Strojové učení a dolování dat v geografii

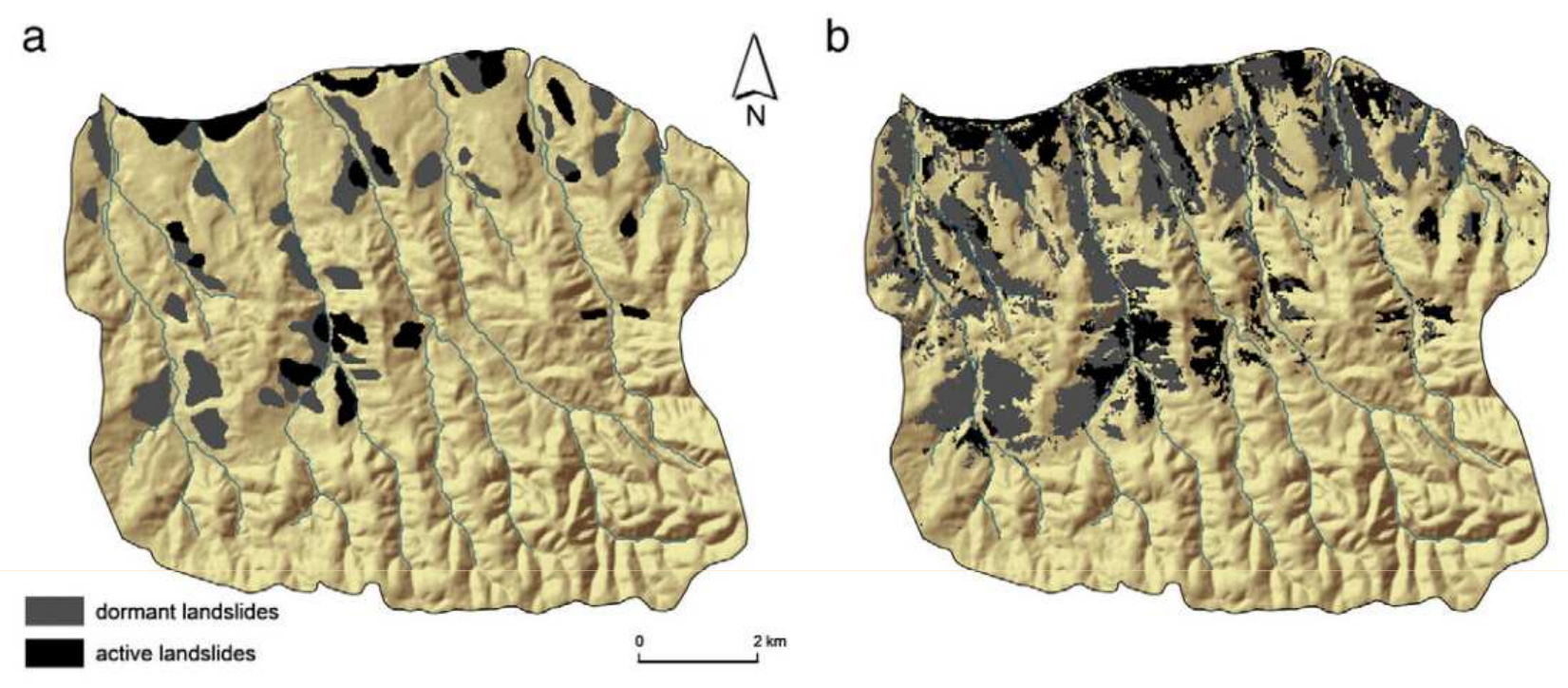
Vybrané partie dolování dat 2016/17

Jan Šimbera

simberaj@natur.cuni.cz

Kde v geografii?

- Získávání prostorově podrobných dat
  - Prostorová dezagregace
  - Analýza dat dálkového průzkumu
  - Big data processing (geotweets)
- Geomodelování
  - Přírodní rizika a procesy
  - Šíření druhů
  - Mobilita populace



GEOMODELOVÁNÍ

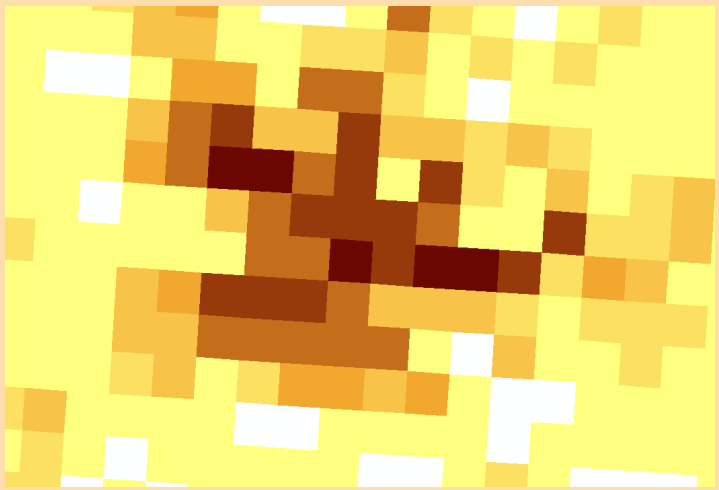
Pravděpodobnost sesuvů

- Marjanović (2014)
- Per-pixelová segmentace a klasifikace
- Topografické a geologické příznaky
  - Druh hornin, hloubka podzemní vody
  - Výška nad závěrovým profilem (= rychlost odtoku)
  - Sklon svahu, křivost, orientace, land cover
- SVM vyšlo jako nejlepší

PROSTOROVÁ DEZAGREGACE

Prostorová dezagregace

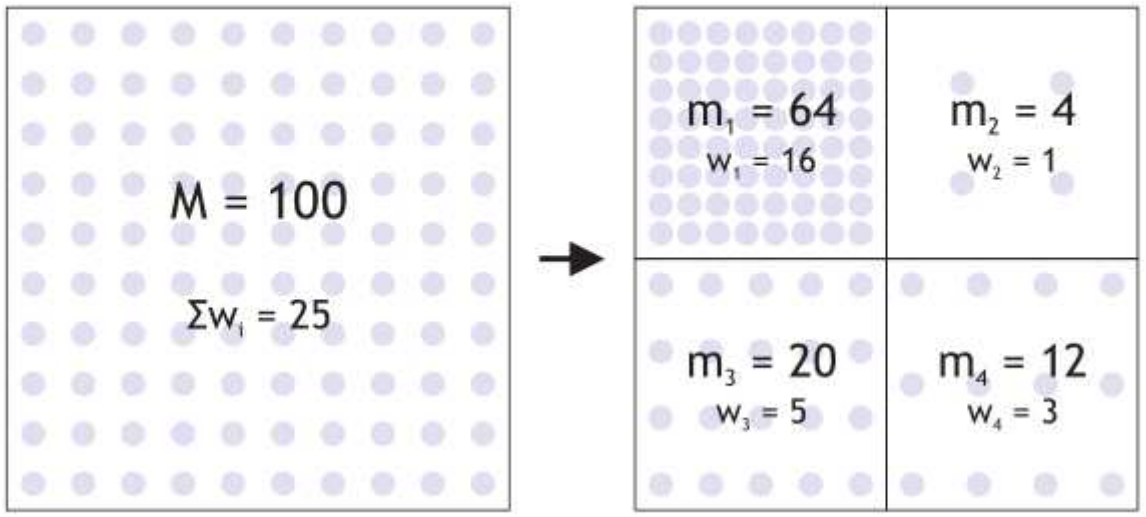
- Odhad prostorově podrobných dat
  - vstup: méně podrobná data (vyšší adm. úroveň) + pomocná data (covariates) – příznaky
- výhoda: je to model, bez problémů s OOÚ



Dezagregace – princip

- rozpočítání na hodnoty podjednotek
- váhy w_i nutno spočítat/namodelovat

$$m_i = \frac{M w_i}{\sum_i w_i}$$



Používané příznaky

- Land cover / land use
  - podíl zelených ploch, tvar, počet a plocha budov
  - hustota barů, bank, ordinací, lékáren…
- Doprava – hustota a zatížení sítí, dostupnost
- Terén – sklon a orientace svahu, nadmořská výška
- Night light imagery, ChÚ, klima, tweety…
- Problém volby okolí
  - Mnoho potenciálních příznaků
  - Dvoufázové trénování

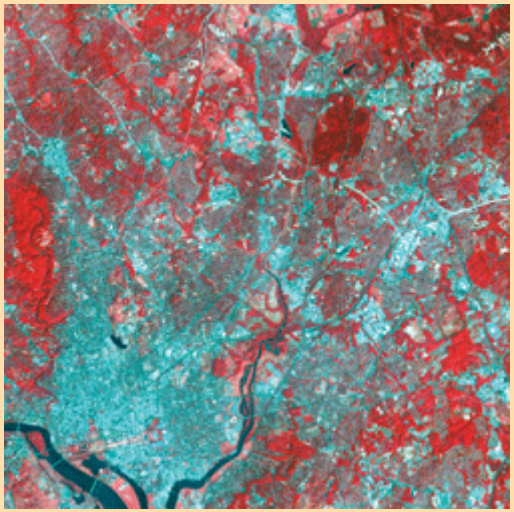
DP

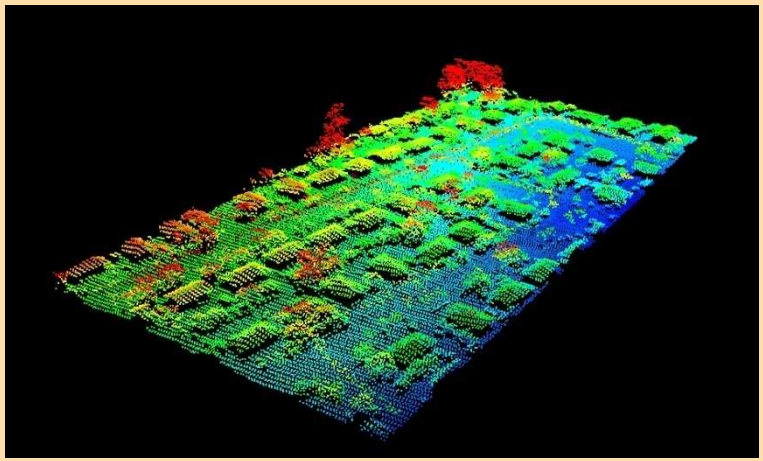
- Dezagragace hustoty zalidnění
- Geostat 1 km grid -> urban block level
- General Regression (Specht 1991)
  - trénováno na Praze (data za adresní body)
  - testováno na Lublani
  - ANN jsem nezvládl ☺
    - Nekompatibilita balíčků, backprop nefungovala
- WorldPop: random forests

DATA DÁLKOVÉHO PRŮZKUMU

Data DPZ

- satelitní a letecké snímky
  - VIS, IR, termální pásmo, radarová data
  - hyperspektrální data: šířka pásma λ – pár nm, desítky pásem -> curse of dimensionality
- letecké laserové skenování (ALS)





Data DPZ – užití

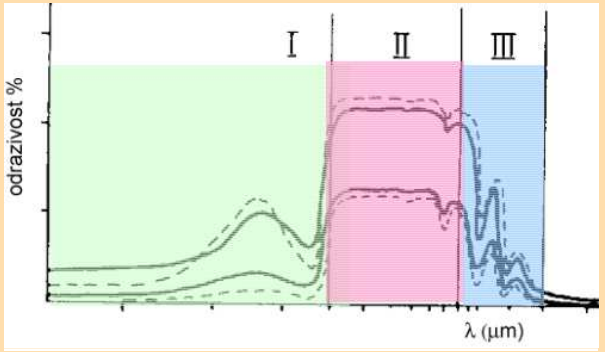
- Klasifikace land coveru
  - change detection…
- Ekologické modelování
  - přítomnost druhů / pravděpodobnost
- Detekce pohybů zemské kůry (mm přesnost)
- ALS -> digitální model terénu

…na rozdíl od klasické image analysis chceme na výstupu také 2D informaci

Land cover klasifikace – příznaky

- Hodnoty z pásem
- Spektrální indexy (NDVI)

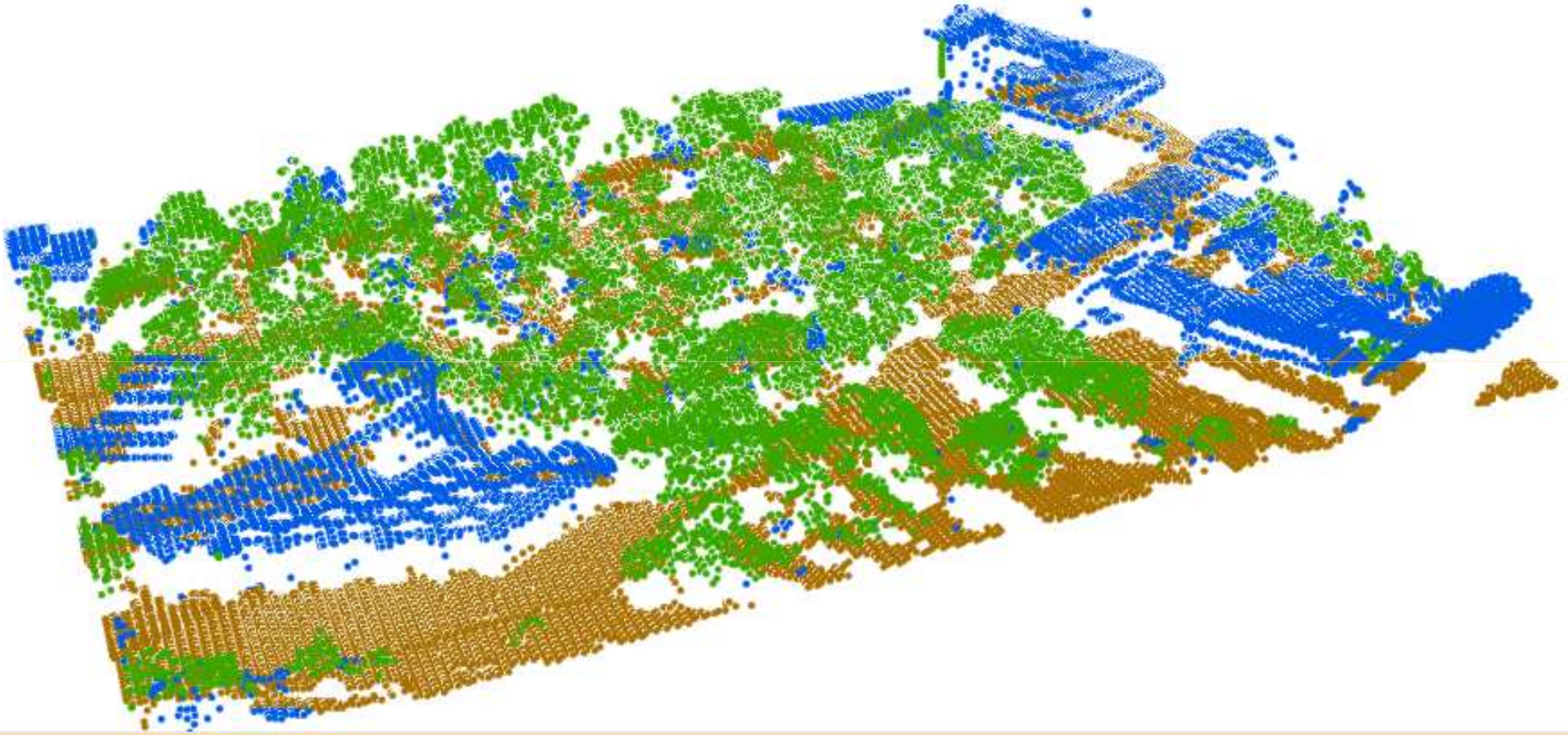
$$NDVI = \frac{NIR - R}{NIR + R}$$



- Object-based Image Analysis
  - segmentace obrazu
    - kritéria: homogenita barev + tvarová kritéria
  - objektové příznaky
    - texturální míry, heterogenita, tvar, sousedství…

ML metody pro klasifikaci

- klasická: Maximum Likelihood
- out-of-the-box cokoliv, co se najde v software (ENVI, ArcGIS, eCognition)
  - ANN (konvenční)
  - SVM
  - Random forests
- Filtrace ALS: nic
  - Pravděpodobně výpočetní náročnost…

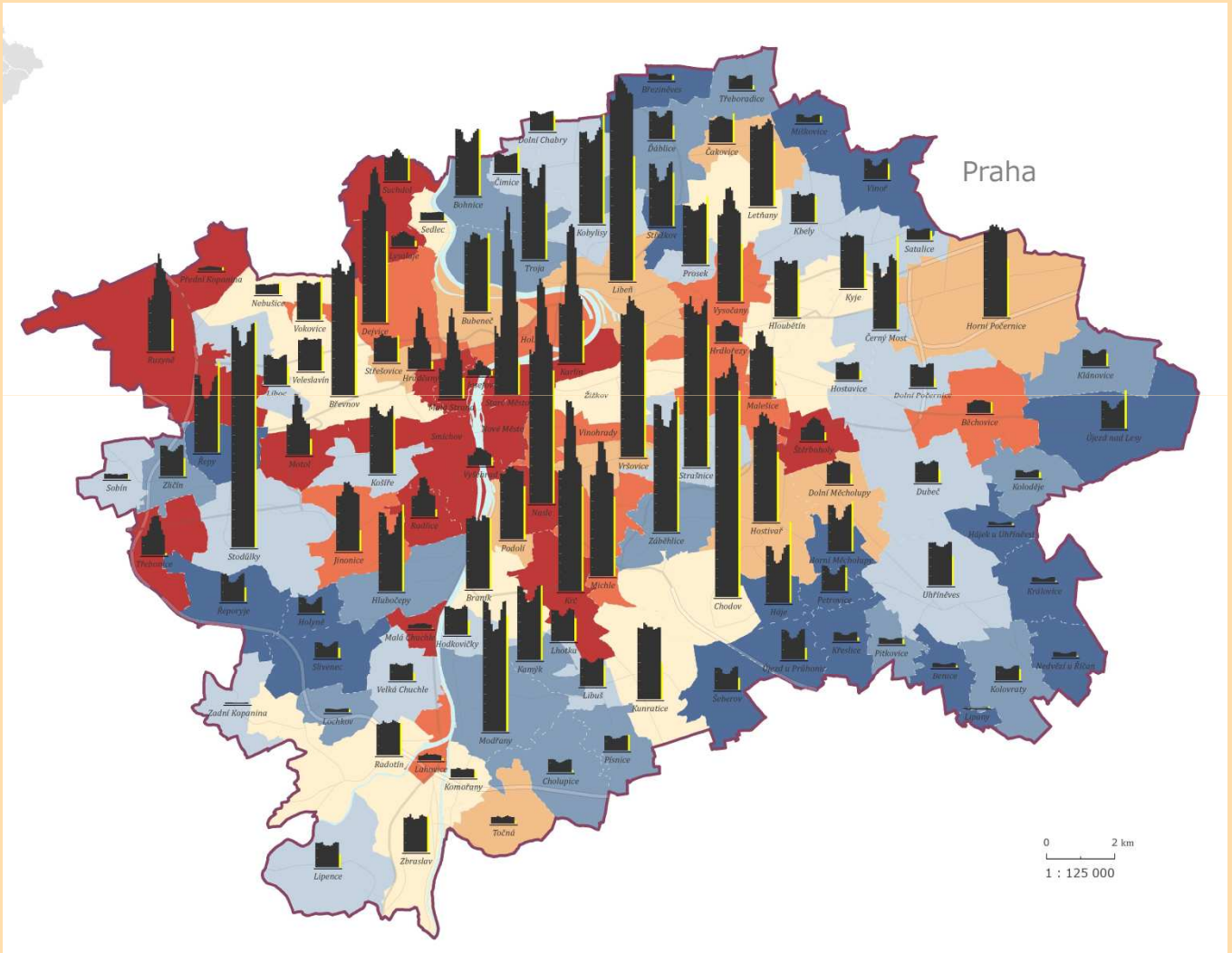


MOBILNÍ SIGNALIZAČNÍ DATA

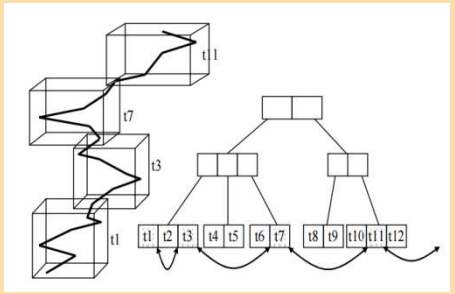
Mobilní signalizační data (CDR)

- Záznamy o aktivitě mobilních telefonů
  - Hovor, zpráva, datový paket
  - Změna buňky
  - Ověření připojení – jednou za čas (30 min–2 h)
- Na hranici big dat
  - 80 GB / 14 dní v ČR (1 operátor)
- Velké problémy s anonymizací
  - Výzkum: bez vazby na demografii
  - Zašumění a posuny času

Statická agregace – rytmy lokalit



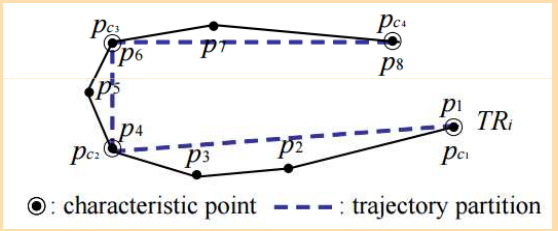
Dolování trajektorií

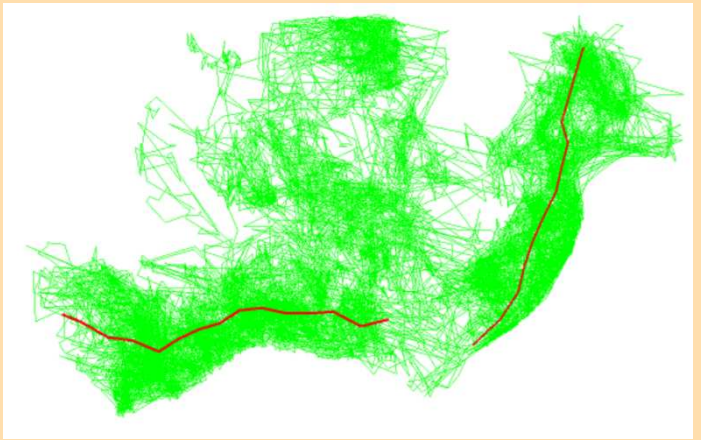


- Moving object databases
  - Rekonstrukce trajektorií
  - Časoprostorové dotazování (TB-strom)
  - Agregace do buněk
- Shlukování
  - Spektrální shlukování nad prům. vzdálenostmi
  - Sémantické – konverze dle kategorií lokací
    - Jen pro podrobnější data
  - Partition-and-group
  - Frequent itemsets (se směrovou úpravou)

Partition-and-group shlukování trajektorií

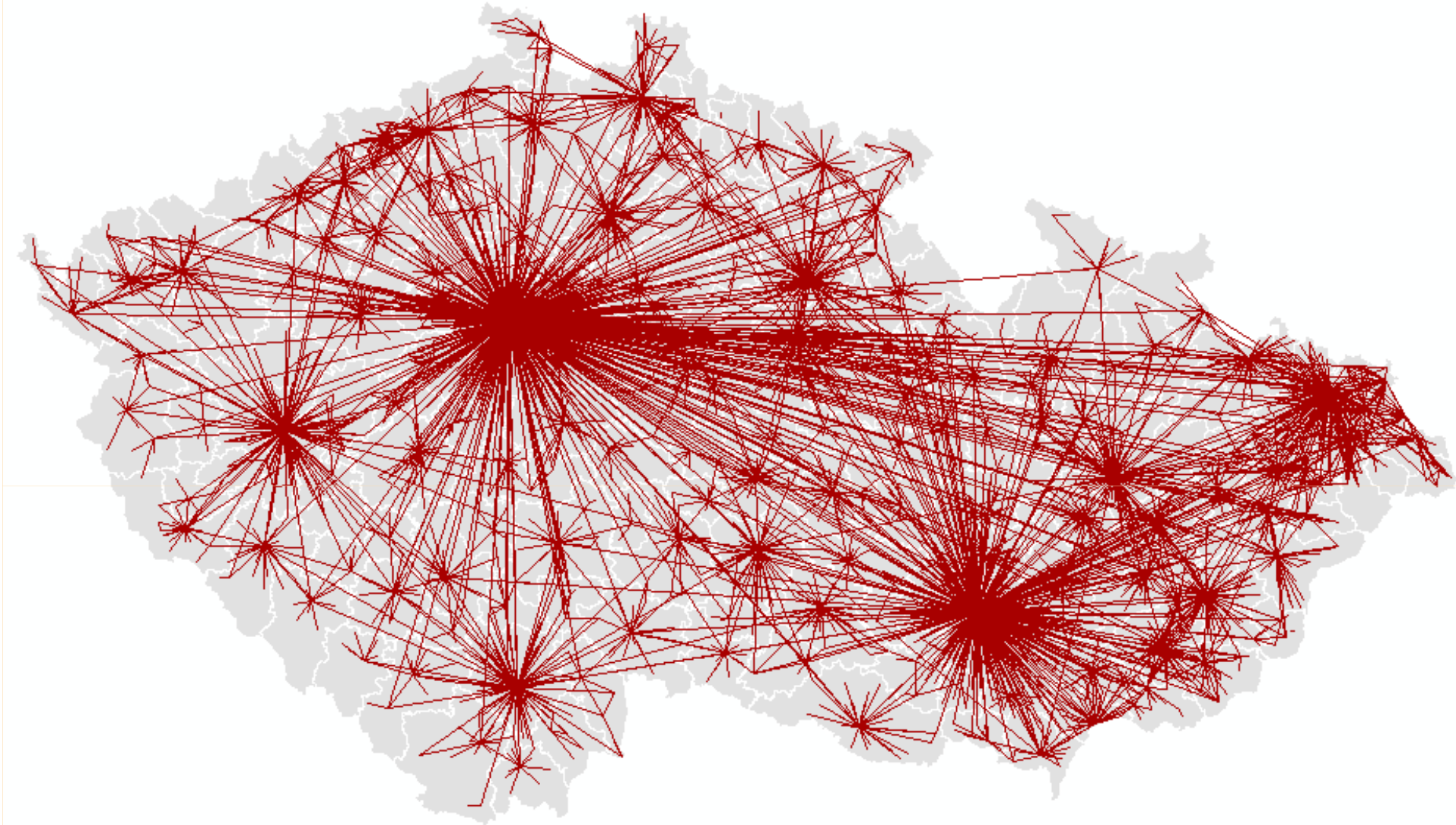
- Partition – segmentace trajektorie
  - Rozdělení dle reprezentativních bodů
    - Výrazné ohyby ~ generalizace linie
- Group – hierarchicky
  - Vzdálenostní funkce: 3 faktory
    - Kolmá vzdálenost
    - Vzdálenost ve směru
    - Rozdíl směrů
- Hlavní trajektorie





Regionalizace

- Shlukování zón dle OD matice do regionů
- Problémy
  - Nesymetrická vzdálenostní matice
  - Definice cílové funkce (při jiných přístupech)
- Subjektivní ad-hoc přístupy
  - Velmi kvalitní (expertní vhled)
  - Replikace – prostor pro ML?
- Dedikované agregační metody
- Genetické algoritmy



NA ZÁVĚR

- Převládají jednoduché metody převodu 2+D informace na 1D a aplikace běžných ML metod
- Prostor zejména u velkých dat
  - Trajektorie,