (2006) ve své studii uvádějí, že rozhodující paradigma při navrhování fáze výběru trénovacích dat je silně založeno na konvenčních statistikách. V tomhle případě se fáze výběru trénovacích dat považuje za cíl, který je zaměřen na odvození přesnosti popisu každé třídy. Odvozené popisné statistiky se pak mohou použít pro přidělování každého pixelu, který nemá definovanou třídu. Tyto pixely jsou pak přiřazeny třídě, s níž má největší podobnost.

Foody and Marthur (2006) proto požadují pro klíčové otázky v návrhu trénovacího datasetu literatura použití statistických přístupů k odvození popisné statistiky každé třídy. Dle autorů hlavními obavami v souvislosti s použitím takového přístupu je to, že získaný vzorek trénovacích dat by měl poskytnout reprezentativní a nestranný popis třídy.

2.2 Množství trénovacích dat

Foody a Mathur (2006) uvádí, že množství trénovacích dat má mít vliv na přesnost klasifikace, přičemž optimální velikost datasetu trénovacích pixelů je obvykle spojena s výběrem klasifikátoru a rovnoměrností dat. Huang et al., (2002) ve své studii tvrdí, že přesnost klasifikace se zvýší s větším množství validačních dat.

Při návrhu a provádění fáze výběru trénovacího datasetu dané klasifikace je třeba vzít v úvahu řadu otázek. Obecně se předpokládá, že cílem fáze výběru trénovacího datasetu je v podstatě definovat přesný model tříd. Pro každou třídu by proto trénovací dataset měl poskytnout reprezentativní popis. To vyžaduje, aby místa trénování byla dostatečně početná a byla rozložená po celé zájmové oblasti.

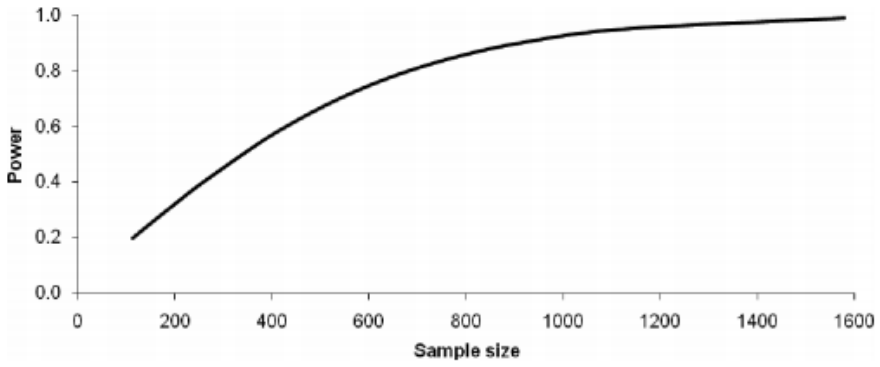
Velikost souboru trénovacích dat je na druhé straně spojená s klasifikátorem, který má být použit a také s charakteristikami souboru dat, který má být zařazen. Například klasifikátor maximální pravděpodobnosti pomocí střední a kovariační matice shrnuje spektrální odezvu každé třídy, zatímco vícevrstvá neuronová síť percepton používá přímo každý trénovací pixel.

Na rozdíl od klasických klasifikátorů klasifikátor SVM jako hraniční klasifikátor používá jen pixely na hranicích tříd, protože tyto pixely jsou nejlepší pro umístění tzv. hyperplane, která odděluje třídy. Proto se předpokládá, že minimální velikost trénovacího souboru je menší pro SVM klasifikátor než pro klasické klasifikátory, zatímco přesnost zůstává vysoká (Foody and Mathur, 2004).

Velká část literatury je založena na faktu, že zvýšení přesnosti klasifikace je závislé na velkém množství čistých trénovacích pixelů. Foody and Mathur (2006) se snažili o přímé srovnání mezi dvěma přístupy k trénování. Zjistili, že klasifikace za použití stejného počtu trénovacích pixelů pro každou třídu v souvislosti s analýzou založenou na smíšených spektrálních odezvách přinesla podobnou přesnost klasifikace jako další analýzy, které byly založeny na čistých pixelech v každé třídě.

Výsledky tedy ukazují, že je možné použít malý trénovací dataset založený na smíšených spektrálních odezvách a dostaneme stejný výsledek klasifikace s podobnou přesností jako u jednoho trénovacího datasetu s použitím většího počtu čistých pixelů získaných konvenčním způsobem (Foody and Mathur, 2006).

Foody (2009) ve své studii zkoumal výběr dostatečného množství trénovacích dat. Pozornost věnuje stanovení finální míry přesnosti klasifikace. Dle Foodyho by mělo být množství trénovacích dat závislé na požadované míře přesnosti klasifikace. Zkoumal tři míry přesnosti. Ve všech třech případech zjistil, že přesnost klasifikace je vždy větší, pokud je trénovacích dat více viz obrázek 3.



Obrázek 3: Vztah mezi množstvím trénovacích dat a výslednou přesností klasifikace

Často se uvádí, že počet trénovacích pixelů pro každou třídu by měl obsahovat nejméně 10–30násobek počtu vlnových pásem nebo jiných diskriminačních proměnných použitých v analýze. Je samozřejmě žádoucí, aby jen pixely, které opravdu popisují danou třídu, byly v téhle třídě zahrnuty. Takto vybrané pixely pro účely výběru trénovacího datasetu by měly být čistými členy příslušné třídy. K dosažení takového výsledku se často záměrně maskují nebo vylučují hranice oblasti, kde může dojít ke smíchání spektrálních reakcí tříd (Foody and Mathur, 2006).

Zhu a Gallant (2016) zkoumali jaké množství trénovacích pixelů je optimálních pro zpřesnění maximální přesnosti. Došli k výsledku, že je potřeba více než 10 000 trénovacích pixelů k tomu, aby bylo zpřesnění klasifikace znatelné.

Dále Pal and Mathur (2006) uvádějí pravidlo, kterým je definován vztah mezi množstvím trénovacích pixelů a množstvím validačních pixelů. Toto pravidlo je definováno tak, že by měl existovat alespoň dvojnásobek validačních pixelů, než je trénovacích pixelů.

Kang (2015) zkoumal velikost trénovacího datasetu na přesnost klasifikace podle měnící se velikosti trénovacího datasetu. Klasifikace byly spuštěny s různými velikostmi trénovacího datasetu. Pro každé opakování bylo vygenerováno 100 náhodných bodů pro validaci a zbylých 400 bodů bylo označeno jako trénovacích. Z těchto bodů vytvořily různé podmnožiny pro výcvik z 90 % až na 30 % dat a provedli tak 25 klasifikací. Z výsledků této studie je evidentní, že když se množství trénovacích dat zvyšuje, chyba klasifikace se obecně snižuje, proto by značně rostoucí velikost trénovacího datasetu měla vést k zvýšení přesnosti klasifikace.

Hromádková (2015) také zkoumala množství trénovacích pixelů a jejich vliv na přesnost SVM klasifikace. Závěr, který udělala je, že se zvyšujícím se počtem trénovacích pixelů se také zvyšuje přesnost klasifikace. Vzhledem k tomu, že větší množina trénovacích dat má větší šanci na zahrnutí podpůrných vektorů, které definují skutečné hranice, a proto by měly poskytovat větší přesnost (Huang et al., 2002). Navíc scénáře výběru trénovacích dat se stejnou equal sample rate poskytují lepší celkovou přesnost než scénáře equal sample size.

Přesto, když Hromádková (2015) zkoumala chybovou matici, zjistila, že equal sample rate scénáře podceňovaly méně hojné třídy a pixely těchto tříd přiřazovaly k třídám s nejvíce pixely. Toto chování bylo pozorováno i v publikaci Huang et al. (2002) nejen pro algoritmus SVM.

2.3 Rozmístění trénovacích dat

Rozmístění dat může mít také velký vliv na přesnost klasifikace, například trénovací polygony na hranicích daného zkoumaného území, pro které máme nasbíraná data, mohou mít obrovský vliv. Například pixely, které jsou spektrálně daleko od středu většiny pixelů pro danou třídu a pixely umístěné v těsné blízkosti hranic daného zkoumaného území lze identifikovat jako odlehlé (Radoux et al., 2014).

Jonese a Vaughana (2010) a Huanga et al. (2002) prohlásili, že trénovací pixely, umístěné příliš blízko sebe, nesou stejnou korelovanou informaci a nemohou tedy poskytnout dostatečnou přesnost. Proto se nedoporučuje používat návrh vzorkování založeném na povrchu tak jak odpovídá realitě, a to i tehdy, je-li to považováno za nejméně náročný návrh vzorkování při zvažování materiálních a lidských zdrojů. Naopak Zhu a Gallant (2016) zjistili, že zlepšení přesnosti klasifikace je založeno právě na rovnoměrném rozložení dat přibližně až o 8 %.

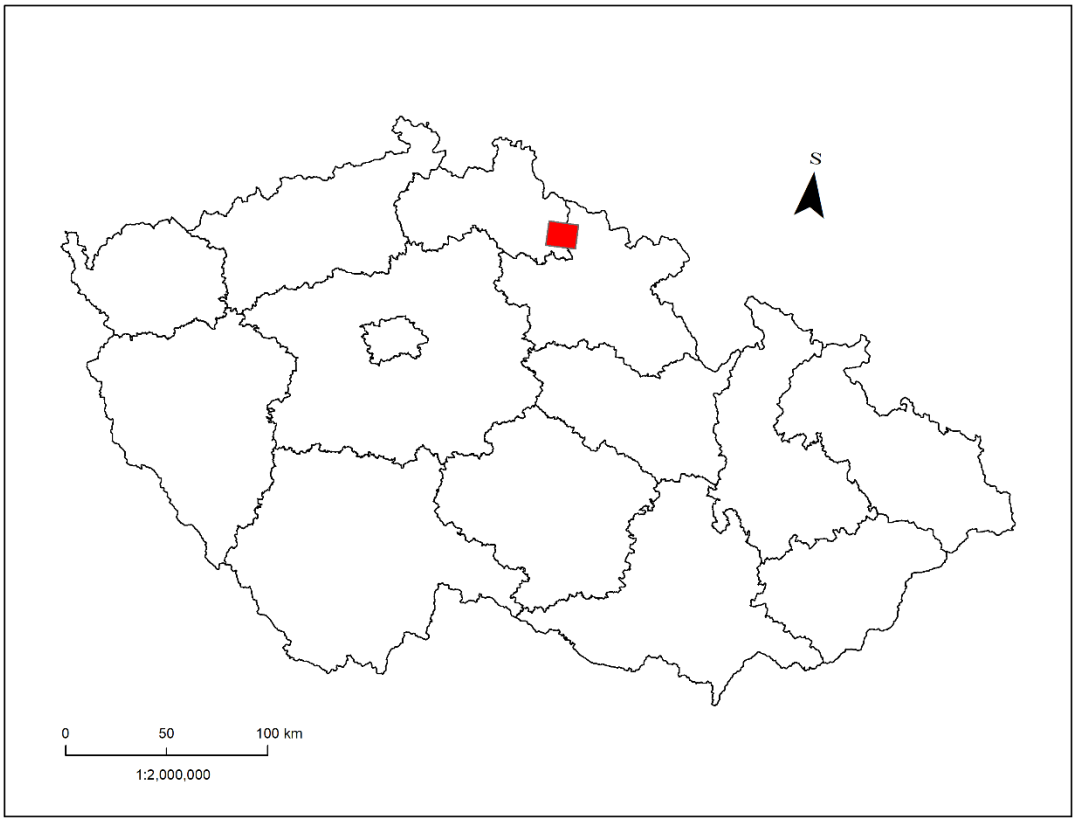
Foody and Mathur (2004, 2006) testovali vliv hraničních trénovacích dat na přesnost klasifikace SVM, a to buď pomocí zeměpisných hranic, nebo hranic definovaných fyzickými proměnnými. Bylo zjištěno, že tato účelně vybraná trénovací data mohou poskytovat vyšší nebo srovnatelnou přesnost než konvenčně definované tréninkové soubory. Odzkoušení prostorového vzorkování "reálného života" bylo testováno, protože je výhodné. Všechny testované strategie výběru byly prováděny stratifikovaným způsobem.

Určení hraničních pixelů dobře popsala Hromádková (2015). Aby autorka vytvořila množinu trénovacích dat s pixely na hranicích tříd, nejprve identifikovala samotnou hranici. Následně všechny pixely do 1 metru od vnější čáry polygonu považovala za hraniční. V dalším kroku z hraničních pixelů náhodně vybrala určité množství pixelů a výběr uložila jako novou vektorovou vrstvu s tréninkovými pixely. Zbývající pixely opět uložila jako validační.

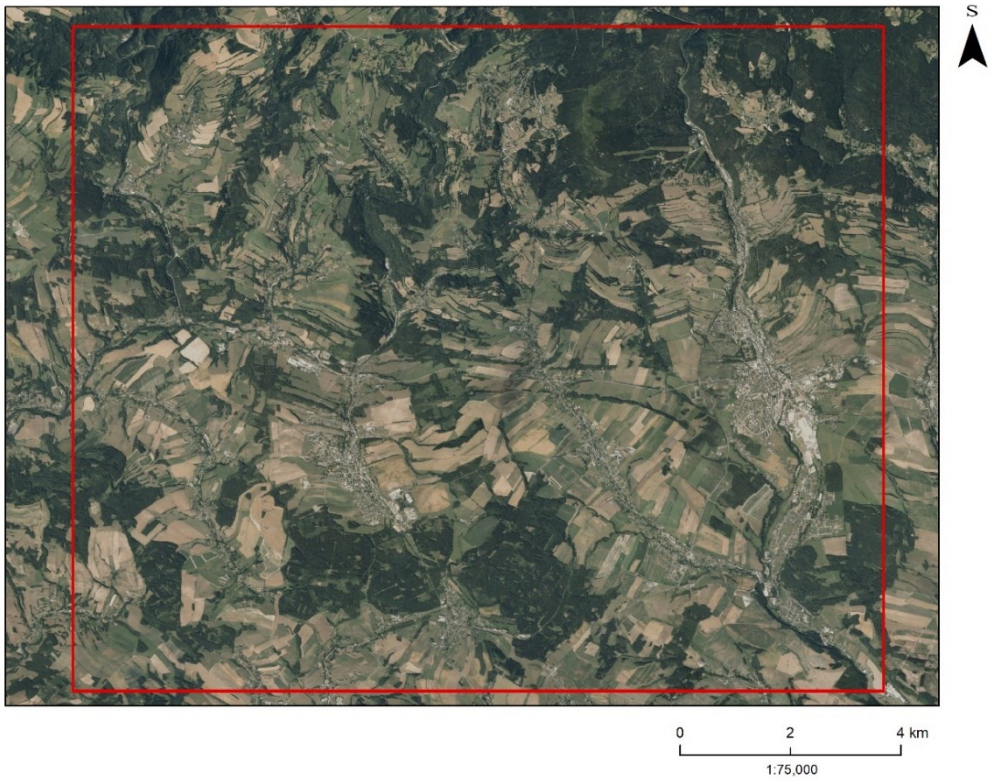
Autorka vycházela z již zmíněné hypotézy od Foodyho and Mathura (2004,2006), že algoritmus SVM je schopen klasifikovat datové sady se stejnou nebo vyšší přesností, když se použijí hraniční trénovací pixely. Výsledky od Hromádkové (2015) opravdu potvrzují tuto hypotézu. Přesnost získaná použitím trénovacích pixelů na hranicích byla srovnatelná s přesností získanou při použití náhodně vybraných tréninkových pixelů. Tyto výsledky byly pravděpodobně získány, protože podpůrné vektory, které jsou používány algoritmem SVM k vytvoření oddělující hyperplane, jsou skutečně umístěny na hranicích polygonů spíše než v centrech těchto polygonů (Foody and Mathur 2004,2006).

3 CHARAKTERISTIKA ÚZEMÍ

Pro experimenty s trénovacími a validačními daty bylo vybráno zájmové území, které se nachází na hranici Libereckého a Královehradeckého kraje v podhůří Krkonoš. Zájmové území se konkrétně nachází v okolí měst Vrchlabí a Jilemnice, polohu území a krajinný pokryv lze vidět na obrázcích 4 a 5. Území má rozměry přibližně 10 x 15 km a celková rozloha činí 177,5 km^2. Největší plochu zaujímají lesy - přibližně 37 % území, trvalý travní porost zaujímá 36 % rozlohy, orná půda bez vegetace 13 %, orná půda s vegetací 6 %, zástavba 7 % a vodní plochy 0,2 % rozlohy.



Obrázek 4: Poloha zájmového území v ČR



Obrázek 5: Zájmové území – ortofoto

4 DATA A METODIKA

4.1 Použitý software

Jedním z využitých softwarů je software ENVI. Tento software vznikl především za účelem zpracování družicových snímků, ale umožňuje také práci s radarovými nebo jinak nasnímanými daty dálkového průzkumu Země. Využití softwaru tkví v přednosti vizualizace velkého objemu dat, která je následně možné analyzovat, klasifikovat a detekovat změny (Harris Geospatial Solutions, 2018).

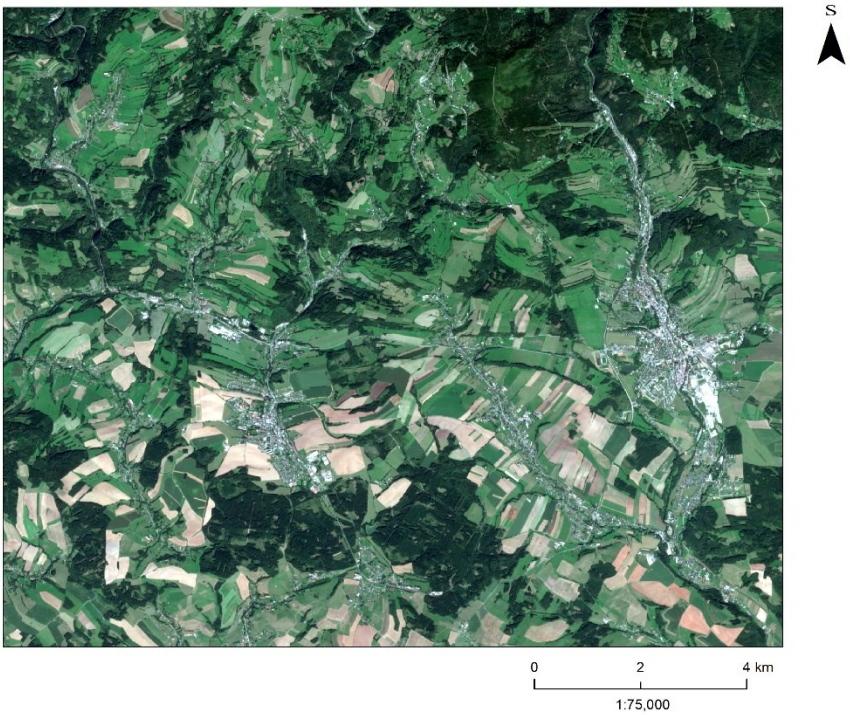
Pro zpracování snímků Sentinel byl také využit software SNAP. Jedná se o volně dostupný balíček nástrojů, který slouží ke snadnému zobrazení RGB obrazu, jeho ortorektifikaci, reprojekci nebo převzorkování (ESA, 2018).

V rámci diplomové práce byl dále využit software ArcMap. Což je jeden z nejpopulárnějších komerčních softwarů z řad ArcGIS, který je provozován firmou ESRI. Slouží pro zpracování mapových podkladů, prostorových analýz a k editaci dat (ARCDATA,2018).

4.2 Využitá data

V diplomové práce byl využit snímek Sentinel-2A. Volně dostupný snímek byl stažen po registraci na webu Sentinel SciHub (https://scihub.copernicus.eu). Hlavním kritériem výběru snímku byl nízký podíl oblačnosti a také datum jeho pořízení, které by mělo být co nejblíže termínu, který se shoduje s termínem vzniku aktuálního ortofota pořízeného pro zájmové území. Snímky byly tedy vybírány pro rok 2016 od května do října 2016 (tak aby byl snímek bez sněhové pokrývky a nejlépe v průběhu vegetační sezóny). Ve zmíněném období byl nalezen pouze jeden snímek, který splňoval podmínky.

Snímek využívaný pro účely této diplomové práce je snímek produktu level 1 C z termínu 28.8.2016, který je pro dané území zcela bezoblačný.



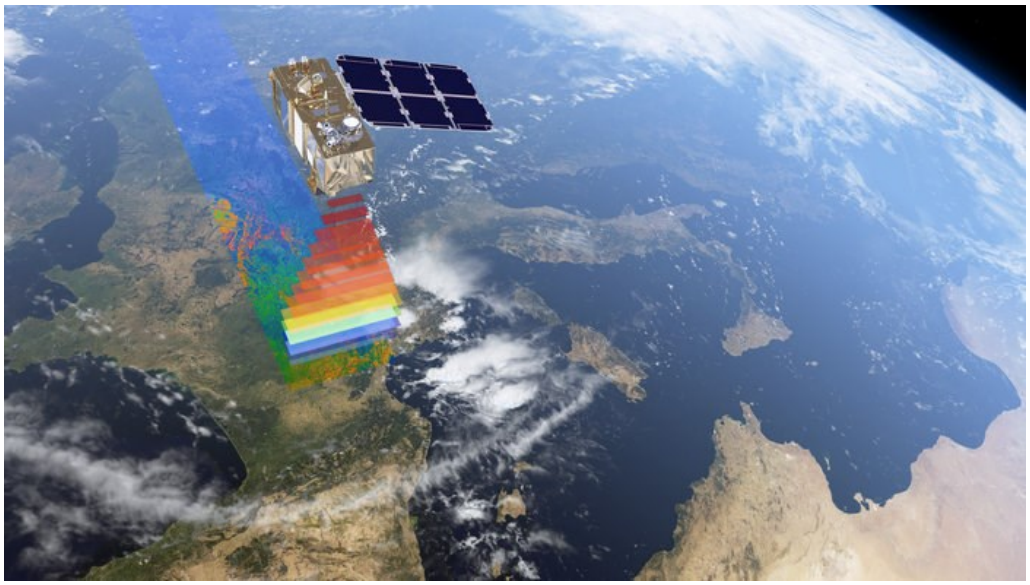
Obrázek 6: Zájmové území – snímek S-2A

Kódové označení využitého snímku:

S2A_OPER_PRD_MSIL1C_PDMC_20160828T210754_R022_V20160827T101022_20160827T101025.

4.3 Definice dat Sentinel-2A

Podle Radoux (2016) zažívá dálkový průzkum země novou éru vyznačující se velkým množstvím volně dostupných satelitních dat, která nabízejí vysoké spektrální rozlišení jak ve viditelném, tak infračerveném spektru a jejich snímání v krátkém časovém horizontu. Sentinel 2 je mise programu Copernicus, která poskytuje multispektrální snímky s vysokým rozlišením a nebývale velkou šířkou záběru. Sentinel 2 tvoří dvě družice na stejné oběžné sráze s posunem 180 °. Samotná družice prolétne nad stejným místem na Zemi jednou za 10 dní, dvě družice na jednou za 5 dní (to platí na rovníku, ve vyšších zeměpisných šířkách se tato doba zkracuje) čímž je dosaženo vysokého časového rozlišení.



Obrázek 7: Družice Sentinel 2 (ESA b, 2017)

Jak uvádí (Drusch et al., 2012), Sentinel-2 je vybaven sensorem Multispectral imager (MSI), který snímá ve 13 spektrálních pásmech a prostorové rozlišení má 10, 30 nebo 60 m, v závislosti na spektrálním pásmu. Využitý snímek Sentinel-2A je již ortorektifikovaný snímek po atmosférické korekci. Každý produkt Level 2-A se skládá z dlaždic o rozměru 100 km^2 v kartografickém zobrazení UTM / WGS84.

V porovnání s ostatními multispektrálními daty je jeho přínos právě v kombinaci vysokého prostorového a spektrálního rozlišení s 290 km širokým záběrem sensoru. Také díky pásmům v červeném okraji viditelného spektra jsou tato data vhodná především pro monitoring krajinného pokryvu (Gisat, 2017).

Tabulka 1: Spektrální pásma Sentienl-2A (dle Gisat,2017)

označení pásma	název pásma	prostorové rozlišení	rozsah od (µm)	rozsah do (µm)
1	Coastal aerosol	60	0,433	0,453
2	Blue	10	0,4575	0,5225
3	Green	10	0,5425	0,5775
4	Red	10	0,65	0,68
5	Vegetation red edge	20	0,6978	0,7125
6	Vegetation red edge	20	0,7325	0,7475
7	Vegetation red edge	20	0,773	0,793
8	NIR	20	0,7845	0,8995
8a	Vegetation red edge	10	0,855	0,875
9	Water vapour	60	0,935	0,955
10	SWIR-Cirrus	60	1,365	1,395
11	SWIR	20	1,565	1,655
12	SWIR	20	2,1	2,28

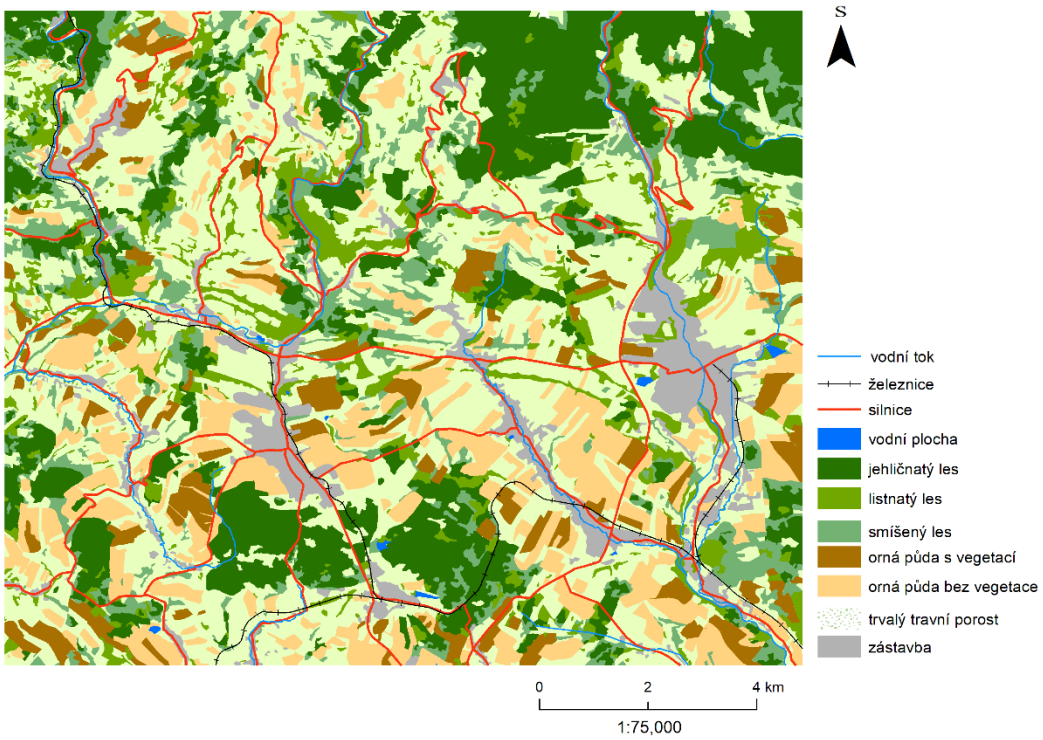
4.4 Předzpracování dat

Daný snímek Sentinel 2A byl oříznut pomocí funkce Resize data podle vektorové vrstvy hranice zájmového území. V rámci předzpracování dat byla vytvořena tematická vektorová vrstva na podkladu snímku. Pro tvorbu přesné tematické vektorové vrstvy byla využita služba WMS ČÚZK dostupná z geoportálu ČÚZK.

Pro vytvoření tematické vektorové vrstvy je potřeba stanovit legendu tříd zájmového území a pro hodnocení přesnosti klasifikace je potřeba mít referenční a trénovací data. Jako referenční data byla využita vytvořená tematická vektorová vrstva. Pro vytvoření vektorové vrstvy byla stanovena následující legenda, viz tabulka 2.

Tabulka 2: Legenda pro tvordu tematické vektorové vrstvy

Třída
Travní porost
Jehličnatý les
Listnatý les
Smíšený les
Orná půda bez vegetace
Orná půda s vegetací
Vodní plocha
Zástavba



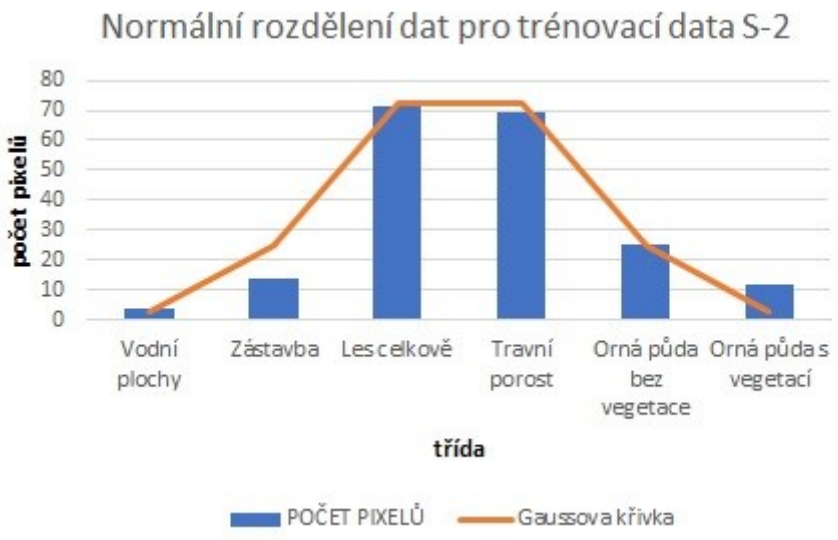
Obrázek 8: Tematický vektorový výstup manuální klasifikace snímku na základě vizuální interpretace

4.5 Analýza dat

Na základě rešerše literatury a zdrojů byly vybrány klasifikátory Maximum Likelihood a SVM jako metody pro řízenou klasifikaci.

4.5.1 Testování vlivu podílu/množství trénovacích a validačních dat na přesnost klasifikace metodou Maximum Likelihood

Metoda MLC vyžaduje tzv. normální rozdělení trénovacích dat, které je charakterizováno Gaussovou křivkou (Jensen, 2005). Trénovací množiny, které byly vybrány náhodným stratifikovaným výběrem, reprezentované počty pixelů by měly po zanesení do grafu zobrazovat tvar Gaussovy křivky. Z trénovacích množin byly spočítány směrodatné odchylky, které byly potřeba pro spočítání normálního rozdělení pravděpodobnosti. Dále byl zvolen interval, na kterém bylo spočítáno normální rozdělení dat. Pomocí hodnot z distribuční tabulky normálního rozdělení dat (ČVUT, 2018) byly vypočítány jednotlivé hodnoty normálního rozdělení, které byly následně vynásobeny příslušným koeficientem tak, aby se hodnoty přiblížily hodnotám počtů pixelů pro jednotlivé třídy (Kuthan, 2019). Výsledek lze vidět na grafu č. 1. Vidíme, že data podléhají normálnímu rozdělení dat.



Graf 1: normální rozdělení trénovacích dat reprezentované Gaussovou křivkou

Pro trénovací data byla také otestována jejich odlišitelnost v SW ENVI pomocí nástroje ROI Separability. Dá se předpokládat, že trénovací plochy s větší separabilitou přinesou také lepší výsledky při klasifikaci. Hodnoty míry separability se pohybují v rozmezí 0 – 2 pro každou dvojici tříd a čím je hodnota vyšší, tím je zajištěna lepší separabilita (Harris Geospatial Solutions, 2018). V průběhu testu byly testovány třídy dle legendy viz tabulka 2. Míra separability však byla pro dvojici travní porost a orná půda s vegetací na hodnotě 0.3359. Z tohoto důvodu byla tato dvojice tříd, stejně jako všechny druhy lesa, sloučena do jedné a test separability tříd byl proveden pro nově vzniklou legendu viz tabulka č. 3, která byla označena jako finální pro vstup do dalších pokusů.

Tabulka 3: finální legenda

	Třída	**Podíl plochy**
1	Les	37 %
2	Zástavba	7 %
3	Půda s vegetací	42 %
4	Půda bez vegetace	13 %
5	Vodní plocha	1 %

Hodnoty separability tak byly v tomto případě podstatně lepší. Nejlepší míry separability dosáhly dvojice les – zástavba, orná půda bez vegetace – les, zástavba – vodní plocha, orná půda bez vegetace – vodní plocha. Nejhůře pak dopadla dvojice půda s vegetací – půda bez vegetace viz tabulka 4.

Tabulka 4: výsledky testu separability

	1	2	3	4	5
1	x	2.0000	1.7550	2.0000	1.7621
2	2.0000	x	1.8088	1.9664	2.0000
3	1.7550	1.8088	x	1.6785	1.9998
4	2.0000	1.9664	1.6785	x	2.0000
5	1.7621	2.0000	1.9998	2.0000	x

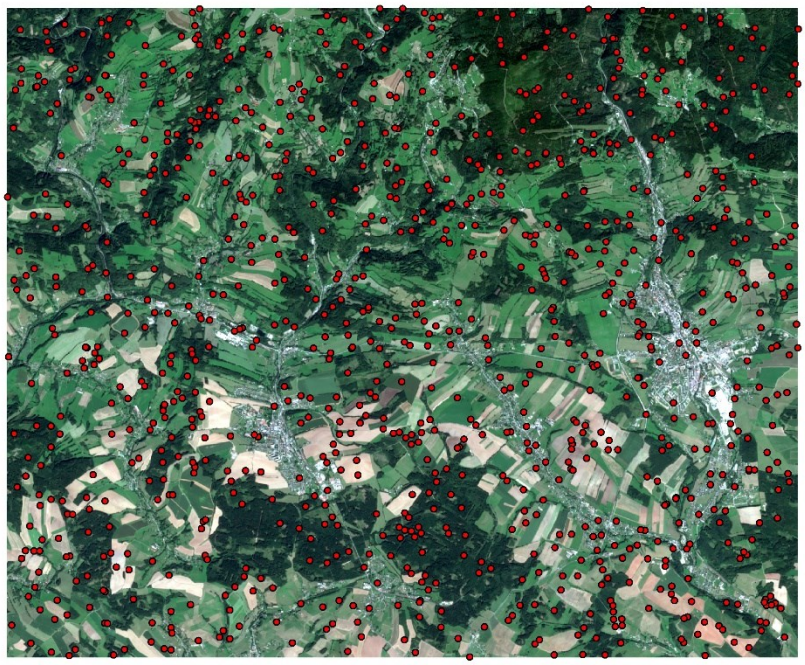
V rámci analýzy směřující k optimalizaci tvorby trénovacího a validačního datasetu byl nejprve testován měnící se podíl trénovacích a validačních dat. Na základě literatury Foodyho (2009) bylo určeno, že pro tento test bude dostatečné množství 1000 bodů, které budou stratifikovaně rozděleny do jednotlivých tříd, výsledek rozdělení viz tabulka 5.

Tabulka 5: stratifikované množství bodů v jednotlivých třídách

	Třída	**Počet bodů**
1	Les	370
2	Zástavba	70
3	Půda s vegetací	420
4	Půda bez vegetace	130
5	Voda	10

Body byly stratifikovaně vybrány pomocí skriptu v příloze 1. Pro každou třídu byl funkcí Create Random Points vygenerován daný počet bodů a pomocí funkce

Calculate Management Field byly každému bodu přiřazeny atributy určující třídu legendy k další klasifikaci.



Obrázek 9: Příklad rozmístění bodů pro trénování či validaci

Následně byly tyto body podílově rozděleny na trénovací a validační množinu dat. Trénovací data byla vybírána od počtu 50 bodů (zároveň tedy bylo 950 validačních bodů) po kroku 5 % (po 25 bodech) až do počtu 950 bodů (zároveň tedy 50 validačních bodů). Body byly vždy vybírány stratifikovaně dle podílu plochy dané třídy viz skript příloha 2. Pro zachování celočíselných počtů bodů, byly nejméně početné třídy zaokrouhlovány směrem nahoru. Následně byl sečten počet těchto bodů. Pro nejpočetnější třídu byl počet bodů vytvořen jako doplněk do celkového požadovaného počtu bodů pro daný krok.

Následná klasifikace Maximum Likelihood byla provedena funkcí MLC Classify. Klasifikován byl výše specifikovaný snímek Sentinelu 2A. Klasifikace pro každý podíl trénovacích a validačních bodů proběhla 1000x.

4.5.2 Testování vlivu změny velikosti validačního datasetu na stabilitu výsledku hodnocení přesnosti klasifikace metodou Maximum Likelihood

Veškerá data byla validována s využitím kontrolních množin, které byly vytvořeny doplňkem trénovacích dat do 1000. Pro počet trénovacích bodů, při kterém bude výsledek s nejvyšší hodnotou přesnosti klasifikace, bude proveden další test. Tento test bude pro daný a stálý počet trénovacích bodů měnit pouze počet validačních bodů, pro ověření vlivu množství validačních dat na přesnost klasifikace. Pro tento test byl skript v příloze 4 doplněn o parametr množství validačních dat. Validační data byla také stratifikovaně rozdělena do tříd dle podílu plochy. Následně byla vypočítána chybová matice pomocí funkce Compute Confusion Matrix. K validaci byly využity parametry celková přesnost, Kappa koeficient, zpracovatelská přesnost a uživatelská přesnost. Skript tedy jako výsledek vypsal 1000x všechny parametry pro 1000 klasifikací.

4.5.3 Testování vlivu podílu/množství trénovacích a validačních dat na přesnost klasifikace metodou Support Vector Machine

K otestování, zda pro klasifikační algoritmy Maximum Likelihood a Support Vector Machine je optimální podíl trénovacích a validačních dat stejný, byl upraven skript viz příloha 3. Trénovací body byly opět podílově rozděleny na trénovací a validační množinu dat. Trénovací data byla vybírána od počtu 375 trénovacích bodů (zároveň tedy 625 validačních bodů) po kroku 10 % (po 50 bodech) až do počtu 275 trénovacích bodů a následně do počtu 475 trénovacích bodů. Podíl 375 trénovacích a 625 validačních bodů byl zvolen v návaznosti na výsledky získané pro metodu Maximum Likelihood (viz kapitola 5 Výsledky) Klasifikace SVM byla provedena úpravou skriptu, kdy funkce pro tvorbu algoritmu Maximum Likelihood byla nahrazena funkcí Train Support Vector Machine Classifier. Klasifikace byla provedena bez využití segmentace dat.

5 VÝSLEDKY

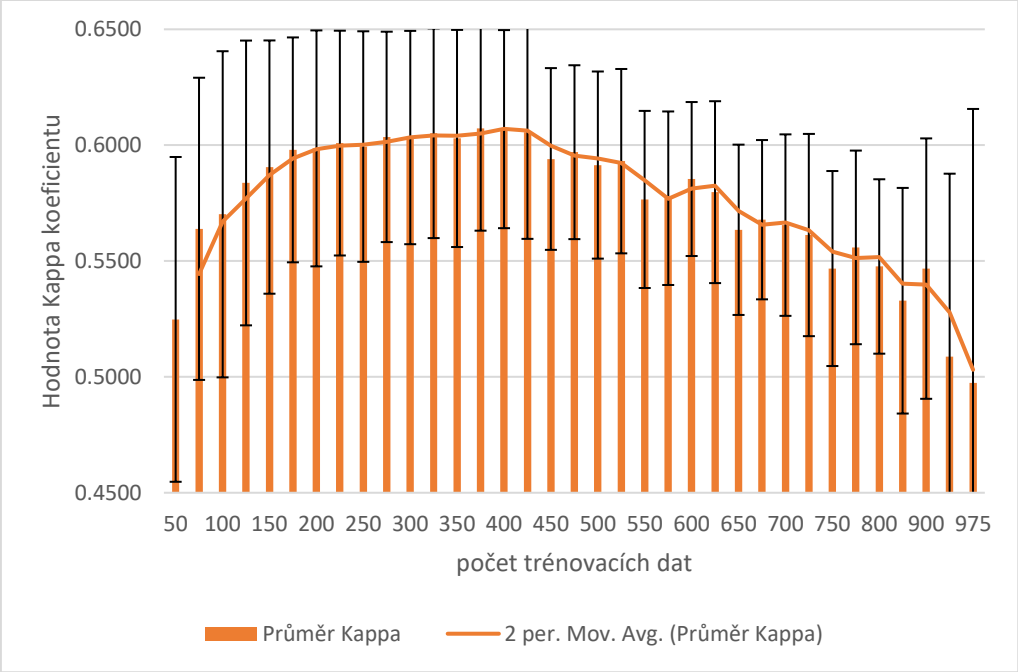
5.1 Testování vlivu podílu/množství trénovacích a validačních dat na přesnost klasifikace metodou Maximum Likelihood

Jak je popsáno v metodice (kapitola 4), pro testování trénování a validaci klasifikace bylo pracováno se souborem 1000 bodů, které byly stratifikovaně rozděleny dle zastoupení jednotlivých kategorií legendy v rámci zájmového území. Poměr trénovacích a validačních bodů byl měněn po 25 trénovacích bodech. Validační data byla vytvořena jako doplněk (k trénovacím datům) do 1000. V analýze bylo využito celkem 38 různých poměrů trénovacích a validačních dat. Klasifikace byla pro každý poměr spuštěna 1000x. Proces byl prováděn s využitím skriptu (viz Příloha 1, 2, 3 a 4) a celkem bylo vytvořeno 38 000 výsledků klasifikace. Přesnost klasifikace pro každý počet trénovacích bodů (a příslušný doplněk validačních bodů) byla hodnocena s využitím kappa koeficientu, parametru celková přesnost, přesnost zpracovatelská a přesnost uživatelská. V následující kapitole jsou popsány výsledky testů, které jsou pro názornější orientaci znázorněny v grafech. V grafech je na ose x vyčísleno množství trénovacích bodů.

5.1.1 Kappa koeficient a celková přesnost klasifikace MLC

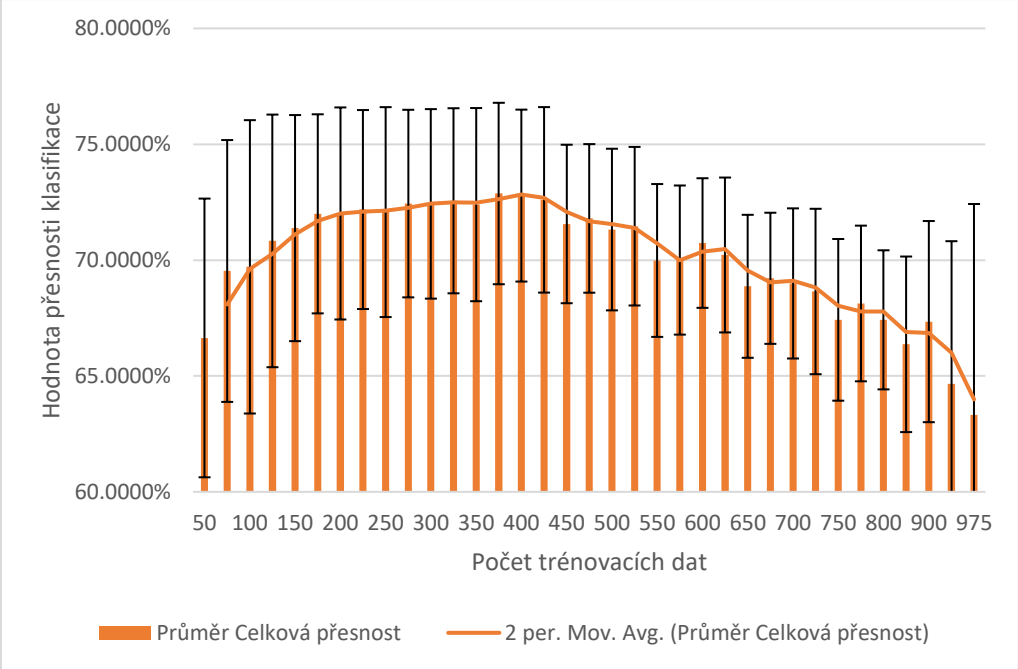
Ke zhodnocení přesnosti aplikovaného klasifikačního algoritmu, který pracoval s trénovacími množinami, byl využit software ArcMap, v němž byla počítána chybová matice. Ta porovná soubor kontrolních množin a následně vypočte celkovou přesnost a Kappa koeficient.

Výsledné hodnoty kappa koeficientu, které lze vidět v grafu 2, se pohybují v rozmezí od 0,50 po hodnotu 0,61. Nejnižší hodnoty Kappa koeficientu bylo dosaženo pro poměr 975 trénovacích bodů, hodnota koeficientu byla 0,50, pro 50 trénovacích bodů byl koeficient 0,53 a pro 25 trénovacích bodů byla již hodnota koeficientu nulová. Naopak nejvyšší hodnoty Kappa koeficientu (0,61) bylo dosaženo pro poměr 375 trénovacích a 625 validačních bodů.



Graf 2: průměr Kappa koeficient klasifikace MLC

Průběh parametru celkové přesnosti klasifikace lze vidět v grafu 3. Hodnoty se pohybují v rozmezí od 63,32 % dosahující maximálně 72,88 %. Nejnižší hodnoty celkové přesnosti bylo dosaženo pro poměr 975 trénovacích a 25 validačních bodů. Pro 50 trénovacích bodů a 950 validačních byla hodnota přesnosti 66,64 % a pro 25 trénovacích bodů a 975 validačních byla již hodnota opět nulová. Naopak nejvyšší hodnoty přinesl poměr 375 trénovacích ku 625 validačním bodům. Stejně jako v grafu 2, kde je zobrazen průměr hodnoty Kappa koeficientu, jsou v grafu celkové přesnosti patrné zlomy. Pokud máme trénovacích bodů méně než 75 a naopak více než 900, je již výsledná hodnota celkové přesnosti klasifikace na velice nízké hodnotě. Z grafu lze vidět, že přesnost klasifikace s větším počtem trénovacích bodů roste až do maxima, kterého výsledky celkové přesnosti dosahují pro 375 trénovacích dat. Následně hodnoty celkové přesnosti klasifikace s narůstajícím počtem trénovacích bodů klesají. Nejvyšší hodnoty směrodatné odchylky, které lze vidět z grafu 3, dosahují hodnoty 0,09 pro klasifikace s 900 a více trénovacími body. Nejnižších hodnot směrodatné odchylky dosahuje klasifikace pro 675 trénovacích a 325 validačních bodů a to hodnoty 0,03. Pro klasifikaci poměru 375 trénovacích a 625 validačních bodů, kde byla celková přesnost nejvyšší, byla hodnota směrodatné odchylky 0,04.



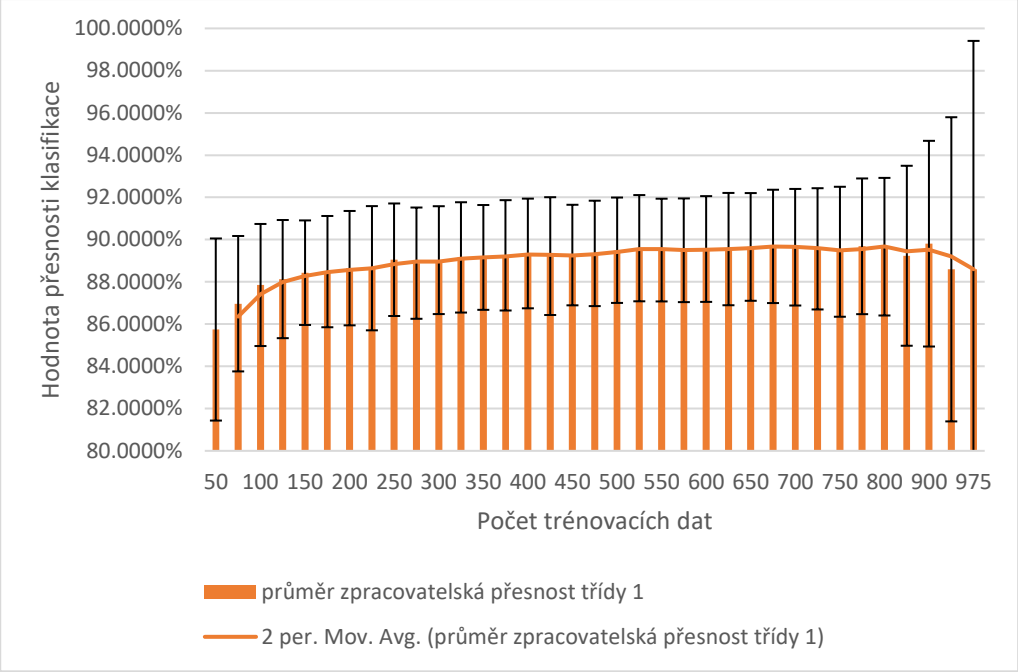
Graf 3: průměr celková přesnost klasifikace MLC

5.1.2 Zpracovatelská a uživatelská přesnost třídy les klasifikace MLC

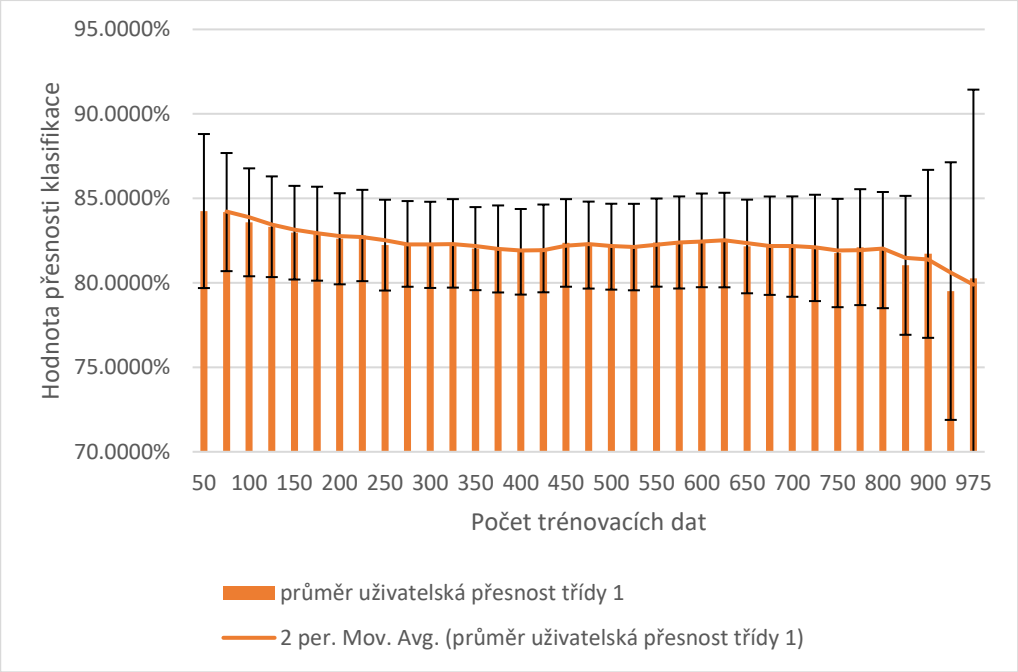
Zpracovatelská přesnost, která se na rozdíl od celkové přesnosti počítá pro každou třídu zvlášť, je poměr mezi správně klasifikovanými pixely a pixely použitými pro testování dané třídy. Průběh hodnocení zpracovatelské přesnosti klasifikace pro třídu les znázorňuje graf 4. Hodnoty se pohybují v rozmezí od 85,74 % dosahující až 89,81 %. Nejnižší hodnoty zpracovatelské přesnosti bylo dosaženo pro poměr 50 trénovacích bodů a 950 validačních bodů. Pro tento poměr bylo v třídě les pro trénování vytvořeno 19 trénovacích bodů a 351 validačních bodů. Naopak nejvyšší hodnoty přinesl poměr 900 trénovacích ku 100 validačním bodům. Stejně jako z grafu 2, kde je zobrazen průměr hodnoty Kappa koeficientu, můžeme vidět zlom v hodnotách. Pro tuto třídu nastává zlom ve zpracovatelské přesnosti pouze pokud máme trénovacích bodů méně jak 75. Pro tuto třídu platí, že pokud se počet trénovacích bodů zvyšuje, zpracovatelská přesnost narůstá. Zpracovatelská přesnost klasifikace pro poměr 375 trénovacích ku 625 validačním bodům byla 89,26 %.

Směrodatné odchylky dosahují nejvyšších hodnot, pokud máme trénovacích bodů více jak 900 a směrodatná odchylka má v tomto případě hodnotu 0,11. Nejnižší směrodatné odchylky bylo dosaženo pro 450 trénovacích bodů. Její hodnota byla 0,02.

Uživatelská přesnost udává, s jakou pravděpodobností pixel zařazený do určité třídy tuto třídu doopravdy představuje. Uživatelskou přesnost klasifikace pro třídu les znázorňuje graf 5. Hodnoty se pohybují v rozmezí od 79,51 % po přesnost 84,25 %. Nejnižší hodnoty uživatelské přesnosti bylo dosaženo pro poměr 950 trénovacích ku 50 validačním bodům. Pro tento poměr byla přesnost klasifikace 79,52 %. Nejvyšších hodnot uživatelské přesnosti pak bylo dosaženo pro poměr 50 trénovacích ku 950 validačním bodům, hodnota přesnosti byla 84,25 %. Zlom v datech uživatelské přesnosti není nijak extrémně viditelný. Trend je však snižující se s narůstajícím počtem trénovacích bodů a následně se trend mírně snižuje s ubývajícím počtem trénovacích dat. Uživatelská přesnost pro poměr 375 trénovacích ku 625 validačním bodům byla 82,01 %. Směrodatné odchylky byly opět nejvyšší pro klasifikace s 900 a více trénovacími body. Směrodatná odchylka je nejvyšší pro 975 trénovacích a 25 validačních bodů a její hodnota dosahuje 0,11.



Graf 4: průměr zpracovatelské přesnosti třídy les klasifikace MLC

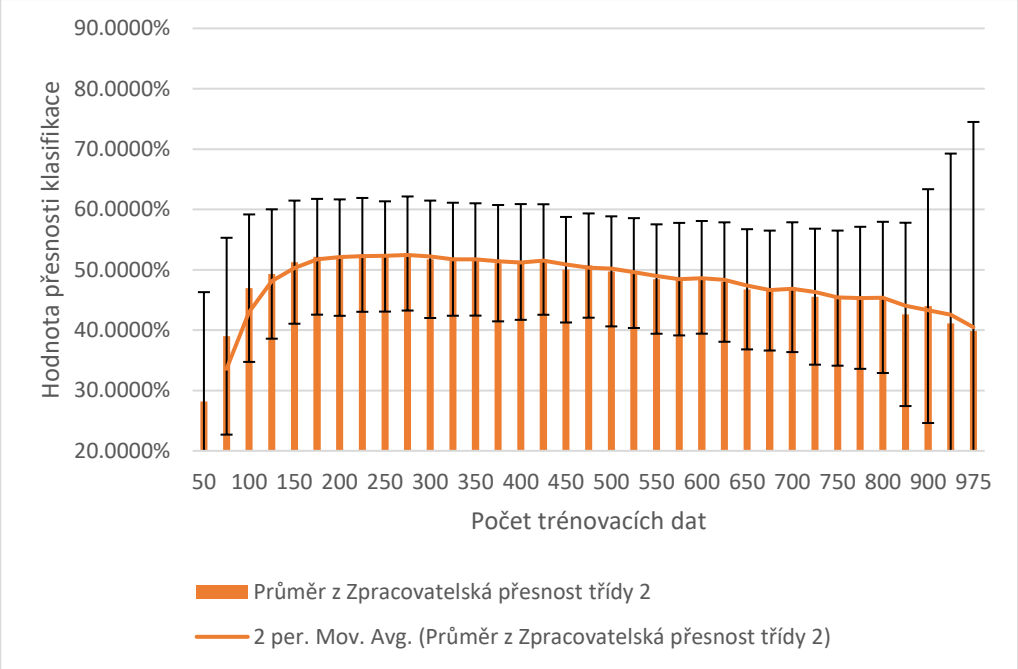


Graf 5: průměr uživatelské přesnosti třídy les klasifikace MLC

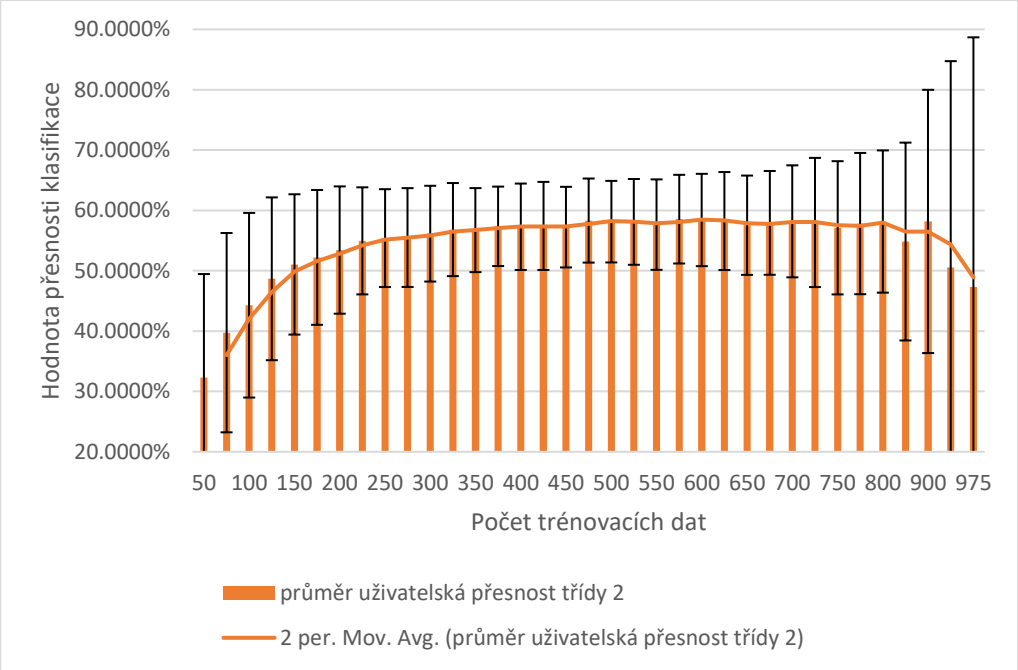
5.1.3 Zpracovatelská a uživatelská přesnost třídy zástavba klasifikace MLC

Zpracovatelskou přesnost klasifikace pro třídu zástavba znázorňuje graf 6. Hodnoty se pohybují v rozmezí od 28,18 % a dosahují až 52,70 %. Nejnižší hodnoty zpracovatelské přesnosti bylo dosaženo pro poměr 50 trénovacích ku 950 validačním bodům. Naopak nevyšší hodnoty přinesl poměr 275 trénovacích ku 725 validačním bodům. Zlom v hodnotách lze vidět pouze v případě kdy máme pro trénování méně jak 100 bodů. Zpracovatelská přesnost klasifikace pro poměr 375 trénovacích ku 625 validačním bodům byla 51,10 % což v porovnání s nejvyšší přesností pro poměr 275 ku 725 je srovnatelná hodnota. Směrodatné odchylky dosahují nejvyšších hodnot, pokud máme trénovacích bodů méně jak 100 a to hodnot až 0,18. Pro více jak 900 trénovacích bodů dosahuje směrodatná odchylka až hodnoty 0,35. Nejnižší směrodatné odchylky je dosaženo pro poměr 475 trénovacích a 525 validačních bodů, odchylka je pro tento poměr 0,09.

Uživatelská přesnost klasifikace pro třídu zástavba je znázorněna v grafu 7. Hodnoty se pohybují v rozmezí od 32,31 % až po přesnost 58,55 %. Nejnižší hodnoty uživatelské přesnosti bylo dosaženo pro poměr 50 trénovacích ku 950 validačním bodům. Nejvyšších hodnot uživatelské přesnosti pak bylo dosaženo pro poměr 575 trénovacích ku 425 validačním bodům. Zlom v datech uživatelské přesnosti lze pozorovat se snižujícím se počtem trénovacích dat. Uživatelská přesnost pro poměr 375 trénovacích ku 625 validačním bodům byla 57,37 %. Směrodatné odchylky byly opět nejvyšší pro klasifikace s 900 a více trénovacími body. Směrodatná odchylka je nejvyšší pro 975 trénovacích a 25 validačních bodů a její hodnota dosahuje 0,41. Nejnižší směrodatné odchylky dosahují výsledky pro poměr 375 trénovacích a 625 validačních bodů, odchylka je pro tento poměr 0.07.



Graf 6: průměr zpracovatelské přesnosti třídy zástavba klasifikace MLC

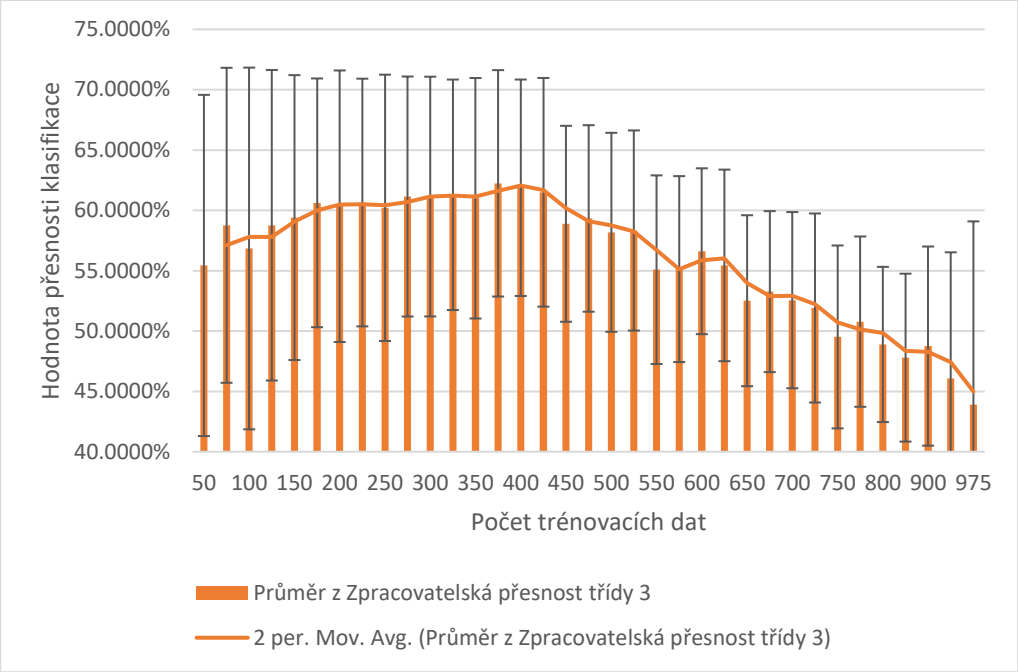


Graf 7:průměr uživatelské přesnosti třídy zástavba klasifikace MLC

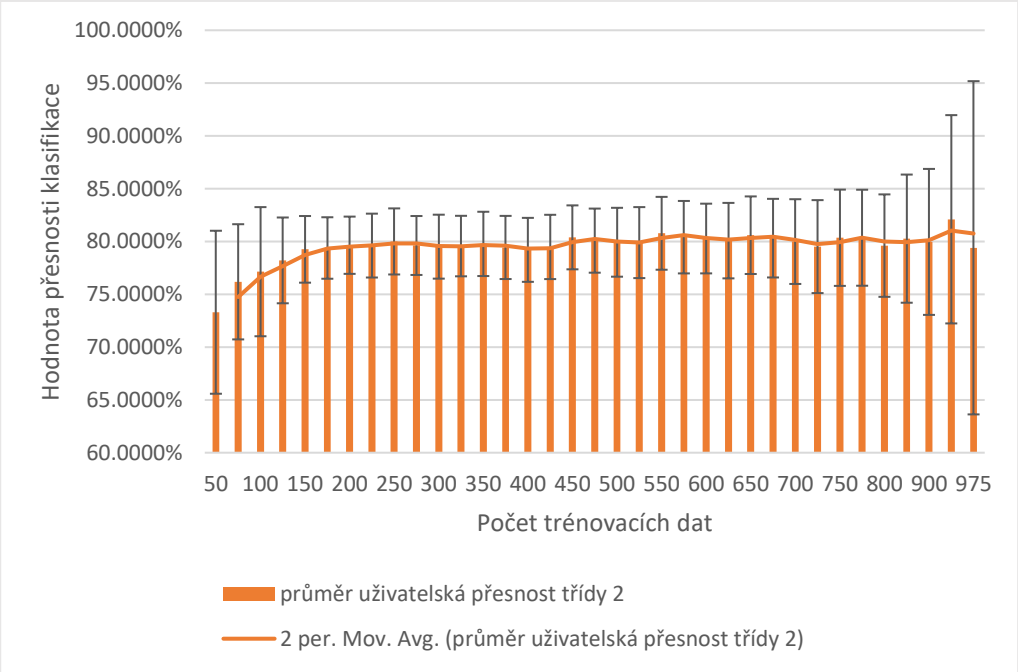
5.1.4 Zpracovatelská a uživatelská přesnost třídy půda s vegetací klasifikace MLC

Zpracovatelskou přesnost klasifikace pro třídu půda s vegetací zachycuje graf 8. Hodnoty se pohybují v rozmezí od 43,91 % a dosahují až 62,24 %. Nejnižší hodnoty zpracovatelské přesnosti bylo dosaženo pro poměr 975 trénovacích ku 25 validačním bodům. Naopak nejvyšší hodnoty přinesl poměr 375 trénovacích ku 625 validačním bodům. Zlom v hodnotách zpracovatelské přesnosti pro třídu půdy s vegetací nastává v případě, kdy je trénovacích dat více jak 800. Obecně zpracovatelská přesnost pro tuto třídu klesá s narůstajícím počtem trénovacích dat. Směrodatné odchylky dosahují nejvyšších hodnot, pokud máme trénovacích bodů méně jak 100 a to hodnot až 0,15. Pro více jak 900 trénovacích bodů dosahuje směrodatná odchylka až hodnoty 0,15. Nejnižší směrodatné odchylky dosahují výsledky pro poměr 800 trénovacích a 200 validačních bodů, odchylka je pro tento poměr 0,06.

Uživatelská přesnost klasifikace pro třídu půda s vegetací je znázorněna v grafu 9. Hodnoty se pohybují v rozmezí od 73,30 % po přesnosti 82,10 %. Nejnižší hodnoty uživatelské přesnosti bylo dosaženo pro poměr 950 trénovacích ku 50 validačním bodům. Nejvyšších hodnot uživatelské přesnosti pak bylo dosaženo pro poměr 50 trénovacích ku 950 validačním bodům. Zlom v datech uživatelské přesnosti lze pozorovat se snižujícím se počtem trénovacích dat. Uživatelská přesnost pro poměr 375 trénovacích ku 625 validačním bodům byla 79,44 %. Směrodatné odchylky byly opět nejvyšší pro klasifikace s 900 a více trénovacími body. Směrodatná odchylka je nejvyšší pro 975 trénovacích a 25 validačních bodů a její hodnota dosahuje 0,16. Nejnižší směrodatné odchylky dosahují výsledky pro poměr 375 trénovacích a 625 validačních bodů, odchylka je pro tento poměr 0,03.



Graf 8: průměr zpracovatelské přesnosti třídy půda s vegetací klasifikace MLC

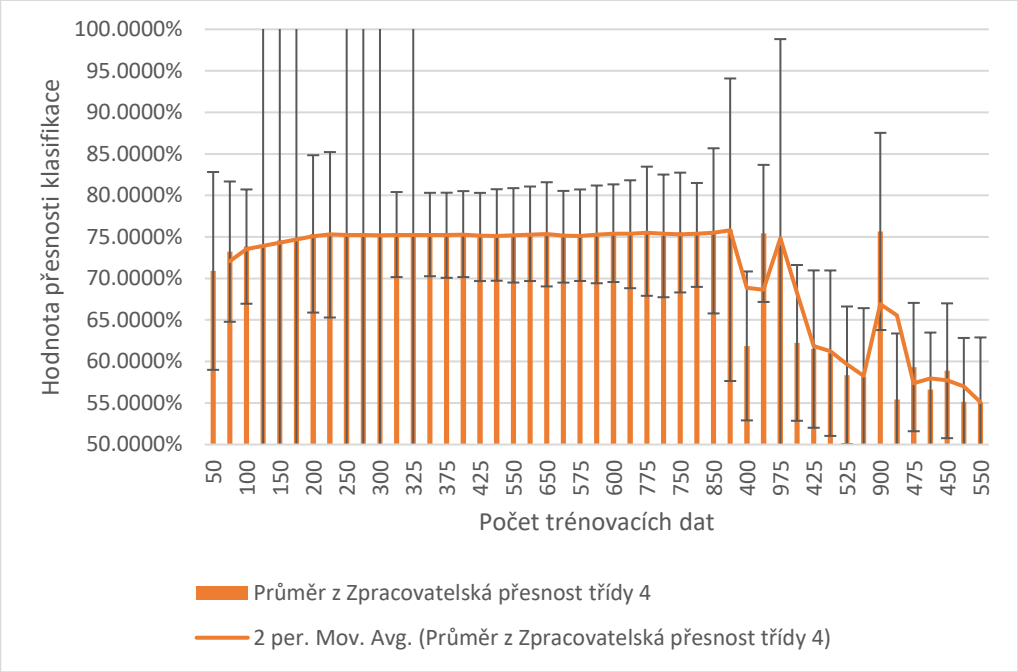


Graf 9: průměr uživatelské přesnosti třídy půda s vegetací klasifikace MLC

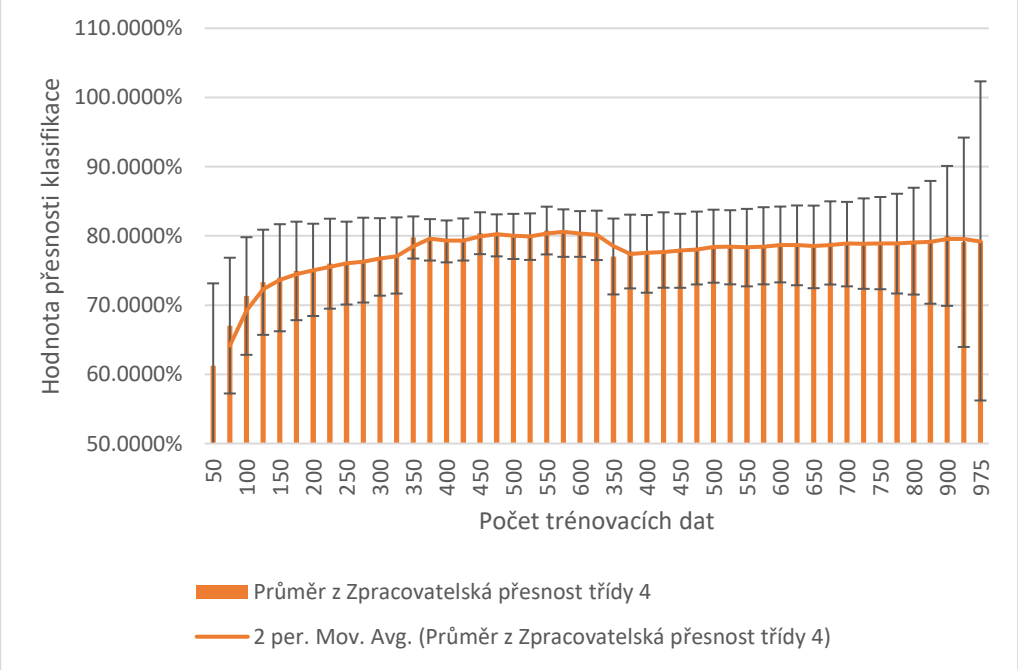
5.1.5 Zpracovatelská a uživatelská přesnost třídy půda bez vegetace klasifikace MLC

Zpracovatelskou přesnost klasifikace pro třídu půda bez vegetace zachycuje graf 10. Hodnoty se pohybují v rozmezí od 55,08 % dosahující až 75,88 %. Nejnižší hodnoty zpracovatelské přesnosti bylo dosaženo pro poměr 550 trénovacích ku 450 validačním bodům. Naopak nejvyšší hodnoty přinesl poměr 950 trénovacích ku 50 validačním bodům. Zlom v hodnotách lze vidět v případě kdy máme pro trénování od 350 do 650 bodů. Směrodatné odchylky dosahují nejvyšších hodnot ze všech tříd. Pokud máme trénovacích bodů 150 dosahuje odchylka hodnoty až 0,92. Pro 375 trénovacích bodů dosahuje směrodatná odchylka hodnoty 0,09. Nejnižší směrodatné odchylky dosahují výsledky pro poměr 400 trénovacích a 600 validačních bodů, odchylka je pro tento poměr 0,05.

Uživatelská přesnost klasifikace pro třídu půda bez vegetace je znázorněna v grafu 11. Hodnoty se pohybují v rozmezí od 61,25 % po přesnosti 80,78 %. Nejnižší hodnoty uživatelské přesnosti bylo dosaženo pro poměr 50 trénovacích ku 950 validačním bodům. Nejvyšších hodnot uživatelské přesnosti pak bylo dosaženo pro poměr 550 trénovacích ku 450 validačním bodům. Zlom v datech uživatelské přesnosti lze pozorovat se snižujícím se počtem trénovacích bodů. Uživatelská přesnost pro poměr 375 trénovacích ku 625 validačním bodům byla 79,44 %. Směrodatné odchylky byly opět nejvyšší pro klasifikace s 900 a více trénovacími body. Směrodatná odchylka je nejvyšší pro 975 trénovacích a 25 validačních bodů a její hodnota dosahuje 0,23. Nejnižší směrodatné odchylky dosahují výsledky pro poměr 375 trénovacích a 625 validačních bodů, odchylka je pro tento poměr 0,03.



Graf 10: průměr zpracovatelské přesnosti třídy půda bez vegetace klasifikace MLC

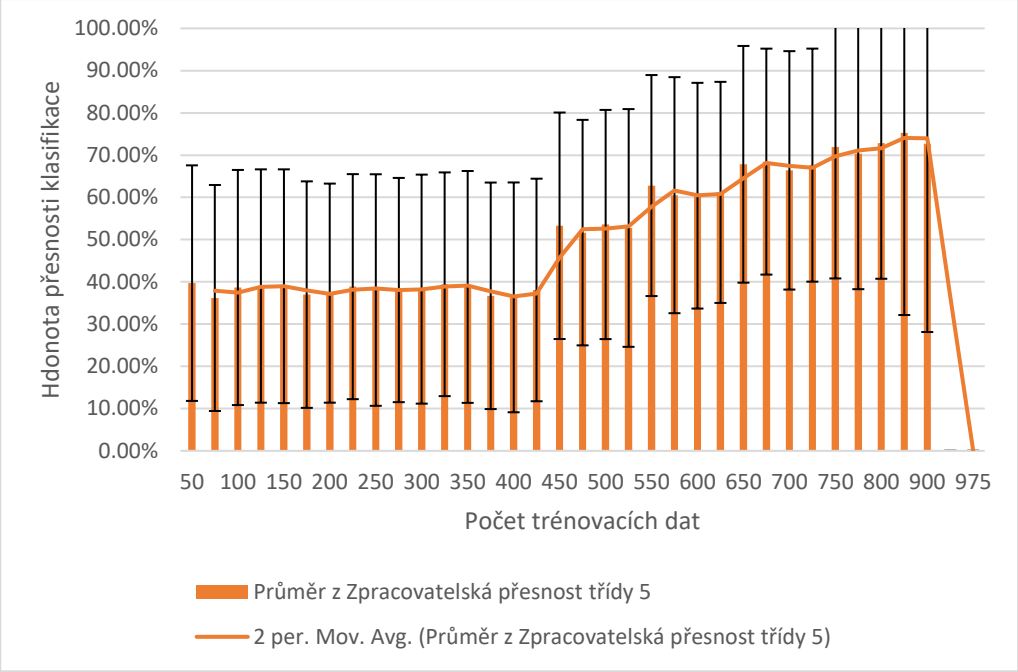


Graf 11: průměr uživatelské přesnosti třídy půda bez vegetace klasifikace MLC

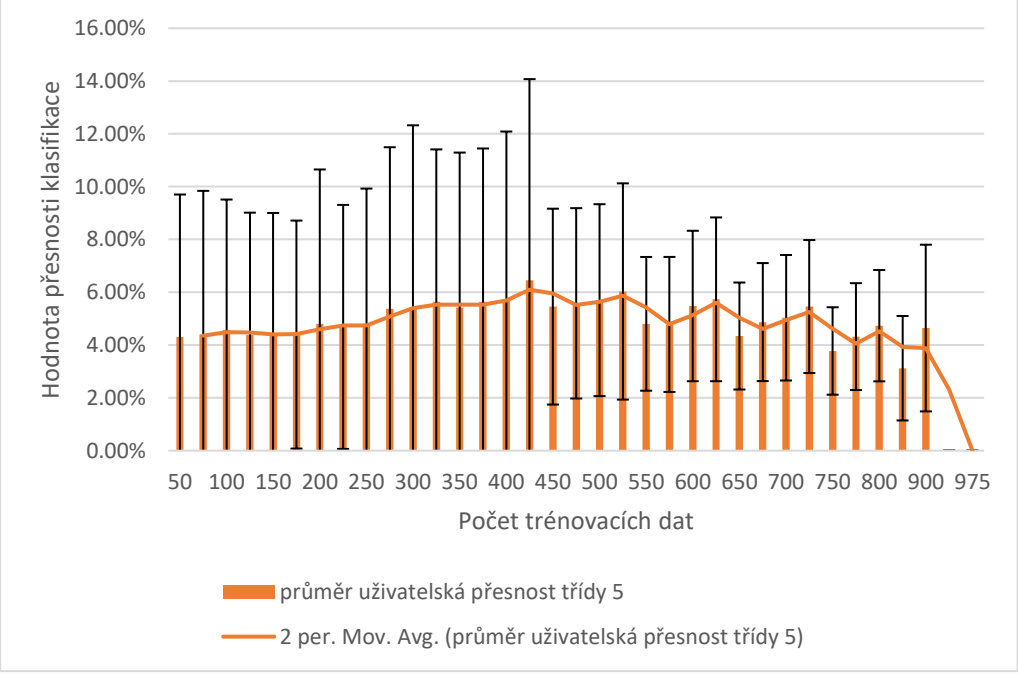
5.1.6 Zpracovatelská a uživatelská přesnost třídy vodní plocha klasifikace MLC

Zpracovatelská přesnost klasifikace pro třídu vodní plocha je v grafu 12. Hodnoty se pohybují v rozmezí od 36,18 % dosahující až 75,30 %. Nejnižší hodnoty zpracovatelské přesnosti bylo dosaženo pro poměr 75 trénovacích ku 925 validačním bodům. Pro tento poměr byla hodnota přesnosti 36.1833 %. Naopak nejvyšší hodnoty přinesl poměr 850 trénovacích ku 150 validačním bodům. Trend v hodnotách zpracovatelské přesnosti třídy vodní plocha lze pozorovat narůstající s vyšším množstvím trénovacích dat. Nejvyšších hodnot dosahují přesnosti pro 850 a 900 trénovacích bodů. Následně pokud máme více jak 925 trénovacích bodů je přesnost klasifikace nulová, z důvodu velice malého počtu validačních dat. V případě, kdy máme trénovacích bodů 75 dosahuje směrodatná odchylka nejnižší hodnoty a to 0,27. Pro 375 trénovacích bodů dosahuje směrodatná odchylka hodnoty 0,27. Nejvyšší směrodatné odchylky dosahují výsledky pro poměr 900 trénovacích a 100 validačních bodů, odchylka je pro tento poměr 0,45.

Uživatelská přesnost klasifikace pro třídu vodní plocha je v grafu 13. Hodnoty se pohybují v rozmezí od 3,12 % po přesnosti 6,46 %. Nejnižších hodnoty uživatelské přesnosti bylo dosaženo pro poměr 850 trénovacích ku 150 validačním bodům. Nejvyšších hodnot uživatelské přesnosti pak bylo dosaženo pro poměr 425 trénovacích ku 575 validačním bodům. V hodnotách uživatelské přesnosti pro třídu vodní plocha lze vidět trend lehce kolísavý narůstající s počtem trénovacích dat, avšak zlom v datech nastává v případě, kdy máme více jak 900 trénovacích bodů, zde uživatelská přesnost klasifikace vychází 0. Uživatelská přesnost pro poměr 375 trénovacích ku 625 validačním bodům byla 5,64 %. Směrodatné odchylky byly nejvyšší pro klasifikace s 50 až 425 trénovacími body. Směrodatná odchylka je nejvyšší pro 425 trénovacích a 25 validačních bodů a její hodnota dosahuje 0,08. Nejnižší směrodatné odchylky dosahují výsledky pro poměr 750 trénovacích a 250 validačních bodů, odchylka je pro tento poměr 0,02.



Graf 12: průměr zpracovatelské přesnosti třídy vodní plocha klasifikace MLC



Graf 13: průměr uživatelská přesnost třídy vodní plocha klasifikace MLC

Tabulka 6: Shrnutí výsledků hodnocení přesnosti pro jednotlivé třídy klasifikace MLC

Třída	Nejvyšší dosažená zpracovatelská přesnost	Poměr trénovacích a validačních dat pro nejvyšší zpracovatelskou přesnost	Nejvyšší dosažená uživatelská přesnost	Poměr trénovacích a validační dat pro nejvyšší uživatelskou přesnost	Nejnižší dosažená zpracovatelská přesnost	Poměr trénovacích a validační dat pro nejnižší zpracovatelskou přesnost	Nejnižší dosažená uživatelská přesnost	Poměr trénovacích a validační dat pro nejnižší uživatelskou přesnost	Zpracovatelská přesnost pro poměr 375 trénovacích a 625 validačních bodů	Uživatelská přesnost pro poměr 375 trénovacích a 625 validačních bodů
lesní plocha	89.81 %	900 / 100	84.25 %	50 / 950	85.74 %	50 / 950	79.51 %	950 / 50	89.26 %	82.01 %
zástavba	52.70 %	275 / 725	58.55 %	575 / 425	28.18 %	50 / 950	32.31 %	50 / 950	51.10 %	57.37 %
půda s vegetací	62.24 %	375 / 625	82.10 %	950 / 50	43.91 %	975 / 25	73.30 %	50 / 950	62.24 %	79.44 %
orná půda	75.87 %	950 / 50	80.78 %	550 / 450	55.08 %	550 / 450	61.25 %	50 / 950	62.24 %	79.44 %
vodní plocha	75.30 %	850 / 150	6.46 %	425 / 575	0 %	950 / 50	0 %	950 / 50	36.70 %	5.64 %

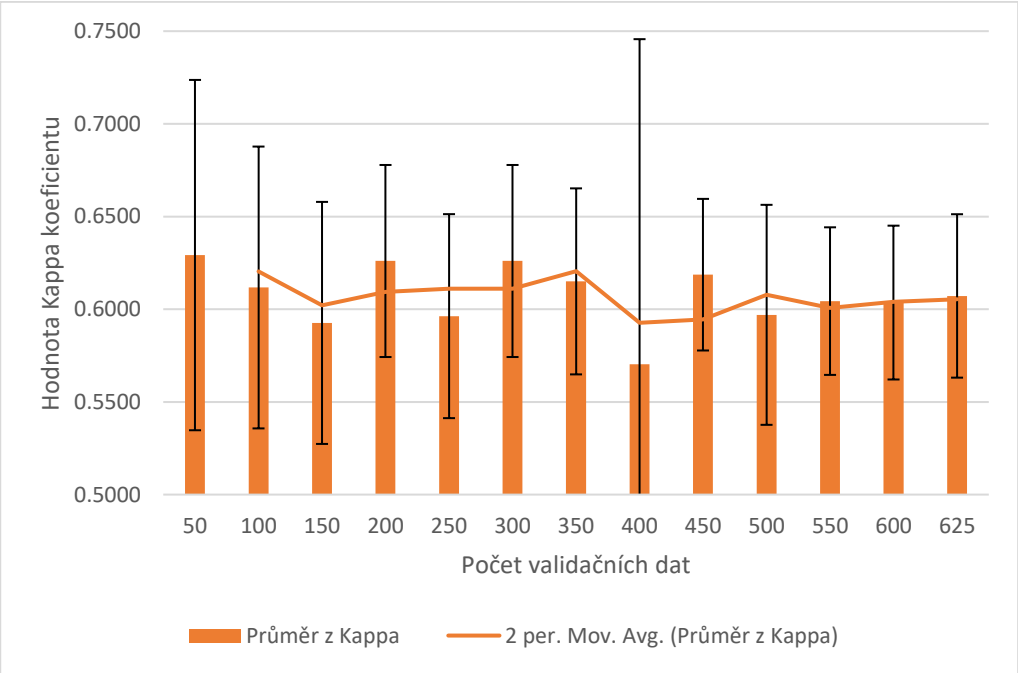
Tabulka 6 shrnuje veškeré výsledky testování vlivu podílu/množství trénovacích a validačních dat na přesnost klasifikace metodou Maximum Likelihood. U poměru 375 trénovacích ku 625 validačním bodům jsou všechny přesnosti jak uživatelské, tak i zpracovatelské dosažené pro jednotlivé třídy nižší než nejvyšší dosažené (s jedinou výjimkou). Zároveň se však ukazuje, že pro kategorie, které byly dobře klasifikovány (s vysokou uživatelskou a zpracovatelskou přesností – např. lesní ploch, půda s vegetací, půda bez vegetace) jsou přesnosti dosažené pro poměr 375 trénovacích a 625 validačních bodů pouze o málo nižší než hodnoty nejvyšší dosažené uživatelské či zpracovatelské přesnosti. Naopak je z tabulky zřejmé, že pro kategorie, které byly klasifikovány s horší přesností (např. vodní plochy, zástavba), je v případě poměru 375 trénovacích a 625 validačních bodů přesnost znatelně nižší.

5.2 Testování vlivu změny velikosti validačního datasetu na stabilitu výsledku hodnocení přesnosti klasifikace metodou Maximum Likelihood

Na základě výsledků pokusů s podílem trénovacích a validačních dat, byl určen poměr s nevyšší přesností. Poměr, který dosahoval nejvyšších hodnot přesnosti klasifikace byl 375 trénovacích ku 625 validačním bodům. Počet 375 trénovacích bodů zůstal v dalším testu zachován. Měněn byl počet validačních dat a to od 50 bodů do 625 s cílem vyhodnotit, jaký je vliv změny velikosti validačního datasetu na stabilitu výsledku hodnocení celkové přesnosti klasifikace. Výsledky sledovaných parametrů se nachází v následujících kapitolách a jsou pro názornost prezentovány v grafech. V grafech je nyní na ose x uveden počet validačních bodů.

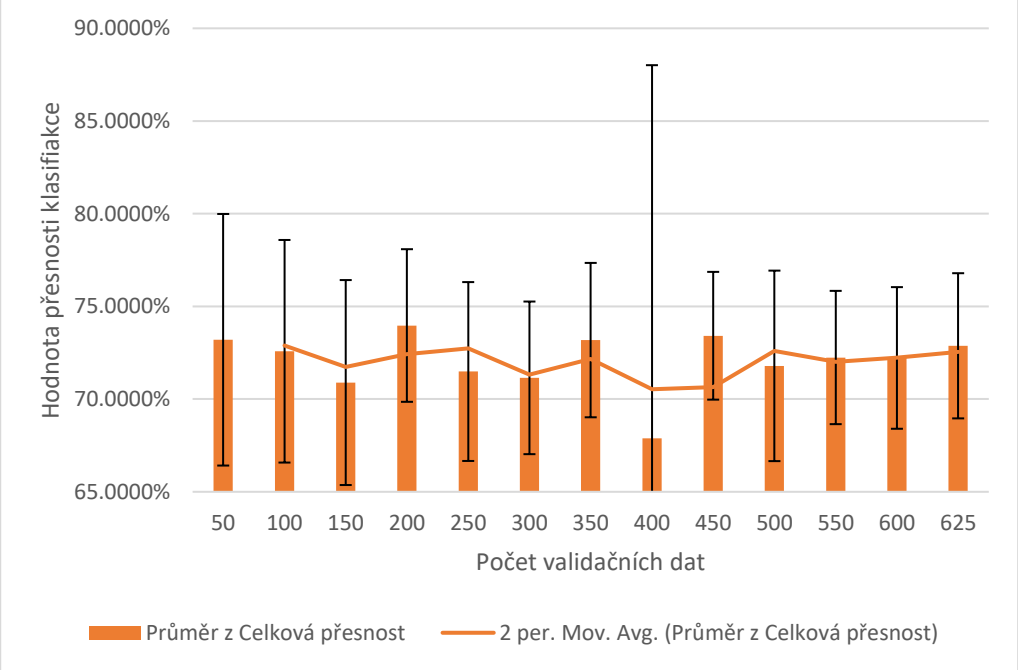
5.2.1 Kappa koeficient a celková přesnost při změně velikosti validačního datasetu klasifikace MLC

Výsledné hodnoty kappa koeficientu, které lze vidět v grafu 14, se pohybují v rozmezí od 0,57 po hodnotu 0,63. Rozdíly jsou tedy relativně malé. Nejnižší hodnoty Kappa koeficientu bylo dosaženo pro poměr 375 trénovacích bodů ku 400 validačním bodům, hodnota koeficientu byla 0,57. Pro 375 trénovacích bodů ku 150 validačním bodům byl koeficient 0.59 a pro 25 trénovacích bodů byla již hodnota koeficientu nulová. Naopak nejvyšší hodnoty Kappa koeficientu (0,63) bylo dosaženo pro poměr 375 trénovacích a 50 validačních bodů. V grafu lze vidět zlom – v případě, kdy máme validačních bodů téměř stejně jako trénovacích, je již výsledná hodnota Kappa koeficientu velice nízká. Nejvyšší hodnoty směrodatné odchylky, které lze vidět z grafu 14, dosahují klasifikace pro poměr 375 trénovacích ku 400 validačním bodům. Hodnota směrodatné odchylky je pro tento počet validačních bodů 0.18.



Graf 14: průměr Kappa koeficient při změně velikosti validačního datasetu klasifikace MLC

Průběh parametru celkové přesnosti klasifikace lze vidět v grafu 15. Hodnoty se pohybují v rozmezí od 67,88 % dosahující maximálně 73,97 %. Rozdíl je tedy více než 6 %. Nejnižší hodnoty celkové přesnosti bylo dosaženo pro poměr 375 trénovacích a 400 validačních bodů. Pro 375 trénovacích bodů a 150 validačních byla hodnota přesnosti 70,89 %. Naopak nejvyšší hodnoty přinesl poměr 375 trénovacích ku 200 validačním bodům. Stejně jako v grafu 14, kde je zobrazen průměr hodnoty Kappa koeficientu, je v grafu celkové přesnosti patrný zlom. Pokud máme validačních bodů méně než 150, je již výsledná hodnota celkové přesnosti klasifikace na velice nízké hodnotě. Z grafu lze vidět, že přesnost klasifikace s větším počtem validačních bodů roste až do maxima, kterého výsledky celkové přesnosti dosahují. Nejvyšší hodnoty směrodatné odchylky, které lze vidět z grafu 15, dosahují hodnoty 0,20 pro klasifikace s 400 validačními body. Nejnižších hodnot směrodatné odchylky dosahuje klasifikace pro 375 trénovacích a 450 validačních bodů a to hodnoty 0,03. Pro klasifikaci poměru 375 trénovacích a 625 validačních bodů, kde byla celková přesnost nejvyšší, byla hodnota směrodatné odchylky 0,04.

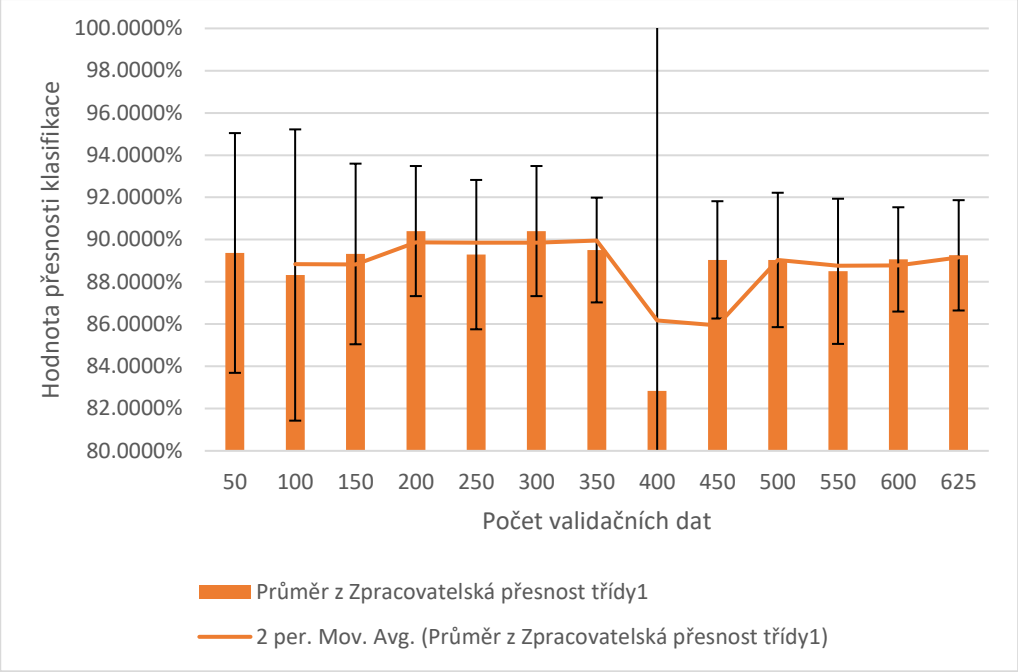


Graf 15: průměr Celková přesnost při změně velikosti validačního datasetu klasifikace MLC

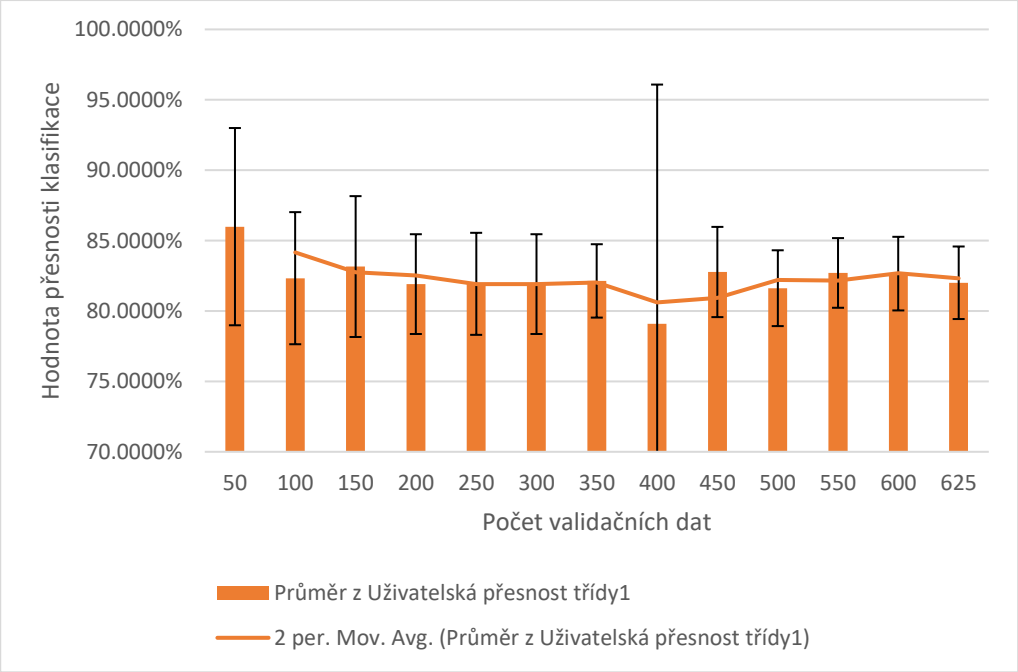
5.2.2 Zpracovatelská a uživatelská přesnost třídy les při změně velikosti validačního datasetu klasifikace MLC

Průběh hodnocení zpracovatelské přesnosti klasifikace pro třídu les znázorňuje graf 16. Hodnoty se pohybují v rozmezí od 82,84 % dosahující až 90,41 %. Nejnižší hodnoty zpracovatelské přesnosti bylo dosaženo pro poměr 375 trénovacích bodů a 400 validačních bodů. Naopak nejvyšší hodnoty přinesl poměr 375 trénovacích ku 300 validačním bodům. Zpracovatelská přesnost klasifikace pro poměr 375 trénovacích ku 625 validačním bodům byla 89,26 %. Směrodatné odchylky dosahují nejvyšších hodnot, pokud máme 375 trénovacích bodů a 400 validačních bodů, směrodatná odchylka má v tomto případě hodnotu 0,24. Nejnižší směrodatné odchylky bylo dosaženo pro 600 validačních bodů. Její hodnota byla 0,02.

Uživatelskou přesnost klasifikace pro třídu les znázorňuje graf 17. Hodnoty se pohybují v rozmezí od 79,08 % po přesnost 85,99 %. Nejnižší hodnoty uživatelské přesnosti bylo dosaženo pro poměr 375 trénovacích ku 500 validačním bodům. Nejvyšších hodnot uživatelské přesnosti pak bylo dosaženo pro poměr 375 trénovacích ku 50 validačním bodům. Zlom v datech uživatelské přesnosti není nijak extrémně viditelný. Trend je však snižující se s narůstajícím počtem validačních bodů. Uživatelská přesnost pro poměr 375 trénovacích ku 625 validačním bodům byla 82,01 %. Směrodatné odchylky byly nejvyšší pro klasifikace s 400 validačními body. Směrodatná odchylka byla nejnižší pro klasifikace s 550 validačními body a její hodnota byla 0,03.



Graf 16: průměr zpracovatelské přesnosti třídy les při změně velikosti validačního datasetu klasifikace MLC

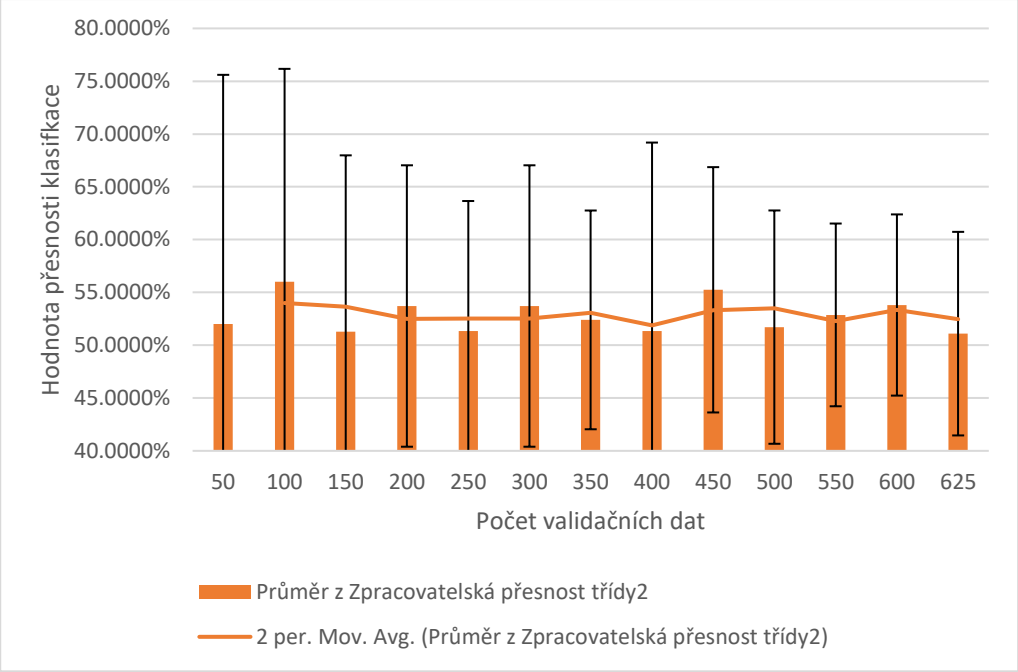


Graf 17: průměr uživatelské přesnosti třídy les při změně velikosti validačního datasetu klasifikace MLC

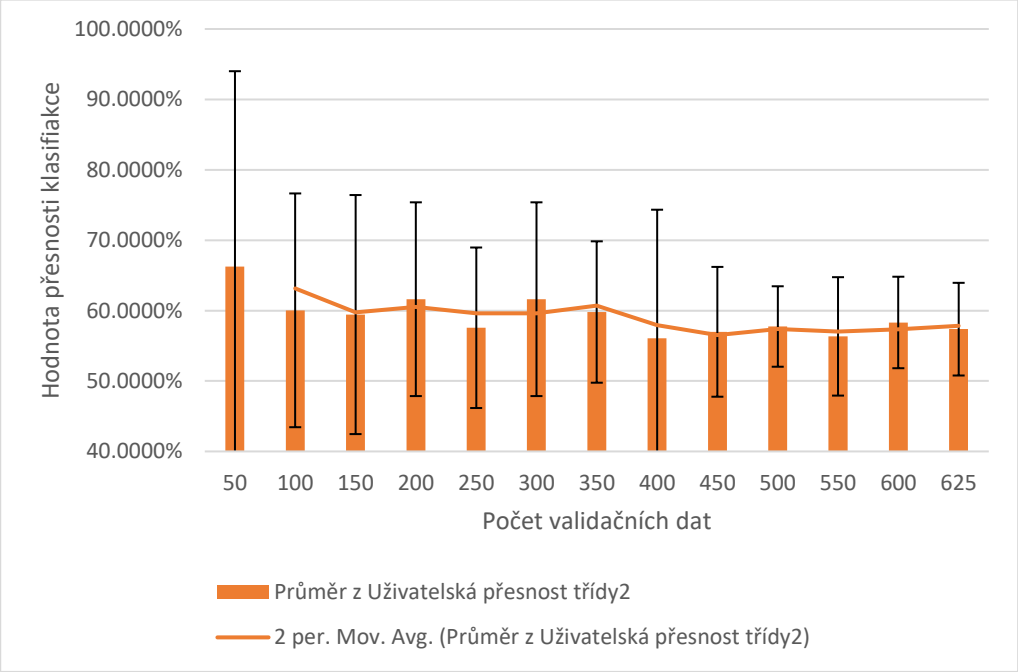
5.2.3 Zpracovatelská a uživatelská přesnost třídy zástavba při změně velikosti validačního datasetu klasifikace MLC

Zpracovatelskou přesnost klasifikace pro třídu zástavba znázorňuje graf 18. Hodnoty se pohybují v rozmezí od 51,10 % a dosahují až 56,00 %. Nejnižší hodnoty zpracovatelské přesnosti bylo dosaženo pro poměr 375 trénovacích ku 625 validačním bodům. Naopak nevyšší hodnoty přinesl poměr 375 trénovacích ku 100 validačním bodům. Pro tento poměr byla hodnota celkové přesnosti 56,00 %. Zpracovatelská přesnost pro poměr 375 trénovacích ku 625 validačním bodům byla 51,10 %. Směrodatné odchylky dosahují nejvyšších hodnot, pokud máme trénovacích bodů méně jak 100 a to hodnot až 0,20. Pro 50 validačních bodů dosahuje směrodatná odchylka až hodnoty 0,24. Nejnižší směrodatné odchylky je dosaženo pro poměr 375 trénovacích a 600 validačních bodů, odchylka je pro tento poměr 0,09.

Uživatelská přesnost klasifikace pro třídu zástavba je znázorněna v grafu 19. Hodnoty se pohybují v rozmezí od 56,06 % až po přesnost 66,26 %. Nejnižší hodnoty uživatelské přesnosti bylo dosaženo pro poměr 375 trénovacích ku 400 validačním bodům. Nejvyšších hodnot uživatelské přesnosti pak bylo dosaženo pro poměr 375 trénovacích ku 50 validačním bodům, hodnota přesnosti byla 66.26 %. Uživatelská přesnost pro poměr 375 trénovacích ku 625 validačním bodům byla 57,37 %. Směrodatné odchylky byly opět nejvyšší pro klasifikace s 900 a více trénovacími body. Směrodatná odchylka je nejvyšší pro 375 trénovacích a 50 validačních bodů a její hodnota dosahuje 0,28. Nejnižší směrodatné odchylky dosahují výsledky pro poměr 375 trénovacích a 500 validačních bodů, odchylka je pro tento poměr 0,06.



Graf 18: průměr zpracovatelské přesnosti třídy zástavba při změně velikosti validačního datasetu klasifikace MLC

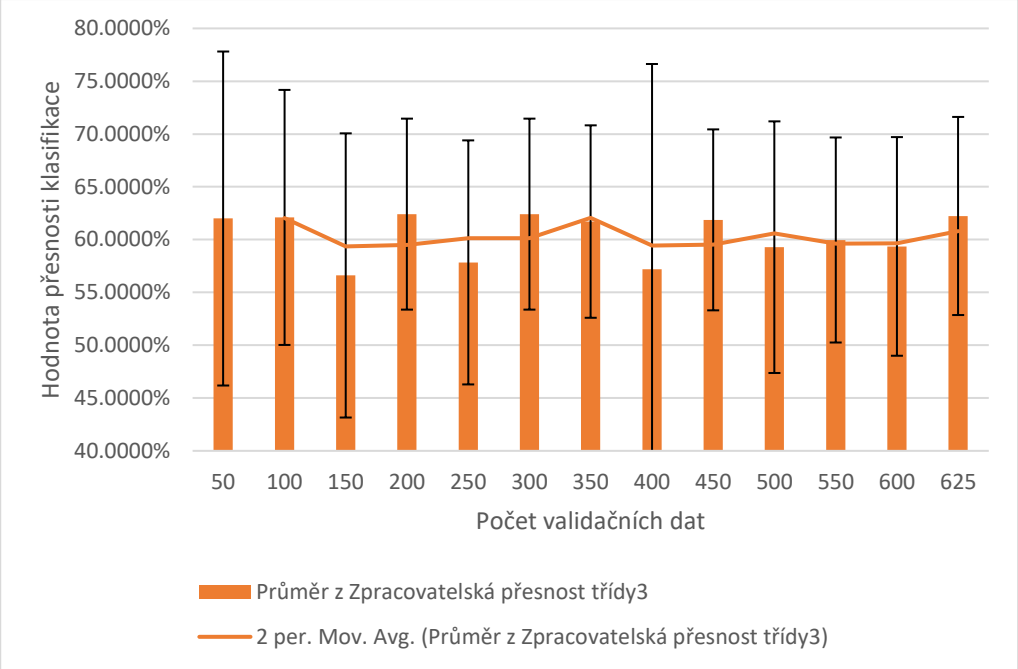


Graf 19: průměr uživatelské přesnosti třídy zástavba při změně velikosti validačního datasetu klasifikace MLC

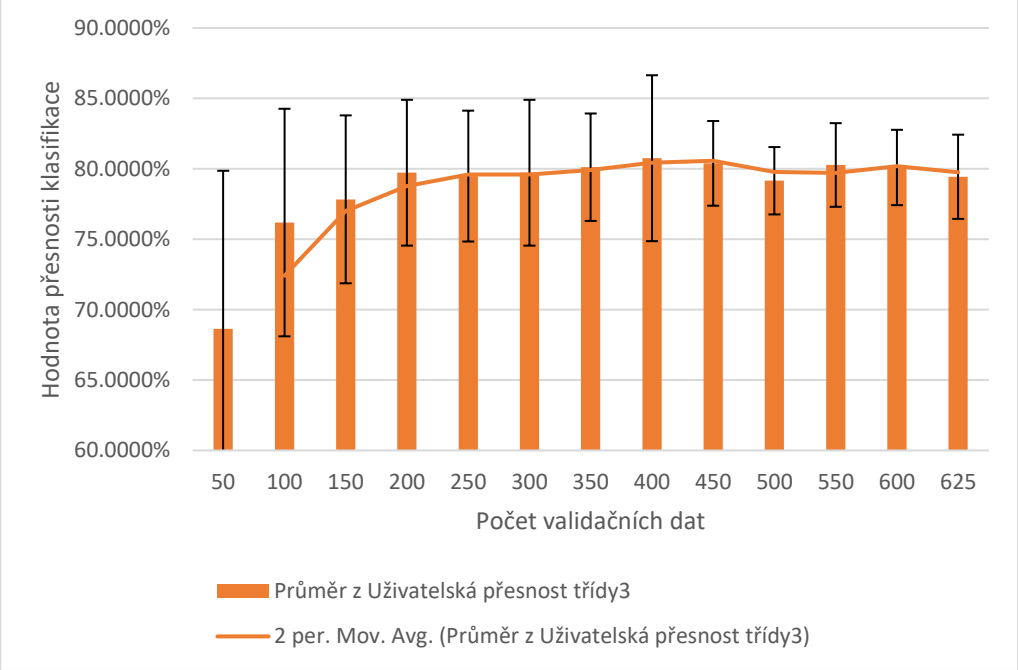
5.2.4 Zpracovatelská a uživatelská přesnost třídy půda s vegetací při změně velikosti validačního datasetu klasifikace MLC

Zpracovatelskou přesnost klasifikace pro třídu půda s vegetací zachycuje graf 20. Hodnoty se pohybují v rozmezí od 56,61 % a dosahují až 62,42 %. Nejnižší hodnoty zpracovatelské přesnosti bylo dosaženo pro poměr 375 trénovacích ku 150 validačním bodům. Naopak nejvyšší hodnoty přinesl poměr 375 trénovacích ku 300 validačním bodům. Pro tento poměr byla hodnota celkové přesnosti 62,42 %. Zpracovatelská přesnost třídy půda s vegetací pro 375 trénovacích a 625 validačních bodů byla 62,24 %. Směrodatná odchylka dosahuje nejvyšší hodnoty, pokud máme validačních bodů 400 a to hodnoty 0,19. Nejnižší směrodatné odchylky dosahuje výsledek pro poměr 375 trénovacích a 450 validačních bodů, odchylka je pro tento poměr 0,09. Směrodatná odchylka zpracovatelské přesnosti třídy půda s vegetací pro 375 trénovacích a 625 validačních bodů byla 0,09.

Uživatelská přesnost klasifikace pro třídu půda s vegetací je znázorněna v grafu 21. Hodnoty se pohybují v rozmezí od 68,63 % po přesnosti 80,75 %. Nejnižší hodnoty uživatelské přesnosti bylo dosaženo pro poměr 375 trénovacích ku 50 validačním bodům. Nejvyšších hodnot uživatelské přesnosti pak bylo dosaženo pro poměr 375 trénovacích ku 400 validačním bodům, hodnota přesnosti byla 80,75 %. Uživatelská přesnost třídy půda s vegetací pro 375 trénovacích a 625 validačních bodů byla 79,44 %. Směrodatné odchylky byly opět nejvyšší pro klasifikace s 900 a více trénovacími body. Směrodatná odchylka dosahuje nejvyšší hodnoty, pokud máme validačních bodů 50 a to hodnoty 0,11. Nejnižší směrodatné odchylky dosahuje výsledek pro poměr 375 trénovacích a 500 validačních bodů, odchylka je pro tento poměr 0,02. Směrodatná odchylka uživatelské přesnosti třídy půda s vegetací pro 375 trénovacích a 625 validačních bodů byla 0,03.



Graf 20: průměr zpracovatelské přesnosti třídy půda s vegetací při změně velikosti validačního datasetu klasifikace MLC

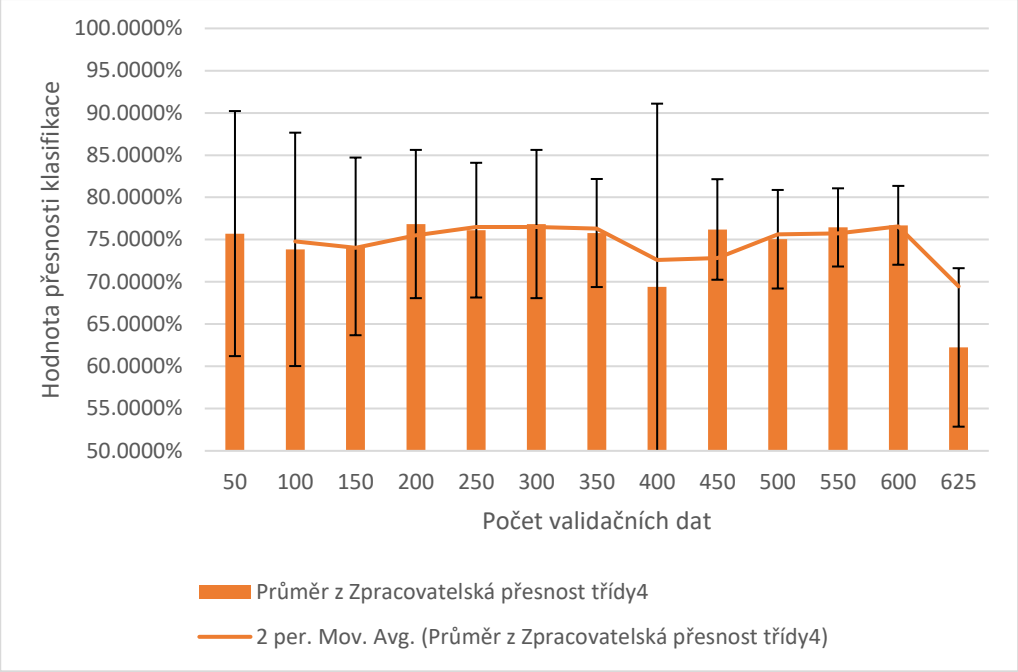


Graf 21: průměr uživatelské přesnosti třídy půda s vegetací při změně velikosti validačního datasetu klasifikace MLC

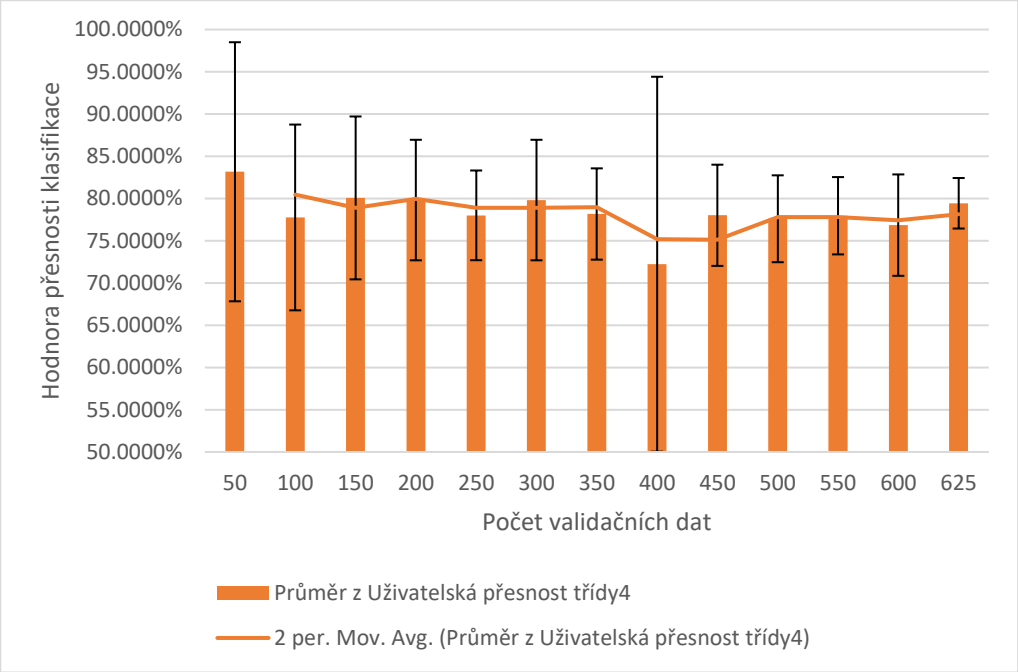
5.2.5 Zpracovatelská a uživatelská přesnost třídy půda bez vegetace při změně velikosti validačního datasetu klasifikace MLC

Zpracovatelskou přesnost klasifikace pro třídu půda bez vegetace zachycuje graf 22. Hodnoty se pohybují v rozmezí od 62,24 % dosahující až 76,85 %. Nejnižší hodnoty zpracovatelské přesnosti bylo dosaženo pro poměr 375 trénovacích ku 625 validačním bodům. Naopak nejvyšší hodnoty přinesl poměr 375 trénovacích ku 300 validačním bodům. Pro tento poměr byla hodnota celkové přesnosti 76,85 %. Směrodatná odchylka dosahuje nejvyšší hodnoty, pokud máme validačních bodů 400 a to hodnoty 0,22. Nejnižší směrodatné odchylky dosahuje výsledek pro poměr 375 trénovacích a 550 validačních bodů, odchylka je pro tento poměr 0,05. Směrodatná odchylka zpracovatelské přesnosti třídy půda s vegetací pro 375 trénovacích a 625 validačních bodů byla 0,09.

Uživatelská přesnost klasifikace pro třídu půda bez vegetace je znázorněna v grafu 23. Hodnoty se pohybují v rozmezí od 72,23 % po přesnosti 83,17 %. Nejnižší hodnoty uživatelské přesnosti bylo dosaženo pro poměr 375 trénovacích ku 50 validačním bodům. Nejvyšších hodnot uživatelské přesnosti pak bylo dosaženo pro poměr 375 trénovacích ku 625 validačním bodům, hodnota přesnosti byla 83,17 %. Uživatelská přesnost pro poměr 375 trénovacích ku 625 validačním bodům byla 79,44 %. Směrodatná odchylka dosahuje nejvyšší hodnoty, pokud máme validačních bodů 400 a to hodnoty 0,22. Nejnižší směrodatné odchylky dosahuje výsledek pro poměr 375 trénovacích a 625 validačních bodů, odchylka je pro tento poměr 0,03.



Graf 22: průměr zpracovatelské přesnosti třídy půda bez vegetace při změně velikosti validačního datasetu klasifikace MLC

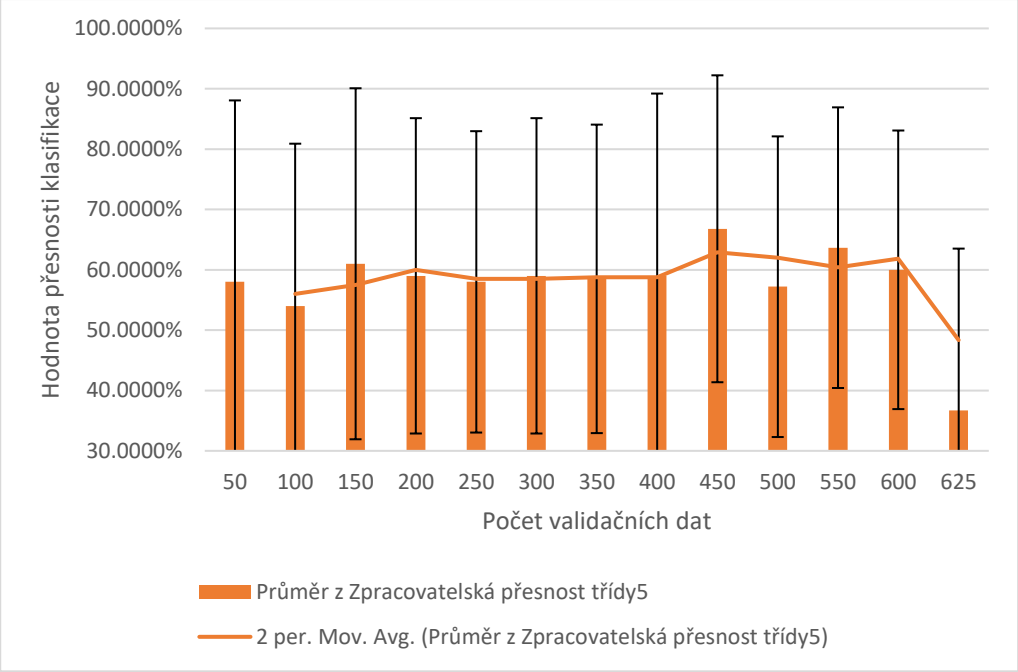


Graf 23: průměr uživatelské přesnosti třídy půda bez vegetace při změně velikosti validačního datasetu klasifikace MLC

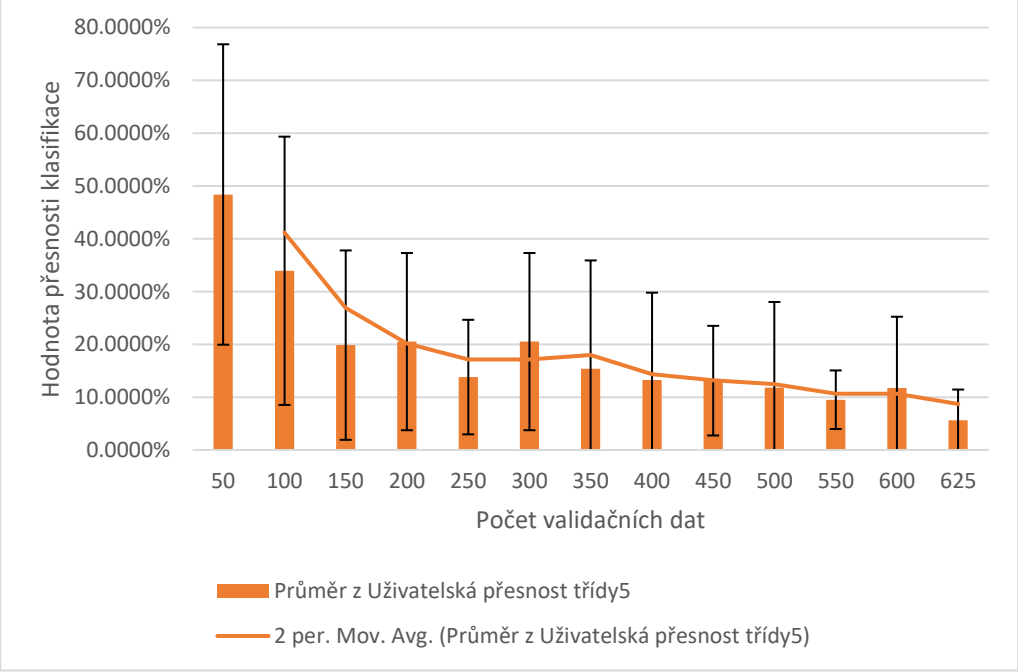
5.2.6 Zpracovatelská a uživatelská přesnost třídy vodní plocha při změně velikosti validačního datasetu klasifikace MLC

Zpracovatelská přesnost klasifikace pro třídu vodní plocha je v grafu 24. Hodnoty se pohybují v rozmezí od 36,70 % dosahující až 66,80 %. Nejnižší hodnoty zpracovatelské přesnosti bylo dosaženo pro poměr 375 trénovacích ku 625 validačním bodům. Naopak nejvyšší hodnoty přinesl poměr 375 trénovacích ku 450 validačním bodům. Pro tento poměr byla hodnota celkové přesnosti 66,80 %. Směrodatná odchylka dosahuje nejvyšší hodnoty, pokud máme validačních bodů 400 a to hodnoty 0,30. Nejnižší směrodatné odchylky dosahuje výsledek pro poměr 375 trénovacích a 600 validačních bodů, odchylka je pro tento poměr 0,23. Směrodatná odchylka zpracovatelské přesnosti třídy půda s vegetací pro 375 trénovacích a 625 validačních bodů byla 0,27.

Uživatelská přesnost klasifikace pro třídu vodní plocha je v grafu 25. Hodnoty se pohybují v rozmezí od 5,64 % po přesnosti 48,38 %. Nejnižších hodnoty uživatelské přesnosti bylo dosaženo pro poměr 375 trénovacích ku 625 validačním bodům. Nejvyšších hodnot uživatelské přesnosti pak bylo dosaženo pro poměr 375 trénovacích ku 50 validačním bodům. Směrodatná odchylka dosahuje nejvyšší hodnoty, pokud máme validačních bodů 50 a to hodnoty 0,29. Nejnižší směrodatné odchylky dosahuje výsledek pro poměr 375 trénovacích a 550 validačních bodů, odchylka je pro tento poměr 0,06. Směrodatná odchylka zpracovatelské přesnosti třídy půda s vegetací pro 375 trénovacích a 625 validačních bodů byla 0,58.



Graf 24: průměr zpracovatelské přesnosti třídy půda vodní plocha při změně velikosti validačního datasetu klasifikace MLC



Graf 25: průměr uživatelské přesnosti třídy půda vodní plocha při změně velikosti validačního datasetu klasifikace MLC

Tabulka 7: Shrnutí výsledků hodnocení přesnosti pro jednotlivé třídy při změně velikosti validačního datasetu klasifikace MLC

Třída	Nejvyšší dosažená zpracovatelská přesnost	Počet validačních bodů pro nejvyšší zpracovatelskou	Nejvyšší dosažená uživatelská přesnost	Počet validačních bodů pro nejvyšší uživatelsko u přesnost	Nejnižší dosažená zpracovatelská přesnost	Počet validačních bodů pro nejnižší zpracovatelskou	Nejnižší dosažená uživatelská přesnost	Počet validačních bodů pro nejnižší uživatelsko u přesnost	Zpracovatelská přesnost pro poměr 375 trénovacích a 625	Uživatelská přesnost pro poměr 375 trénovacích a 625
lesní plocha	90.41 %	300	85.99 %	50	82.84 %	400	79.08 %	400	89.26 %	82.01 %
zástavba	56.00 %	100	66.26 %	50	51.10 %	625	56.05 %	400	51.10 %	57.37 %
půda s vegetací	62.41 %	300	80.75 %	400	56.61 %	150	68.63 %	50	62.24 %	79.44 %
orná půda bez	76.85 %	300	83.17 %	50	62.24 %	625	72.23 %	400	62.24 %	79.44 %
vodní plocha	66.80 %	450	48.38 %	50	36.70 %	625	5.64 %	625	36.70 %	5.64 %

Pro celkovou přesnost klasifikace i uživatelskou a zpracovatelskou přesnost jednotlivých kategorií se v tabulce 7 ukázalo, že změna velikosti validačního datasetu při zachování stabilní velikosti trénovacího datasetu má relativně významný vliv na stabilitu výsledku hodnocení přesnosti klasifikace metodou Maximum Likelihood. Rozdíl 6 % v celkové přesnosti klasifikace lze považovat za relativně velký. U jednotlivých kategorií jsou často rozdíly ještě větší. Pro každou kategorii se našel určitý počet bodů nižší než 625, při němž bylo dosaženo vyšší uživatelské i zpracovatelské přesnosti než v případě 625 bodů.

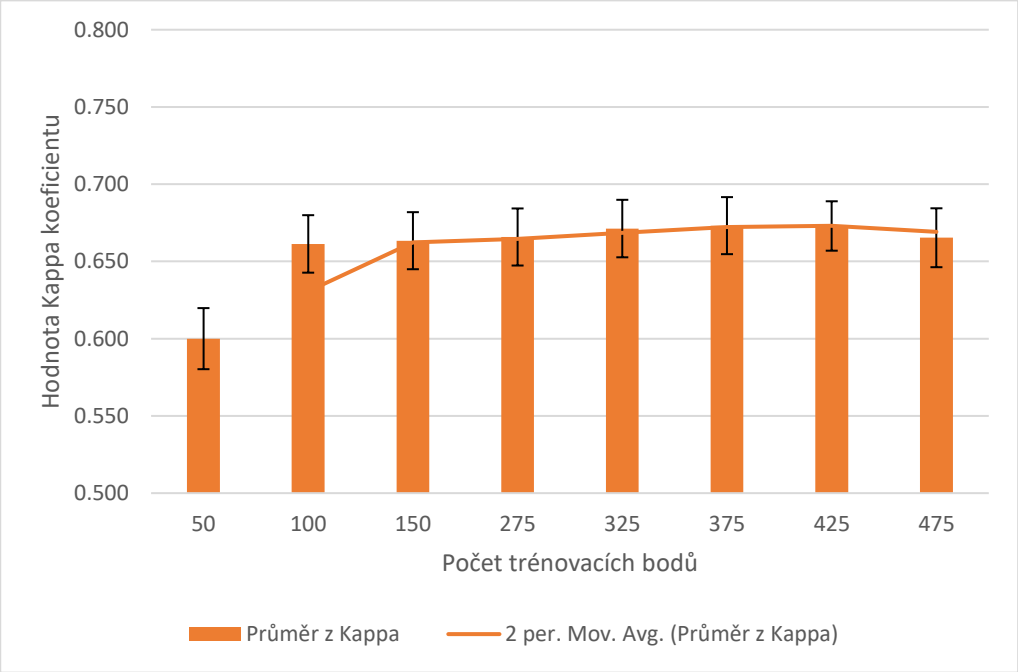
5.3 Testování vlivu podílu/množství trénovacích a validačních dat na přesnost klasifikace metodou Support Vector Machine

Tento test byl proveden pro ověření, zda pro klasifikační algoritmy Maximum Likelihood a Support Vector Machine je optimální podíl trénovacích a validačních dat stejný a zda v případě metody Support Vector Machine stačí k dosažení určité přesnosti méně trénovacích dat než v případě metody Maximum Likelihood. Trénovací data byla vybírána od počtu 375 trénovacích bodů (zároveň tedy 625 validačních bodů) po kroku 10 % (po 50 bodech) až do počtu 275 trénovacích bodů a následně do počtu 475 trénovacích bodů. Podíl 375 trénovacích a 625 validačních bodů byl zvolen v návaznosti na výsledky získané pro metodu Maximum Likelihood. V grafech je nyní na ose x uveden počet trénovacích bodů.

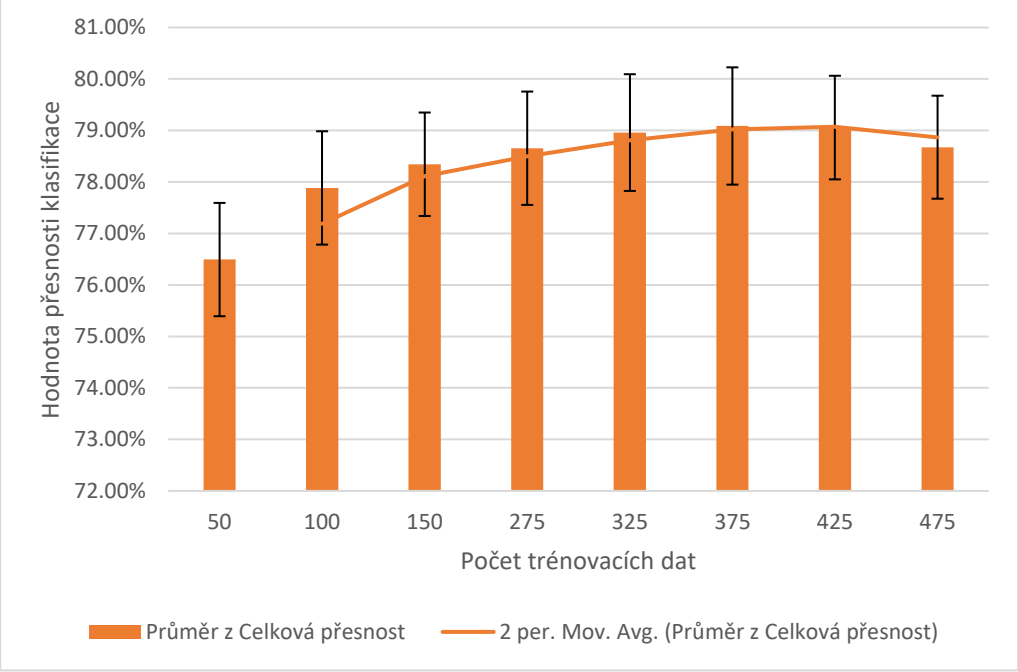
Výsledné hodnoty kappa koeficientu, které lze vidět v grafu 26, se pohybují v rozmezí od 0,600 po hodnotu 0,673. Rozdíl je tedy prakticky zanedbatelný. Nejnižší hodnoty Kappa koeficientu bylo dosaženo pro poměr 50 trénovacích bodů, hodnota koeficientu byla 0,600. Naopak nejvyšší hodnoty Kappa koeficientu (0,673) bylo dosaženo pro poměr 375 trénovacích a 625 validačních bodů. Nejvyšší hodnoty směrodatné odchylky, které lze vidět z grafu 26, dosahují klasifikace pro 475 trénovacích bodů. Hodnota směrodatné odchylky je pro tento počet trénovacích bodů 0,019.

Průběh parametru celkové přesnosti klasifikace lze vidět v grafu 27. Hodnoty se pohybují v rozmezí od 76,49 % dosahující maximálně 79,09 %. Rozdíl je tedy opět minimální a přesnost klasifikace je v případě SVM velice stabilní bez ohledu na podíl trénovacích a validačních dat a také bez ohledu na velikost trénovacího/validačního datasetu. Nejnižší hodnoty celkové přesnosti bylo dosaženo pro poměr 275 trénovacích a 725 validačních bodů. Pro tento poměr byla hodnota přesnosti 78,66 %. Naopak nejvyšší hodnoty přinesl poměr 375 trénovacích ku 625 validačním bodům. Pro tento poměr byla hodnota celkové přesnosti 79,09 %. Stejně jako v grafu 26, kde je zobrazen průměr hodnoty Kappa koeficientu, jsou v grafu celkové přesnosti patrné zlomy.

Pokud máme trénovacích bodů méně než 375 anebo naopak více než 375, je již výsledná hodnota celkové přesnosti klasifikace na nižší hodnotě. Nejvyšší hodnoty směrodatné odchylky, které lze vidět z grafu 27, dosahují hodnoty 0,011 pro klasifikace s 375 trénovacími body. Nejnižších hodnot směrodatné odchylky dosahuje klasifikace pro 475 trénovacích a 525 validačních bodů a to hodnoty 0,010.



Graf 26: průměr Kappa koeficient klasifikace SVM



Graf 27: průměr Celkové přesnosti klasifikace SVM

V případě klasifikace Support Vector Machine bylo stejně jako v případě metody Maximum Likelihood dosaženo nejvyšších hodnot koeficientu Kappa i celkové přesnosti pro poměr 375 trénovacích a 625 validačních bodů. Algoritmem SVM bylo při každém pokusu (testovaném poměru trénovacích a validačních bodů) dosaženo vyšších hodnot celkové přesnosti i Kappa koeficientu než v případě nejlepšího výsledku klasifikace získaného s využitím klasifikačního algoritmu Maximum Likelihood. V případě SVM byly pro všechny testované poměry výsledné hodnoty přesnosti klasifikace vysoké a je tedy možné potvrdit, že v případě metody Support Vector Machine stačí k dosažení určité přesnosti méně trénovacích dat než v případě metody Maximum Likelihood. Lze říct, že i v případě nejnižšího testovaného počtu trénovacích bodů (50) byla celková přesnost v případě SVM relativně srovnatelná s přesností získanou pro nejvyšší testovaný počet trénovacích bodů.

6 DISKUZE

Tato diplomová práce se zabývá optimalizací trénovacího a validačního datasetu pro řízenou klasifikaci dat v DPZ. V rámci řešení diplomové práce jsou v území lesně-luční krajiny v Podkrkonoší prováděny pro dva klasifikační algoritmy experimenty s trénovacími a validačními daty. Práce vychází z předpokladu, že pro dosažení maximální přesnosti v případě klasifikačního algoritmu je ideální podíl 1/3 trénovacích a 2/3 validačních dat (Foody, 2009). Hlavním cílem této diplomové práce bylo testovat vliv podílu/množství trénovacích a validačních dat na přesnossti klasifikace multispektrálních dat senzoru Sentinel-2A s využitím algoritmu Maximum Likelihood.

Pro tyto experimenty s trénovacími a validačními daty bylo vybráno zájmové území, které se nachází na hranici Libereckého a Královehradeckého kraje v podhůří Krkonoš. Pro práci byl využit snímek multispektrálních dat senzoru Sentinel-2A produktu level 1C z termínu 28.8.2016, který je pro dané území zcela bezoblačný. Následně byla vytvořena tematická vektorová vrstva na podkladu snímku S-2A. Pro tvorbu přesné tematické vektorové vrstvy byla využita služba WMS ČÚZK dostupná z geoportálu ČÚZK. Tato tematická vektorová vrstva měla podrobnou legendu, která měla sloužit jako vstup pro další experimenty s trénovacími a validačními body. Pro jednotlivé třídy legendy byla otestována jejich separabilita. Dá se totiž předpokládat, že trénovací plochy s větší separabilitou přinesou také lepší výsledky při klasifikaci. Míra separability však byla pro dvojici travní porost a orná půda s vegetací na hodnotě 0.3359. Z tohoto důvodu byla tato dvojice tříd, stejně jako všechny druhy lesa, sloučena do jedné a test separability tříd byl proveden pro nově vzniklou legendu, která byla označena jako finální pro vstup do dalších pokusů. Pro optimalizaci tvorby trénovacího a validačního datasetu byl nejprve testován podíl trénovacího a validačního datasetu.

Množství trénovacích dat bylo určeno na základě literatury (Foody, 2009), která zkoumá výběr dostatečného množství trénovacích dat. Foody (2009) věnuje pozornost stanovení finální míry přesnosti klasifikace. Na základě této literatury bylo určeno, že pro tento test bude dostatečné množství 1000 bodů, které budou stratifikovaně rozděleny do jednotlivých tříd. Následně byly tyto body podílově rozděleny na trénovací a validační množinu dat. Body byly vždy vybírány stratifikovaně dle podílu plochy dané třídy na celkové rozloze. Problém vznikl při potřebě zachování celočíselných počtů bodů. Z tohoto důvodu byly nejméně početné třídy zaokrouhlovány směrem nahoru. Poté byl sečten počet těchto bodů.

Pro nejpočetnější třídu tak byl počet bodů vytvořen jako doplněk do požadovaného počtu bodů pro daný krok. Pro hodnocení přesnosti klasifikace byla vypočítána chybová matice pomocí funkce Compute Confusion Matrix. K validaci byly využity parametry celková přesnost, Kappa koeficient, zpracovatelská přesnost a uživatelská přesnost. Nevyšších hodnot celkové přesnosti klasifikace MLC a Kappa koeficientu bylo dosaženo pro podíl 375 trénovacích a 625 validačních dat. Celková přesnost pro tento poměr byla 72,88 %. Teorii Foodyho (2009), že pro dosažení nejvyšší přesnosti je ideální podíl 1/3 trénovacích a 2/3 validačních dat, tak výsledky hodnocení celkové přesnosti a Kappa koeficientu pro MLC potvrzují. Avšak výsledné uživatelské a zpracovatelské přesnosti pro jednotlivé třídy nedosáhly v případě tohoto podílu nejvyšších hodnot.

Nejpřesněji klasifikovanou třídou byl les, který byl druhou nejvíce rozsáhlou kategorií v zájmovém území, Relativně dobře byly klasifikovány také třídy orná půda bez vegetace a půda s vegetací.

Nejnižší výsledky zpracovatelské a uživatelské přesnosti byly získány pro nejméně zastoupené třídy. Hodnoty zpracovatelské přesnosti třídy zástavba dosahují oproti jiným třídám nižších hodnot. Tyto výsledky jsou závislé na vysoké heterogenitě prvků v této třídě a rovněž může mít vliv na výsledky právě velice malé množství bodů (70), které tuto třídu reprezentovaly. Nejvyšší hodnoty zpracovatelské přesnosti bylo dosaženo pro poměr 275 trénovacích ku 725 validačním bodům a to hodnoty 52,70 %. Nejvyšší hodnoty uživatelské přesnosti třídy zástavba bylo také dosaženo pro poměr 275 trénovacích ku 725 validačním bodům a to hodnoty 58,55.

Velice špatné výsledky uživatelské přesnosti (v nejlepším případě 6,46 %) byly získány pro třídu vodní plocha. Tato třída byla zastoupena maximálně 10 body, a navíc byla velice špatně odlišitelná od ploch s vegetací viz obrázek 10. Vodní plochy jsou většinou kategorií, která je klasifikována s vysokou přesností. Lze však říct, že zájmové území nebylo z hlediska vodních ploch vybráno ideálně. Jednak jich v této krajině bylo málo, a ty, které se zde vyskytovaly, byly mělké (zčásti vyschlé) případně byl jejich spektrální signál modifikován příměsemi či přítomností vegetace.

Obrázek 10: znázornění vodních ploch (červeně)

Tyto výsledky potvrzují obecně známé pravidlo (viz např. Kang, 2015), že přesnost klasifikace roste s počtem trénovacích bodů, respektive je špatná při nízkém počtu těchto bodů

Dalším cílem práce bylo pro podíl trénovacích a validačních dat, který přinese nejlepší výsledek klasifikace v případě algoritmu Maximum Likelihood měnit množství validačních dat (při zachování stabilní velikosti trénovacího datasetu) a sledovat vliv změny velikosti validačního datasetu na stabilitu výsledku hodnocení přesnosti klasifikace. Výsledné hodnoty Kappa koeficientu se odlišují pouze o 0,06.

Avšak pro celkovou přesnost klasifikace i uživatelskou a zpracovatelskou přesnost jednotlivých kategorií se ukázalo, že změna velikosti validačního datasetu při zachování stabilní velikosti trénovacího datasetu má relativně významný vliv na stabilitu výsledku hodnocení přesnosti klasifikace metodou Maximum Likelihood.

Poměr 1/3 trénovacích a 2/3 validačních bodů tedy nemusí být univerzálně nejvýhodnějším poměrem při hodnocení přesnosti klasifikace. Výsledky ukazují, že na základě testování nelze učinit jednoznačný závěr o „ideálním" podílu trénovacích a validačních dat. Dále napovídají i to, že „jistota" v případě hodnocení přesnosti klasifikace je velice relativní. Ke změně výsledných hodnot klasifikačních přesností může přispívat množství faktorů a jedním z nich je určitě i počet validačních bodů. Na základě výsledků analýzy ovšem nelze říct, při jakém počtu validačních bodů lze dosáhnout „korektní", resp. „absolutně správné" hodnoty přesnosti. Získání odpovědi na tuto otázku by si vyžádalo provedení dalších testů. Svou roli určitě bude hrát i počet trénovacích bodů, který v této práci testován nebyl. Je zřejmé, že výsledky hodnocení přesnosti klasifikace dat DPZ jsou do určité míry vágní a nepřesné. Lze říct, že do značné míry mohou být ovlivněny právě i množstvím a podílem trénovacích a validačních dat.

Třetím cílem práce bylo ověřit, zda v případě metody Support Vector Machine stačí k dosažení určité přesnosti klasifikace méně trénovacích dat než v případě metody Maximum Likelihood. Na základě výsledků tohoto testu lze říci, že se poměr 375 trénovacích a 625 validačních bodů ukázal jako nejlepší i v případě hodnocení celkové přesnosti klasifikace SVM. Celková přesnost klasifikace pro tento poměr dosáhla hodnoty 79,09 %, což je o více než 6% lepší výsledek, než v případě klasifikace MLC. Zásadní ale je, že i v případě velmi nízkého počtu trénovacích bodů (150, 100, 50) byla celková přesnost klasifikace SVM stále vysoká (kolem 78 %, nejméně 76,5 %). Přesnost klasifikace jednotlivých tříd v tomto případě hodnocena nebyla. V případě MLC celková přesnost klesá pod 70 % od počtu 100 trénovacích bodů a pro 50 trénovacích bodů dosahuje pouze 66,6 %. Je tedy zřejmé, že klasifikační algoritmus SVM dosáhne velmi vysoké celkové přesnosti klasifikace (a podobně stabilní jsou i výsledky pro Kappa koeficient) i s relativně nízkým počtem trénovacích bodů a že pro dosažení stejné přesnosti, jako algoritmus MLC mu stačí méně trénovacích bodů. Výsledek klasifikace může být, ale ovlivněn i dalšími faktory jako je například rozložení trénovacího datasetu. Výsledky pro sledovaný dataset lze ale vzhledem k počtu opakování považovat za spolehlivé. To lze prohlásit o všech testech provedených v této práci.

ZÁVĚR

Tato diplomová práce se zabývala optimalizací trénovacího a validačního datasetu pro řízenou klasifikaci dat v DPZ. Literatura věnovaná strategiím výběru trénovacích datasetů se zaměřuje na tři aspekty, které mají největší vliv na následnou přesnost klasifikace. Hlavními aspekty jsou distribuce neboli podíl a rozdělení trénovacích a validačních dat do tříd, množství vstupních dat a jejich rozmístěn. V rámci řešení diplomové práce jsou v území lesně-luční krajiny v Podkrkonoší prováděny pro dva klasifikační algoritmy experimenty s trénovacími a validačními daty. Práce vychází z předpokladu, že pro dosažení maximální přesnosti v případě klasifikačního algoritmu je ideální podíl 1/3 trénovacích a 2/3 validačních dat (Foody, 2009).

Hlavním cílem práce bylo testovat vliv podílu/množství trénovacích a validačních dat na přesnosti klasifikace multispektrálních dat senzoru Sentinel-2A s využitím algoritmu Maximum Likelihood. Nejvyšší celkové přesnosti klasifikace metodou Maximum Likelihood bylo dosaženo pro podíl 375 trénovacích a 625 validačních bodů. Celková přesnost pro tento podíl byla 72,88 %. Dle výsledků tohoto testu lze říci, že pro tento klasifikační algoritmus je ideální podíl 1/3 trénovacích a 2/3 validačních dat, pokud hodnotíme celkovou přesnost klasifikace, pro jednotlivé třídy se tento předpoklad nepotvrdil.

Druhým cílem práce bylo pro podíl trénovacích a validačních dat, který přinese nejlepší výsledek klasifikace v případě algoritmu Maximum Likelihood, měnit množství validačních dat (při zachování stabilní velikosti trénovacího datasetu) a sledovat vliv změny velikosti validačního datasetu na stabilitu výsledku hodnocení přesnosti klasifikace. Pro celkovou přesnost klasifikace i uživatelskou a zpracovatelskou přesnost jednotlivých kategorií se ukázalo, že změna velikosti validačního datasetu při zachování stabilní velikosti trénovacího datasetu má relativně významný vliv na stabilitu výsledku hodnocení přesnosti klasifikace metodou Maximum Likelihood.

Práce chtěla ověřit také to, zda v případě metody Support Vector Machine stačí k dosažení určité přesnosti méně trénovacích dat než v případě metody Maximum Likelihood. Tento předpoklad byl pro tento dataset potvrzen.

Práce je doplněna o vytvořený skript, který umožňuje automatické experimentování s nastavením trénovacích/validačních dat. Skript by v budoucnu mohl být upraven tak aby mohly být provedeny další testy, které by přispěly k optimalizaci trénovacího a validačního datasetu a v důsledku potom ke zvýšení přesnosti klasifikace. Například by bylo možné experimentovat s množstvím trénovacích dat nebo jejich rozmístěním. Možné by bylo také provést stejné testy v jiném území, které by obsahovalo více vodních ploch a celkově jiný poměr testovaných kategorií. Testovány by mohly být také jiné způsoby výběru trénovacích či validačních dat než v práci použitý stratifikovaný výběr.

SEZNAM POUŽITÉ LITERATURY

ARCDATA (2018): ENVI [online]. [cit. 13. 4. 2019] Dostupné z: https://www.arcdata.cz/produkty/envi

BAATZ, M., SCHAPE, A., 2000. Multiresolution Segmentation – an optimization approach for highquality multi-scale image segmentation. Angewandte Geographische Informations-Verarbeitung [online]. [cit. 13. 4. 2019] 12, str. 12-23. Dostupné z: http://www.ecognition.com/sites/default/files/405_baatz_fp_12.pdf

BELOUSOV A.I., et al. (2002): A flexible classification approach with optimal generalisation performance: support vector machines; chemometrics and Intelligent laboratory systems, 64. 15–25;

BOLSTAD, P., LILLESAND, T. M. (1991): Rapid maximum likelihood classification. Photogrammetric Engineering and Remote Sensing. Vydání 57, č. 1, s. 67–74.

CAMPBELL, J.B., Wynne, R.H. (2011): Introduction to remote sensing, 5th edition, The Guildford press, ISBN 978-1-60918-176-5;

CAMPBELL, J. B. (1996): Introduction to Remote Sensing. Taylor & Francis, London, 622 s.

ČVUT (2018): Distribuční funkce standardního normálního rozdělení [online]. [cit. 13. 4. 2019] Dostupné z: https://mat.fsv.cvut.cz/hala/files/Distribucni_funkce_standardniho_normalniho_rozdeleni.pdf

DOBROVOLNÝ, P. (1998): Dálkový průzkum Země: digitální zpracování obrazu. 1. vyd. Brno: Masarykova univerzita, 208 s. ISBN 80-210-1812-7

DOBROWSKI S.Z., et al. (2008): Mapping mountain vegetation using species distribution modelling, image-based texture analysis and object- based classification; Applied vegetation science, 11(4), 499 – 508;

DRUSCH, M…[et al.] (2012): Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote Sensing of.Environment, 120, s. 25–36. DOI: 10.1016/j.rse.2011.11.026

FOODY, G. M. (2004): Thematic map comparison: evaluating the statistical signifikance of differences in classification accuracy, Photogrammetric Engineering and Remote Sensing, 70, 627-633.

FOODY, G. M. (2008): Harshness in image classification accuracy assessment, International Journal of Remote Sensing, 29, 3137-3158.

FOODY, G. M. (2009): Sample size determination for image classification accuracy assessment and comparison, International Journal of Remote Sensing.

FOODY, G.M., Mathur, A. (2006): The use of small training sets containing mixed pixels for accurate hard image classification: training on mixed spectral responses for classification by a SVM. Remote Sens. Environ. 103 (2), 179–189.

FOODY, G.M (1995): Land cover classification by an artificial neural network with ancillary information. Int. J. Geographical Inf. Syst. 9 (5), 527–542.

FOODY, G.M., Mathur, (2004): Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification, Remote sensing of environment, 93, 107–117.

FOODY, G.M., Arora, M.K. (1997): An evaluation of some factors affecting the accuracy of classification by an artificial neural network. Int. J. Remote Sens. 18 (4), 799–810.

GALLEGO, F. J. (2005). Stratified sampling of satellite images with a systematic grid of points. Journal of Photogrammetry and Remote Sensing, 59, 369–376.

GISAT (2017): Sentinel-2A [online]. [cit. 13. 4. 2019] Dostupné z: http://www.gisat.cz/content/cz/dpz/prehled-druzicovych-systemu/satelite/sentinel-2-a

HALOUNOVÁ, L., PAVELKA, K., 2005. Dálkový průzkum Země. Praha: České vysoké učení technické. 1. vyd. 192 str. ISBN 978-80-01-03124-7.

HARRIS GEOSPATIAL SOLUTIONS (2006): Calculate Confussion Matrices (Using ENVI) [online]. [cit. 13. 4. 2019]. Dostupné z: http://www.harrisgeospatial.com/Support/HelpArticlesDetail/TabId/219/ArtMID/900/ArticleID/4059/4059.aspx

HARRIS GEOSPATIAL SOLUTIONS (2016): Supplying a proper data scaling factor in ENVI Maximum Likelihood Classification [online]. [cit. 13. 4. 2019] Dostupné z: http://www.harrisgeospatial.com/docs/CalculatingConfusionMatrices.html

HARRIS GEOSPATIAL SOLUTIONS (2018): Region of Interest (ROI) Tool [online]. [cit. 13. 4. 2019] Dostupné z: https://www.harrisgeospatial.com/docs/RegionOfInterestTool.html

HROMÁDKOVÁ, L. (2015): Classification of meadow vegetation in the Krkonoše Mts. using aerial hyperspectral data and support vector machines classifier. Diplomová práce. Univerzita Karlova v Praze, Přírodovědecká fakulta. 137 s.

HUANG, C., et al. (2002): An assessment of support vector machines for land cover classification; International journal of remote sensing, 23/4; 725–749;

IMAN, R. L., & SONOVER, W. J. (1980). Small sample sensitivity analysis techniques for computer models, with an application to risk assessment. Communications in Statistics - Theory and Methods, A9, 1749–1874.

JENSEN, R. J. (2005): Introductory Digital Image Processing: A Remote Sensing Perspective. Upper 101 Saddle River, N.J: Prentice Hall. 3. vyd. 544 str. ISBN 978-0131453616.

JONES H.G, VAUGHAN (2010): Remote sensing of vegetation: Principles, techniques and applications; Oxford: Oxford university press, 2010, ISBN 978-0-19-920779-4;

KANG, F. (2015): System probabilistic stability analysis of soil slopes using Gaussian process regression with latin hypercube sampling Computers and Geotechnics 63, 13–25

KOLÁŘ, J. (1990): Dálkový průzkum Země. 1. vyd. Praha: SNTL – Nakladatelství technické literatury, 170 s. Populární přednášky o fyzice, sv. 35. ISBN 80-03-00517-5.

KOLÁŘ, J.; HALOUNOVÁ, L.; PAVELKA, K. (1997): Dálkový průzkum Země 10. 1. vyd. Praha: Vydavatelství ČVUT. 1997, 164 s. ISBN: 80-01-01567-X

KUTHAN, T. (2019): Klasifikace vybraných zemědělských plodin v modelovém území kunohorska s využitím časové řady dat Sentinel-2A a Planetscope. Diplomová práce. Univerzita Karlova v Praze, Přírodovědecká fakulta. 107 s.

LILLESAND, T. M., KIEFER, R. W. (1994): Remote sensing and image interpretation. John Wiley & Sons. New York, Chichester, Brisbane, Toronto, Singapore, 750 s.

LIN, Y.P. (2008): Geostatistical Approaches and Optimal Additional Sampling Schemes for Spatial Patterns and Future Samplings of Bird Diversity. Global Ecol. Biogeogr. 17, 175–188.

McKay, M. D., Beckman, R. J., & Conover, W. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics, 21, 239–245.

MILLARD, K. (2015): On the Importance of Training Data Sample Selection in Random Forest Image Classification: A Case Study in Peatland Ecosystem Mapping, ISSN 2072-4292

MOUNTRAKIS, G.; IM, J.; OGOLE, C. (2011): Support vector machines in remote sensing: A review. ISPRS Journal of Photogrammetry and Remote Sensing. Vydání 66, č. 3, s. 247–259.

NOVAK, D. (2009): Correlation control in small-sample Monte Carlo type simulations I. A simulated annealing approach. Probabilistic Engineering Mechanics, Vol. 24, Issue 3, 2009, pp. 452 462

PAL M. and MATHER P.M. (2006): Some issues in the classification of DAIS hyperspectral data, International Journal of Remote Sensing, 27, 2895-2916

RADOUX, J., et al (2014): Automated training sample extraction for global land cover mapping. Remote Sens. 6 (5), 3965–3987.

RADOUX, J. et al (2016): Sentinel-2's Potential for Sub-Pixel Landscape Feature Detection. Remote sensing 8, 2-28.

SHA Z., et al. (2009): Using a hybrid fuzzy classifier (HFC) to map typical grassland vegetation in Xilin River Basin, Inner Mongolia, China; International journal of remote sensing; 29:8, 2317–2337;

SCHMIDTLEIN S., SASSIN J. (2004): Mapping of continous floristic gradients in grasslands using hyperspectral imagery; Remote sensing of environment 92, 126–138;

SVOBODA, T. a HILAR, M. (2013): PravděpodobnostnÍ analýzy metodou Latin Hypercube Sampling 3G Consulting Engineers, Prague, Czech Republic

YU, Q. et al. (2006): Object-based detailed vegetation classification with airborne high spatial resolution remote sensing imagery. Photogrammetric Engineering & Remote Sensing, 72, č. 7, s. 799–811.

YU-PIN Lin, et al. (2011): Monitoring and identification of spatiotemporal landscape changes in multiple remote sensing images by using a stratified conditional latin hypercube sampling approach and geostatistical simulation. Environ Monit Assess 177:353–373

YU-PIN Lin, et al. (2009): Remote Sensing Data with the Conditional Latin Hypercube Sampling and Geostatistical Approach to Delineate Landscape Changes Induced by Large Chronological Physical Disturbances, Sensor Journal, ISSN 1424-8220, 148-174

PŘÍLOHY

Příloha 1 – Skript pro vygenerování 1000 vstupních bodů, stratifikovaně rozdělených do tříd

```
import arcpy



arcpy.env.overwriteOutput = 1

arcpy.env.workspace = "E:/B/NaturCuni/Diplomka/vrstvy"



"lesy_dissolve"

arcpy.CreateRandomPoints_management(out_path="E:/B/Natur-
Cuni/Diplomka/body_arcmap", out_name="body500_les", constrain-
ing_feature_class="lesy_dissolve", constraining_extent="0 0 250
250", number_of_points_or_field="370", minimum_allowed_distance="0
Meters", create_multipoint_output="POINT", multipoint_size="0")

"body500_les"

arcpy.CalculateField_management(in_table="body500_les",
field="CID", expression="1", expression_type="VB", code_block="")



"voda_dissolve"

arcpy.CreateRandomPoints_management(out_path="E:/B/Natur-
Cuni/Diplomka/body_arcmap", out_name="body500_voda", constrain-
ing_feature_class="voda_dissolve", constraining_extent="0 0 250
250", number_of_points_or_field="10", minimum_allowed_distance="0
Meters", create_multipoint_output="POINT", multipoint_size="0")

"body500_voda"

arcpy.CalculateField_management(in_table="body500_voda",
field="CID", expression="5", expression_type="VB", code_block="")



"ornaBEZ_dissolve"

arcpy.CreateRandomPoints_management(out_path="E:/B/Natur-
Cuni/Diplomka/body_arcmap", out_name="body500_orna", constrain-
ing_feature_class="ornaBEZ_dissolve", constraining_extent="0 0 250
250", number_of_points_or_field="130", minimum_allowed_distance="0
Meters", create_multipoint_output="POINT", multipoint_size="0")
```

```
"body500_orna"

arcpy.CalculateField_management(in_table="body500_orna",
field="CID", expression="4", expression_type="VB", code_block="")



"puda_s_vege_celkove_dissolve"

arcpy.CreateRandomPoints_management(out_path="E:/B/Natur-
Cuni/Diplomka/body_arcmap", out_name="body500_ornaS", constrain-
ing_feature_class="puda_s_vege_celkove_dissolve", constraining_ex-
tent="0 0 250 250", number_of_points_or_field="420", minimum_al-
lowed_distance="0 Meters", create_multipoint_output="POINT", mul-
tipoint_size="0")

"body500_ornaS"

arcpy.CalculateField_management(in_table="body500_ornaS",
field="CID", expression="3", expression_type="VB", code_block="")



"zastavba_dissolve"

arcpy.CreateRandomPoints_management(out_path="E:/B/Natur-
Cuni/Diplomka/body_arcmap", out_name="body500_zastavba", constrain-
ing_feature_class="zastavba_dissolve", constraining_extent="0 0 250
250", number_of_points_or_field="70", minimum_allowed_distance="0
Meters", create_multipoint_output="POINT", multipoint_size="0")

"body500_zastavba"

arcpy.CalculateField_management(in_table="body500_zastavba",
field="CID", expression="2", expression_type="VB", code_block="")



arcpy.Merge_management(in-
puts="body500_les;body500_zastavba;body500_or-
naS;body500_orna;body500_voda", output="E:/B/Natur-
Cuni/Diplomka/body_arcmap/body500_vse.shp", field_mappings='CID
"CID" true true false 10 Long 0 10 ,First,#,body500_les,CID,-1,-
1,body500_zastavba,CID,-1,-1,body500_ornaS,CID,-1,-
1,body500_orna,CID,-1,-1,body500_voda,CID,-1,-1')
```

PŘÍLOHA 2 – script pro vytvoření klasifikace Maximum Likelihood

```
import arcpy, random, sys



pocet_trenovacich = int(sys.argv[1])

pokus = int(sys.argv[2])



arcpy.env.overwriteOutput = 1

cesta = "C:/Users/Edmyon/Desktop/Diplomka"

databaze = cesta + "/validace.gdb"

dataset = databaze + "/UTM33"

arcpy.env.workspace = dataset



zakladni_body = "body1000_vse"

grnd_true_attrib = "GrndTruth"



#psti = [0.371754081, 0.071711226, 0.423855961, 0.131571104,
0.001107627]

psti = [0.37, 0.07, 0.42, 0.13, 0.01]



def select_random(fc, attrib, value, count, output):

    # Vybere z $fc nahodnych $count zaznamu takovych, ze maji hod-
notu atributu $attrib rovnu $value a ulozi se jako $output.
```

```
 # pridame pole
    try:
        arcpy.AddField_management(fc, "SEL", "SHORT")
    except:
        pass

    arcpy.CalculateField_management(fc, "SEL", 0)
    # zjistime, kolik je zaznamu splnujicich podminku
    arcpy.MakeFeatureLayer_management(fc, "tmp", attrib + "="+
str(value))
    res = arcpy.GetCount_management("tmp")
    avail_cnt = int(res.getOutput(0))
    arcpy.Delete_management("tmp")

    if count > avail_cnt:
    # chci jich vybrat vic, nez jich je
        indices = avail_cnt * [1]
        # vyberu vsechny

    else:
        sel_idx = random.sample(range(0, avail_cnt), count)
        indices = avail_cnt * [0]
        for i in sel_idx:
            indices[i] = 1
            # pole priznaku: ma-li byt vybran, bude v nem 1
```

```
    cur = arcpy.UpdateCursor(fc, attrib + "="+ str(value))
    i = 0
    for row in cur:
        row.setValue("SEL", indices[i])
        cur.updateRow(row)
        i += 1
    del cur


    arcpy.MakeFeatureLayer_management(fc, "tmp")
    arcpy.SelectLayerByAttribute_management("tmp", "NEW_SELECTION",
"SEL = 1")
    arcpy.CopyFeatures_management("tmp", output)
    arcpy.Delete_management("tmp")


    arcpy.DeleteField_management(output, "SEL")
    arcpy.DeleteField_management(fc, "SEL")


def get_table_value(table, row_no, column_name):
    # Vybere z tabulky $table hodnotu na radku $row_no (cislovani
radku od 1) a ve sloupci s nazvem $column_name


    cur = arcpy.SearchCursor(table)
    i = 1
    while i <= row_no:
        row = cur.next()
        i += 1


    ret_val = row.getValue(column_name)
    del cur
```

```
    return ret_val

#program zacina tady


try:

    arcpy.CheckOutExtension("spatial")



    select_random(zakladni_body, grnd_true_attrib, 1,
int(round(pocet_trenovacich * psti[0])), "tren_clas_1")

    select_random(zakladni_body, grnd_true_attrib, 2,
int(round(pocet_trenovacich * psti[1])), "tren_clas_2")

    select_random(zakladni_body, grnd_true_attrib, 4,
int(round(pocet_trenovacich * psti[3])), "tren_clas_4")

    select_random(zakladni_body, grnd_true_attrib, 5,
int(round(pocet_trenovacich * psti[4])), "tren_clas_5")




    arcpy.Merge_manage-
ment("tren_clas_1;tren_clas_2;tren_clas_4;tren_clas_5", "treno-
vaci_mezikrok")




    pocet = arcpy.GetCount_management("trenovaci_mezikrok")

    pocet_int = int(float(pocet.getOutput(0)))

    odecist = pocet_trenovacich - pocet_int



    select_random(zakladni_body, grnd_true_attrib, 3, odecist,
"tren_clas_vygen3")

    arcpy.Merge_management("tren_clas_vygen3;trenovaci_mezikrok",
"trenovaci")



    arcpy.Erase_analysis(zakladni_body,"trenovaci", cesta + "/vali-
dacni.shp", 0.10)
```

```
    arcpy.gp.CreateSignatures_sa(databaze + "/snimek", "trenovaci",
cesta + "/signature_body.GSG", "COVARIANCE", grnd_true_attrib)



    arcpy.gp.MLClassify_sa(databaze + "/snimek", cesta + "/signa-
ture_body.GSG", databaze + "/klas_snimek", "0.0", "EQUAL", "", "")









    arcpy.gp.UpdateAccuracyAssessmentPoints_sa(databaze +
"/klas_snimek", cesta + "/validacni.shp", cesta + "/validacni_up-
date.shp", "CLASSIFIED")




    arcpy.gp.ComputeConfusionMatrix_sa(cesta + "/validacni_up-
date.shp", databaze + "/matice")



    kappa = get_table_value(databaze + "/matice", 8, "Kappa")

    celkova_presnost = get_table_value(databaze + "/matice", 7,
"U_Accuracy")

    zpracovatelska_presnost1 = get_table_value(databaze +
"/matice", 7, "C_1")

    zpracovatelska_presnost2 = get_table_value(databaze +
"/matice", 7, "C_2")

    zpracovatelska_presnost3 = get_table_value(databaze +
"/matice", 7, "C_3")

    zpracovatelska_presnost4 = get_table_value(databaze +
"/matice", 7, "C_4")

    zpracovatelska_presnost5 = get_table_value(databaze +
"/matice", 7, "C_5")
```

```
    uzivatelska_presnost1 = get_table_value(databaze + "/matice",
1, "U_Accuracy")

    uzivatelska_presnost2 = get_table_value(databaze + "/matice",
2, "U_Accuracy")

    uzivatelska_presnost3 = get_table_value(databaze + "/matice",
3, "U_Accuracy")

    uzivatelska_presnost4 = get_table_value(databaze + "/matice",
4, "U_Accuracy")

    uzivatelska_presnost5 = get_table_value(databaze + "/matice",
5, "U_Accuracy")




    print pocet_trenovacich, pokus, kappa, celkova_presnost,
zpracovatelska_presnost1, zpracovatelska_presnost2, zpraco-
vatelska_presnost3, zpracovatelska_presnost4, zpracovatelska_pres-
nost5, uzivatelska_presnost1, uzivatelska_presnost2, uzi-
vatelska_presnost3, uzivatelska_presnost4, uzivatelska_presnost5

except:

    print pocet_trenovacich, pokus, " chyba"
```

PŘÍLOHA 3 – script pro vytvoření klasifikace SVM

```
import arcpy, random, sys

pocet_trenovacich = int(sys.argv[1])
pokus = int(sys.argv[2])

arcpy.env.overwriteOutput = 1
cesta = "C:\Users\Edmyon\Desktop\Diplomka\data"
arcpy.env.workspace = cesta

zakladni_body = "body1000_vse.shp"
grnd_true_attrib = "GrndTruth"

#psti = [0.371754081, 0.071711226, 0.423855961, 0.131571104,
0.001107627]
psti = [0.37, 0.07, 0.42, 0.13, 0.01]

def select_random(fc, attrib, value, count, output):
    # Vybere z $fc nahodnych $count zaznamu takovych, ze maji hodnotu
atributu $attrib rovnu $value a ulozi se jako $output.

    # pridame pole
    try:
        arcpy.AddField_management(fc, "SEL", "SHORT")
    except:
        pass

    arcpy.CalculateField_management(fc, "SEL", 0)
    # zjistime, kolik je zaznamu splnujicich podminku
    arcpy.MakeFeatureLayer_management(fc, "tmp", attrib + "="+
str(value))
    res = arcpy.GetCount_management("tmp")
    avail_cnt = int(res.getOutput(0))
    arcpy.Delete_management("tmp")


    if count > avail_cnt:
    # chci jich vybrat vic, nez jich je
        indices = avail_cnt * [1]
        # vyberu vsechny

    else:
        sel_idx = random.sample(range(0, avail_cnt), count)
        indices = avail_cnt * [0]
        for i in sel_idx:
            indices[i] = 1
            # pole priznaku: ma-li byt vybran, bude v nem 1


    cur = arcpy.UpdateCursor(fc, attrib + "="+ str(value))
    i = 0
    for row in cur:
        row.setValue("SEL", indices[i])
        cur.updateRow(row)
        i += 1
    del cur

    arcpy.MakeFeatureLayer_management(fc, "tmp")
```

```
    arcpy.SelectLayerByAttribute_management("tmp", "NEW_SELECTION",
"SEL = 1")
    arcpy.CopyFeatures_management("tmp", output)
    arcpy.Delete_management("tmp")

    arcpy.DeleteField_management(output, "SEL")
    arcpy.DeleteField_management(fc, "SEL")

def get_table_value(table, row_no, column_name):
    # Vybere z tabulky $table hodnotu na radku $row_no (cislovani
radku od 1) a ve sloupci s nazvem $column_name

    cur = arcpy.SearchCursor(table)
    i = 1
    while i <= row_no:
        row = cur.next()
        i += 1

    ret_val = row.getValue(column_name)
    del cur

    return ret_val


#program zacina tady

try:
    arcpy.CheckOutExtension("spatial")

    select_random(zakladni_body, grnd_true_attrib, 1,
int(round(pocet_trenovacich * psti[0])), "tren_clas_1.shp")
    select_random(zakladni_body, grnd_true_attrib, 2,
int(round(pocet_trenovacich * psti[1])), "tren_clas_2.shp")
    select_random(zakladni_body, grnd_true_attrib, 4,
int(round(pocet_trenovacich * psti[3])), "tren_clas_4.shp")
    select_random(zakladni_body, grnd_true_attrib, 5,
int(round(pocet_trenovacich * psti[4])), "tren_clas_5.shp")


    arcpy.Merge_management(["tren_clas_1.shp", "tren_clas_2.shp",
"tren_clas_4.shp", "tren_clas_5.shp"], "trenovaci_mezikrok.shp")


    pocet = arcpy.GetCount_management("trenovaci_mezikrok.shp")
    pocet_int = int(float(pocet.getOutput(0)))
    odecist = pocet_trenovacich - pocet_int

    select_random(zakladni_body, grnd_true_attrib, 3, odecist,
"tren_clas_vygen3.shp")
    arcpy.Merge_management(["tren_clas_vygen3.shp", "treno-
vaci_mezikrok.shp"], "trenovaci.shp")

    arcpy.Erase_analysis(zakladni_body, "trenovaci.shp", "vali-
dacni.shp", 0.10)

    # varianta 1 -- vyuziva segmentaci
    # seg_raster = arcpy.gp.SegmentMeanShift_sa("snimek.img",
"snimek_sms.img", 18, 18, 20)
    # arcpy.gp.TrainSupportVectorMachineClassi-
fier_sa("snimek_sms.img", "trenovaci.shp", "svm_params.ecd", "", "",
"COLOR;MEAN")
```

```
    # classified_raster = arcpy.sa.ClassifyRaster("snimek_sms.img",
"svm_params.ecd")
    # classified_raster.save("snimek_classified.img")

    # varianta 2 -- nevyuziva segmentaci
    arcpy.gp.TrainSupportVectorMachineClassifier_sa("snimek.img",
"trenovaci.shp", "svm_params.ecd", "", "", "COLOR;MEAN")
    classified_raster = arcpy.sa.ClassifyRaster("snimek.img",
"svm_params.ecd")
    classified_raster.save("snimek_classified.img")
    # vyberete si variantu 1 nebo variantu 2 a tu druhou smazete/za-
komentujete

    arcpy.gp.UpdateAccuracyAssessmentPoints_sa("snimek_classi-
fied.img", "validacni.shp", "validacni_update.shp", "CLASSIFIED")

    #"validacni_1_update"
    arcpy.gp.ComputeConfusionMatrix_sa("validacni_update.shp",
"matice.dbf")

    kappa = get_table_value("matice.dbf", 8, "Kappa")
    celkova_presnost = get_table_value("matice.dbf", 7, "U_Accuracy")
    zpracovatelska_presnost1 = get_table_value("matice.dbf", 7, "C_1")
    zpracovatelska_presnost2 = get_table_value("matice.dbf", 7, "C_2")
    zpracovatelska_presnost3 = get_table_value("matice.dbf", 7, "C_3")
    zpracovatelska_presnost4 = get_table_value("matice.dbf", 7, "C_4")
    zpracovatelska_presnost5 = get_table_value("matice.dbf", 7, "C_5")
    uzivatelska_presnost1 = get_table_value("matice.dbf", 1, "U_Accu-
racy")
    uzivatelska_presnost2 = get_table_value("matice.dbf", 2, "U_Accu-
racy")
    uzivatelska_presnost3 = get_table_value("matice.dbf", 3, "U_Accu-
racy")
    uzivatelska_presnost4 = get_table_value("matice.dbf", 4, "U_Accu-
racy")
    uzivatelska_presnost5 = get_table_value("matice.dbf", 5, "U_Accu-
racy")


    print pocet_trenovacich, pokus, kappa, celkova_presnost, zpraco-
vatelska_presnost1, zpracovatelska_presnost2, zpracovatelska_pres-
nost3, zpracovatelska_presnost4, zpracovatelska_presnost5, uzi-
vatelska_presnost1, uzivatelska_presnost2, uzivatelska_presnost3, uzi-
vatelska_presnost4, uzivatelska_presnost5
except:
    print pocet_trenovacich, pokus, " chyba"
```

PŘÍLOHA 4 – script pro vytvoření klasifikace MLC s úpravou množství validačních bodů

```
import arcpy, random, sys

pocet_trenovacich = int(sys.argv[1])
pokus = int(sys.argv[2])
pocet_validacnich = int(sys.argv[3])

arcpy.env.overwriteOutput = 1
cesta = "C:/Users/Edmyon/Desktop/Diplomka"
databaze = cesta + "/validace.gdb"
dataset = databaze + "/UTM33"
arcpy.env.workspace = dataset

zakladni_body = "body1000_vse"
grnd_true_attrib = "GrndTruth"

#psti = [0.371754081, 0.071711226, 0.423855961, 0.131571104,
0.001107627]
psti = [0.37, 0.07, 0.42, 0.13, 0.01]

def select_random(fc, attrib, value, count, output):
    # Vybere z $fc nahodnych $count zaznamu takovych, ze maji hodnotu
atributu $attrib rovnu $value a ulozi se jako $output.

    # pridame pole
    try:
        arcpy.AddField_management(fc, "SEL", "SHORT")
    except:
        pass

    arcpy.CalculateField_management(fc, "SEL", 0)
    # zjistime, kolik je zaznamu splnujicich podminku
    arcpy.MakeFeatureLayer_management(fc, "tmp", attrib + "="+
str(value))
    res = arcpy.GetCount_management("tmp")
    avail_cnt = int(res.getOutput(0))
    arcpy.Delete_management("tmp")


    if count > avail_cnt:
    # chci jich vybrat vic, nez jich je
        indices = avail_cnt * [1]
        # vyberu vsechny

    else:
        sel_idx = random.sample(range(0, avail_cnt), count)
        indices = avail_cnt * [0]
        for i in sel_idx:
            indices[i] = 1
            # pole priznaku: ma-li byt vybran, bude v nem 1


    cur = arcpy.UpdateCursor(fc, attrib + "="+ str(value))
    i = 0
    for row in cur:
        row.setValue("SEL", indices[i])
        cur.updateRow(row)
        i += 1
    del cur
```

```
    arcpy.MakeFeatureLayer_management(fc, "tmp")
    arcpy.SelectLayerByAttribute_management("tmp", "NEW_SELECTION",
"SEL = 1")
    arcpy.CopyFeatures_management("tmp", output)
    arcpy.Delete_management("tmp")

    arcpy.DeleteField_management(output, "SEL")
    arcpy.DeleteField_management(fc, "SEL")

def get_table_value(table, row_no, column_name):
    # Vybere z tabulky $table hodnotu na radku $row_no (cislovani
radku od 1) a ve sloupci s nazvem $column_name

    cur = arcpy.SearchCursor(table)
    i = 1
    while i <= row_no:
        row = cur.next()
        i += 1

    ret_val = row.getValue(column_name)
    del cur

    return ret_val


#program zacina tady


arcpy.CheckOutExtension("spatial")

select_random(zakladni_body, grnd_true_attrib, 1,
int(round(pocet_trenovacich * psti[0])), "tren_clas_1")
select_random(zakladni_body, grnd_true_attrib, 2,
int(round(pocet_trenovacich * psti[1])), "tren_clas_2")
select_random(zakladni_body, grnd_true_attrib, 4,
int(round(pocet_trenovacich * psti[3])), "tren_clas_4")
select_random(zakladni_body, grnd_true_attrib, 5, max(4,
int(round(pocet_trenovacich * psti[4]))), "tren_clas_5")


arcpy.Merge_manage-
ment("tren_clas_1;tren_clas_2;tren_clas_4;tren_clas_5", "treno-
vaci_mezikrok")


pocet = arcpy.GetCount_management("trenovaci_mezikrok")
pocet_int = int(float(pocet.getOutput(0)))
odecist = pocet_trenovacich - pocet_int

select_random(zakladni_body, grnd_true_attrib, 3, odecist,
"tren_clas_vygen3")
arcpy.Merge_management("tren_clas_vygen3;trenovaci_mezikrok", "treno-
vaci")

#tady zacnu vybirat validacni

select_random(zakladni_body, grnd_true_attrib, 1, int(round(pocet_val-
idacnich * psti[0])), "vali_clas_1")
select_random(zakladni_body, grnd_true_attrib, 2, int(round(pocet_val-
idacnich  * psti[1])), "vali_clas_2")
```

```
select_random(zakladni_body, grnd_true_attrib, 4, int(round(pocet_val-
idacnich  * psti[3])), "vali_clas_4")
select_random(zakladni_body, grnd_true_attrib, 5, max(4,
int(round(pocet_validacnich  * psti[4]))), "vali_clas_5")


arcpy.Merge_manage-
ment("vali_clas_1;vali_clas_2;vali_clas_4;vali_clas_5", "vali-
dacni_mezikrok")

pocet_vali = arcpy.GetCount_management("validacni_mezikrok")
pocet_int2 = int(float(pocet_vali.getOutput(0)))
odecist2 = pocet_validacnich - pocet_int2

select_random(zakladni_body, grnd_true_attrib, 3, odecist2,
"vali_clas_vygen3")
arcpy.Merge_management("vali_clas_vygen3;validacni_mezikrok", cesta +
"/validacni.shp")

#tady zacina klasifikace

arcpy.gp.CreateSignatures_sa(databaze + "/snimek", "trenovaci", cesta
+ "/signature_body.GSG", "COVARIANCE", grnd_true_attrib)

arcpy.gp.MLClassify_sa(databaze + "/snimek", cesta + "/signa-
ture_body.GSG", databaze + "/klas_snimek", "0.0", "EQUAL", "", "")



#hodnoceni presnosti
arcpy.gp.UpdateAccuracyAssessmentPoints_sa(databaze + "/klas_snimek",
cesta + "/validacni.shp", cesta + "/validacni_update.shp", "CLASSI-
FIED")

#chybova_matice
arcpy.gp.ComputeConfusionMatrix_sa(cesta + "/validacni_update.shp",
databaze + "/matice")

kappa = get_table_value(databaze + "/matice", 8, "Kappa")
celkova_presnost = get_table_value(databaze + "/matice", 7, "U_Accu-
racy")
zpracovatelska_presnost1 = get_table_value(databaze + "/matice", 7,
"C_1")
zpracovatelska_presnost2 = get_table_value(databaze + "/matice", 7,
"C_2")
zpracovatelska_presnost3 = get_table_value(databaze + "/matice", 7,
"C_3")
zpracovatelska_presnost4 = get_table_value(databaze + "/matice", 7,
"C_4")
zpracovatelska_presnost5 = get_table_value(databaze + "/matice", 7,
"C_5")
uzivatelska_presnost1 = get_table_value(databaze + "/matice", 1,
"U_Accuracy")
uzivatelska_presnost2 = get_table_value(databaze + "/matice", 2,
"U_Accuracy")
uzivatelska_presnost3 = get_table_value(databaze + "/matice", 3,
"U_Accuracy")
uzivatelska_presnost4 = get_table_value(databaze + "/matice", 4,
"U_Accuracy")
uzivatelska_presnost5 = get_table_value(databaze + "/matice", 5,
"U_Accuracy")
```

```
print pocet_trenovacich, pokus, kappa, celkova_presnost, zpraco-
vatelska_presnost1, zpracovatelska_presnost2, zpracovatelska_pres-
nost3, zpracovatelska_presnost4, zpracovatelska_presnost5, uzi-
vatelska_presnost1, uzivatelska_presnost2, uzivatelska_presnost3, uzi-
vatelska_presnost4, uzivatelska_presnost5
```