



Zpracování dat DPZ

semestrální projekt

Michal Kovář

ČVUT v Praze, Fakulta stavební, Katedra geomatiky

15. května 2025

Obsah

- 1 Úvod a cíle práce**
- 2 Sběr a příprava dat (Sentinel-2, CORINE, DEM)**
- 3 Trénink modelu (Random Forest)**
- 4 Vylepšování přesnosti (FI, sampling, overfitting)**
- 5 Postprocessing a export map**
- 6 Porovnání: jaro vs. jaro+létó**
- 7 Shrnutí a závěr**
- 8 Otázky a diskuse**

Úvod

- Cíl: mapa využití krajiny
- Data: Sentinel-2, CORINE, DEM
- Multitemporální data
- Porovnání: jaro vs. jaro+létó
- Algoritmus: Random Forest

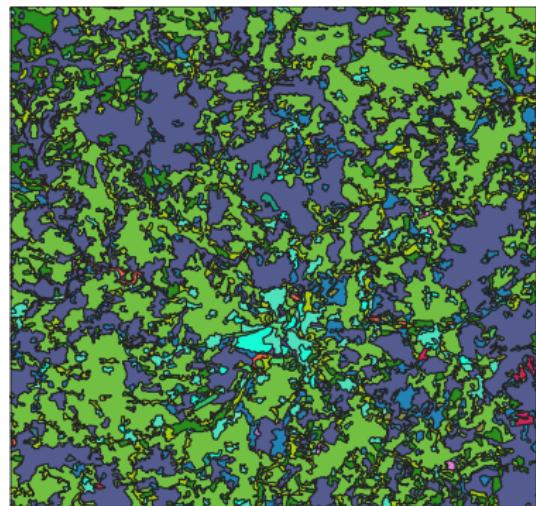
Sběr dat: Výběr lokality

- Lokalita: okolí Plzně
- Heterogenní krajina
- Sezónní změny
- Pokrytí daty: Sentinel-2,
CORINE, DEM



Sběr dat: Vstupy

- Sentinel-2: jaro + léto
- CORINE: trénovací data
- DEM: digitální model terénu



Sběr dat: AOI a preprocessing

- AOI: 50×50 km
- Ořez: Clip by mask
- Resampling: r.resample,
gdalwarp
- Sjednocení CRS, čištění



Příprava dat: Resampling a ořez

- Vstupní pásma: 10 m, 20 m, 60 m
- Převedeno na jednotné rozlišení 10 m
- Ořez podle AOI
- Automatizace pomocí Python skriptů
- Využita knihovna GDAL, funkce gdal.Warp
- Komprese výstupů: LZW

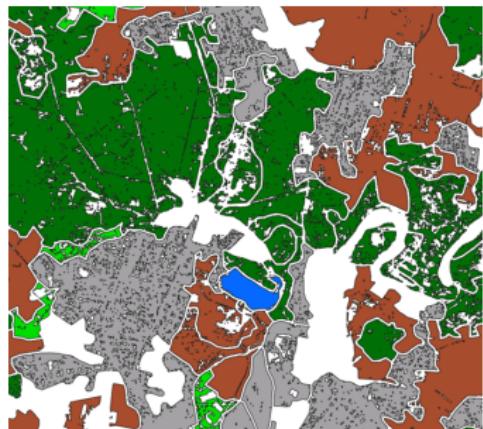
Příprava dat: Tvorba nomenklatury

- CORINE třídy byly tematicky sloučeny.
- Vznikla vlastní nomenklatura (7 tříd + Others).
- Převod pomocí CASE . . . WHEN.

| Kód | Název | CLC třídy |
|-----|-----------------|-------------------------|
| 1 | Artificial Land | 111, 112, 121, 122, 124 |
| 2 | Cropland | 211, 242 |
| 3 | Grassland | 231, 321 |
| 4 | Woodland | 311, 312, 313 |
| 5 | Shrubland | 324, 222 |
| 6 | Water bodies | 511, 512 |
| 99 | Others | 141, 142, 243, 411, 412 |

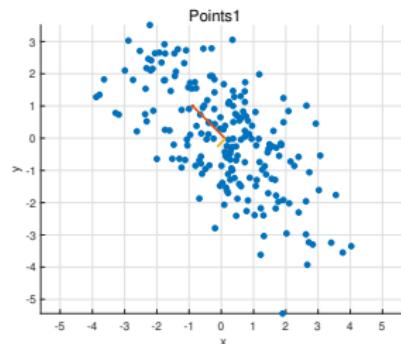
Příprava dat: Čištění polygonů

- Vypočítány indexy: **NDVI**,
NDWI
- $NDVI = (B8 - B4) / (B8 + B4)$
- $NDWI = (B3 - B8) / (B3 + B8)$
- Sloužily k odstranění
vody/vegetace z
neodpovídajících tříd
- Použito maskování, raster
kalkulkator, polygonizace
- Okraje oříznuty pomocí
Buffer (-20m)



Příprava dat: PCA a pomocné kanály

- Použity první dvě komponenty (PC1 a PC2).
- Zachycují 97 % rozptylu dat.
- Dále využity indexy NDVI a NDWI.
- DEM zahrnut jako výškový vstup.



Příprava dat: Export do CSV

- Vytvoření pravidelné bodové vrstvy (rozestup 200 m).
- Spojení bodů s rastrovými hodnotami (Sample raster values).
- Spojení bodů s třídou z CLC (Join attributes by location).
- Export do formátu CSV pro použití v modelu.

Trénink modelu: Úvod do klasifikace

- Cíl: klasifikace land cover podle vstupních dat.
- Použit algoritmus **Random Forest**.
- Výhody: vysoká přesnost, robustnost, nezávislost na rozdělení dat.
- Využita knihovna scikit-learn.

Trénink modelu: Příprava vstupních dat

- CSV obsahuje spektrální kanály, PCA, NDVI, NDWI, DEM.
- X = vstupy (bez label, fid, id).
- y = cílová proměnná (label).
- Rozděleno na trénovací/testovací set (80/20 %) se stratifikací (pro zachování poměru tříd).
- Nastaveno `test_size = 0.2, random_state = 42`

Trénink modelu: Trénování a predikce

- **Klasifikátor:** RandomForestClassifier s `n_estimators = 200, max_depth = None, n_jobs = -1.`
- **Trénink:** `rf.fit(X_train, y_train).`
- **Predikce:** `y_pred = rf.predict(X_test).`
- **Použito random_state pro replikovatelnost.**

Trénink modelu: Přesnost

■ Jaro + léto:

- Testovací přesnost: accuracy = 0.93
- Křížová validace: mean accuracy = 0.93

■ Pouze jaro:

- Testovací přesnost: accuracy = 0.92
- Křížová validace: mean accuracy = 0.92

■ Trénovací sada: accuracy = 1.00 ⇒ **overfitting**

Trénink modelu: F1-score

■ **Jaro + léto:**

- Testovací F1-score: $F1 = 0.75$
- Křížová validace: mean $F1 = 0.76$

■ **Pouze jaro:**

- Testovací F1-score: $F1 = 0.73$
- Křížová validace: mean $F1 = 0.73$

■ **Trénovací sada:** $F1 = 1.00 \Rightarrow \text{overfitting}$

Trénink modelu:

Confusion Matrix

Jaro + léto

| | | | | | |
|------|-------|------|-------|---|-----|
| 1766 | 370 | 6 | 60 | 0 | 0 |
| 171 | 20954 | 542 | 305 | 0 | 0 |
| 8 | 931 | 2093 | 135 | 0 | 0 |
| 13 | 298 | 54 | 17181 | 0 | 0 |
| 0 | 46 | 20 | 204 | 8 | 0 |
| 0 | 0 | 0 | 0 | 0 | 105 |

Pouze jaro

| | | | | | |
|------|-------|------|-------|---|-----|
| 1609 | 512 | 6 | 75 | 0 | 0 |
| 261 | 20778 | 583 | 350 | 0 | 0 |
| 9 | 1153 | 1873 | 132 | 0 | 0 |
| 24 | 368 | 48 | 17106 | 0 | 0 |
| 2 | 50 | 23 | 199 | 4 | 0 |
| 0 | 0 | 0 | 0 | 0 | 105 |

Trénink modelu:

Feature Importance (jaro + léto)

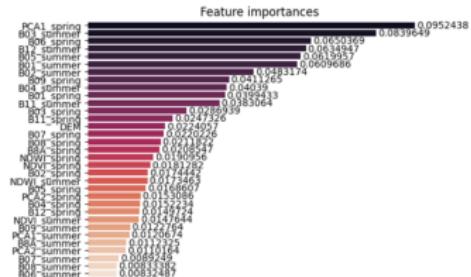
■ Nejdůležitější:

- PCA1_spring, B03_summer,
B06_spring, B12_summer

■ Nejméně důležité:

- B06_summer, B08_summer,
B07_summer, PCA2_summer

■ Letní vstupy měly obecně nižší vliv.



Trénink modelu:

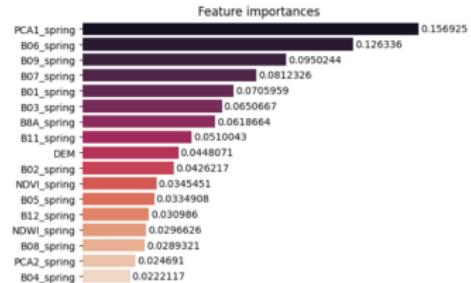
Feature Importance (jaro)

■ Nejdůležitější:

- PCA1_spring, B06_spring,
B09_spring, B07_spring

■ Nejméně důležité:

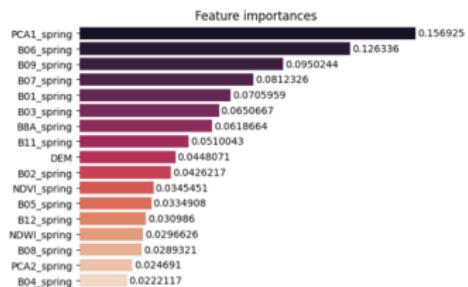
- B04_spring, PCA2_spring,
B08_spring, NDWI_spring



Vylepšování modelu:

Redukce vstupních proměnných

- Na základě hodnot z `feature_importances_` byly identifikovány méně významné vstupy.
- Odstraněno čtyři nejméně důležitá pásma.
- Model byl přeучen se zredukovanou sadou vstupních kanálů.



Vylepšování modelu:

Vyvážení trénovacích dat

- Nerovnoměrné zastoupení tříd mělo negativní vliv na klasifikaci (např. Shrubland a Water výrazně podreprezentovány).
- Pro zvýšení reprezentace minoritních tříd byl použit **oversampling**.
- Výsledkem byl vyrovnanější trénovací

Před vyvážením:

| Třída | Počet vzorků |
|-------|--------------|
| 1 | 11008 |
| 2 | 109863 |
| 3 | 15838 |
| 4 | 87730 |
| 5 | 1388 |
| 6 | 523 |

Po vyvážení:

| Třída | Počet vzorků |
|-------|--------------|
| 1 – 6 | 109863 |

Vylepšování modelu:

Závěrečné vyhodnocení výkonu

Jaro + léto:

- **Accuracy:** 0.97
- **Macro F1-score:** 0.97

| Třída | Precision | Recall | F1 |
|-------|-----------|--------|------|
| 1 | 0.97 | 0.99 | 0.98 |
| 2 | 0.98 | 0.89 | 0.93 |
| 3 | 0.92 | 0.98 | 0.95 |
| 4 | 0.98 | 0.97 | 0.98 |
| 5 | 0.99 | 1.00 | 0.99 |
| 6 | 1.00 | 1.00 | 1.00 |

Pouze jaro:

- **Accuracy:** 0.97
- **Macro F1-score:** 0.97

| Třída | Precision | Recall | F1 |
|-------|-----------|--------|------|
| 1 | 0.96 | 0.99 | 0.97 |
| 2 | 0.97 | 0.87 | 0.92 |
| 3 | 0.92 | 0.97 | 0.95 |
| 4 | 0.97 | 0.97 | 0.97 |
| 5 | 0.99 | 1.00 | 0.99 |
| 6 | 1.00 | 1.00 | 1.00 |

Vylepšování modelu:

Uložení a aplikace modelu

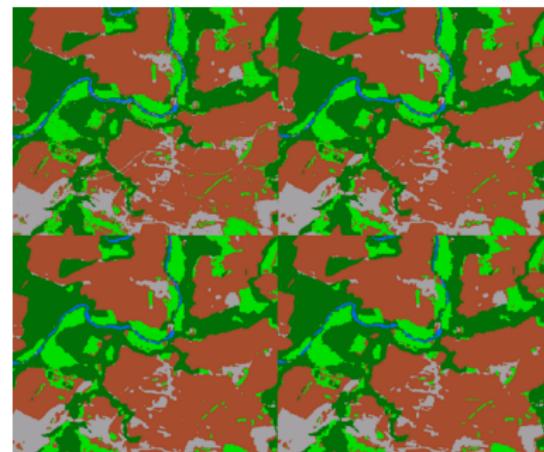
- Model uložen pomocí pickle.
- Aplikace na GeoTIFF raster
(vstup: multitemporální data).
- Pomocí GDAL + NumPy provedena predikce:
 - Načtení, převod na 2D pole
 - Predikce po dávkách
 - Rekonstrukce do rastru



Postprocessing:

Testování filtrů pro odstranění šumu

- Testování filtry:
 - **Majority** – modus hodnot v okolí (SAGA/GRASS)
 - **Median** – medián z okolí (GRASS)
 - **Sieve** – odstranění malých objektů (GDAL)
- Výsledky byly hladší, ale došlo ke ztrátě jemných liniových prvků.
- Z tohoto důvodu nebyly



Postprocessing:

Export mapy z QGIS

- Výsledky byly vizualizovány v **QGIS Layout Manageru**.
- Do mapového výstupu byly přidány základní kartografické náležitosti:
 - Legenda, měřítko, severka
 - Název mapy, autor, zájmová oblast
- Výstupní mapy byly exportovány jako PDF.

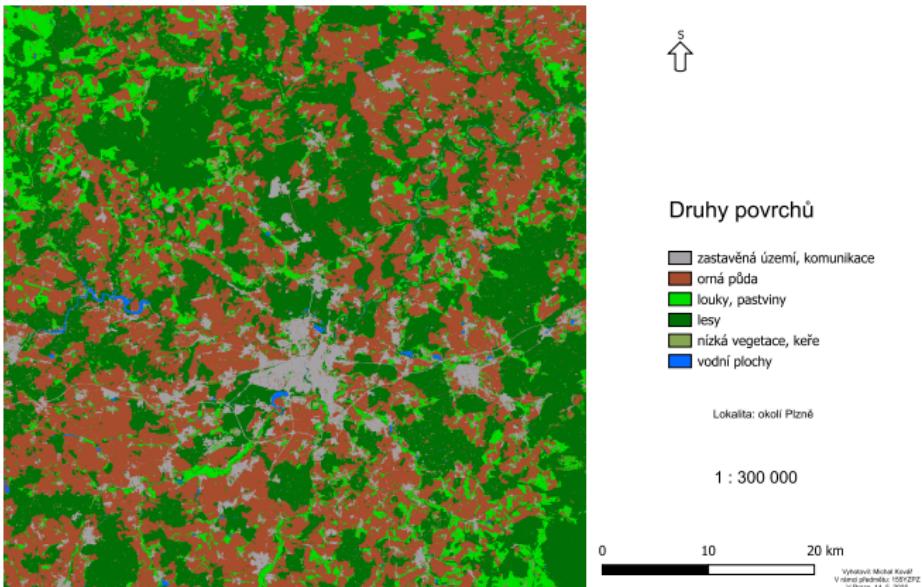
Porovnání klasifikací:

jaro vs. jaro + léto

- Cílem bylo porovnat přesnost a kvalitu klasifikace mezi:
 - **Pouze jarní scénou** – jedna sezóna
 - **Jarní + letní scénou** – multitemporální přístup
- Lepší rozlišení vegetačně proměnlivých tříd.
- Porovnání:
 - pomocí přesnostních metrik (F1, přesnost, recall)
 - vizuálně podle finálních výstupních map

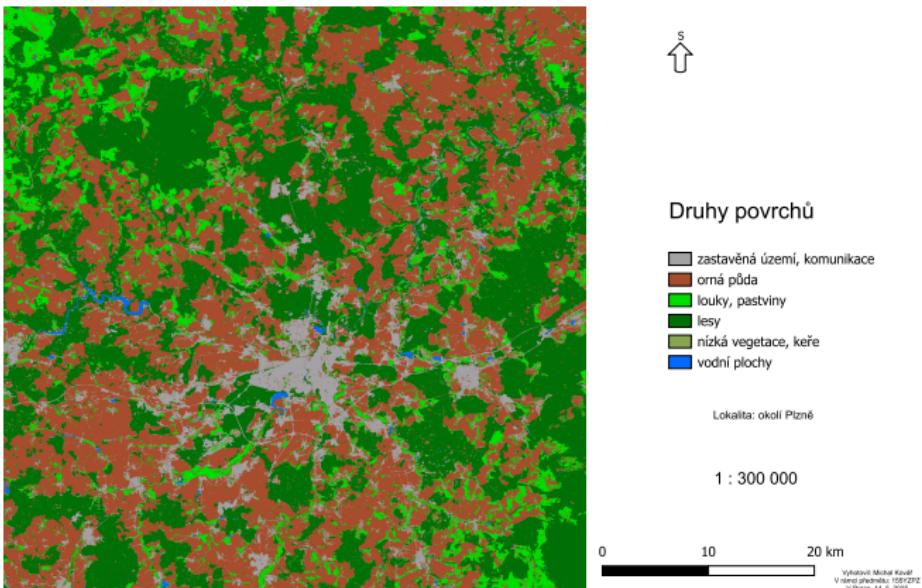
Výstup klasifikace:

Klasifikace jarní+letní scény



Výstup klasifikace:

Klasifikace jarní scény



Závěr

- Byl vytvořen funkční klasifikační model využívající data Sentinel-2, CORINE a model terénu (DEM).
- Multitemporální přístup (jaro + léto) vedl k vyšší přesnosti u sezónně proměnlivých tříd (např. orná půda).
- Přesnost klasifikace byla zvýšena pomocí:
 - redukce nevýznamných vstupních pásem
 - vyvážení trénovacích dat (oversampling)
- Model byl uložen a aplikován na celou oblast zájmu.
- Výsledkem jsou vizualizace z softwaru Qgis

Děkuji za pozornost.

Otázky

Máte nějaké dotazy?

Zdroje

Celý postup byl tvořen podle: BOUČEK, Tomáš (2024)

BOUČEK, Tomáš (2024). Zpracování dat dpz – předmět 155yzpz.

https://geo.fsv.cvut.cz/gwiki/155YZPZ_Zpracovani_dat_DPZ.online,
cit. 2025-05-14.