MUNI SCI

HR EXCELLENCE IN RESEARCH

ESDA
IDW, GPI

6.3.2024

EDA

- Od EDA k ESDA
- Cca v 70. letech byli researcheři frustrováni, že nemohli modelovat bez toho, aniž by se mohli dívat na data (museli přistoupit hned ke statistickému modelování. Regresní analýzy, ANOVA,…)
- John Tuckey (1977) – přišel s nápadem EDA

- GOOD (1983) – „Philosophy of science“ – skvělý článek o tom, co to vlastně EDA, a jak se liší od standardní statistiky
- Přišel s termínem „objevit potenciálně vysvětlitelný pattern“
- Nic se nevysvětlovalo, pouze se objevovaly věci, které by byly potenciálně vysvětlitelné
- Tuckey (1977) řekl, EDA je tzn. „detektivní práce“

  (zkoumáš, vyšetřuješ a učíš se o datech)
- Cílem je pomoci prozkoumat data, ještě předtím, než se vytvoří nějaké předpoklady
- Čísla ukáží hodně, ale je vždy lepší se podívat na grafy, histogram, Box plot, scatter-plot,..)

Průzkumová analýza prostorových dat
Exploratory spatial data analysis (ESDA)

- EDA +
- Přidává prostorovou stránku, což je zásadní část
- Ne pouze mapy jako výsledek, ale prostorová informace je nepostradatelná část průzkumu dat
- Sada nástrojů v prostorové statistice pro analýzy a porozumění prostorových patternů a vztahů uvnitř datasetu
- Cílem je:
  - Odstranění něčeho, co tam třeba nepatří
  - Analýza trendu a jeho případné odstranění
  - Případná transformace rozdělení vstupních dat
  - Analýza rozdělení hodnot

  - Obecně je cílem dosáhnout tzv. AHA! momentu

Čím je ESDA ojedinělá?

- Tradičně se přistupuje k objevování tak, že se první vytvoří hypotéza a poté se použijí data, které hypotézu buď potvrdí nebo vyvrátí, to je tzv. DEDUKTIVNÍ přístup
- Další způsob je přesně opačný, nejdříve se sesbírají data, ze kterých se pak dá něco vyčíst, až poté se vytvoří hypotéza, tzv. INDUKTIVNÍ přístup

- ESDA využívá tzv. ABDAKTIVNÍ přístup
- Deduktivní a indaktivní přístup jsou přesné opaky, zde se využívají oba dohromady (objevím nějaký pattern spolu s hypotézou)
- Nicméně je to neustálé testování, a objevování pomocí interakce s daty (interakce mezi průzkumem dat a lidským vnímáním)

- Typické otázky které se snažíme zodpovědět
  - Kde se určité jevy dějí? (Patterny, shluky, hot spots, cold spots, rozdíly,..)
  - Proč se dějí tam, kde se dějí? (Pomoc při rozhodování)
  - Jak ovlivňuje lokalita sledovaných jevů další aspekty? A jak kontext ovlivňuje co se děje? (Interakce)
  - Kdy by měly být určité věci lokalizovány? (Optimalizace)

Aktivity ESDA

- **Popis prostorového uspořádání dat** (dynamické mapy, grafy,…)
- **Identifikuje atypické prostorové vzorky** – prostorové outliery (Outlier Maps, Box map, percentile map,…)
- Objevuje vzory prostorové závislosti a prostorové heterogenity

- **Vizualizace dat** – grafická reprezentace a souhrn dat
  - spousta rozdílných pohledů (tabulka, graf, mapa,…)
  - klíčem k efektivní vizualizaci je kombinace různých pohledů
- **Interaktivní mapování** – interagujeme s daty, vybíráme různé pohledy, uspořádání, tak aby byla použitá grafika nástrojem k prozkoumání dat
  - hlavní koncepty dynamického mapování jsou „Linking“ a „Brushing“
- **Základní statistické grafy** – ESDA používá spoustu nápadů z EDA uzpůsobeným do vizuální analýzy a dynamické grafiky za použití hlavních konceptů (Linking, brushing)

- Jak udělat dobrou mapu?
  - Mapy dokážou „lhát“ (How to lie with maps, Mark Monmonie (2018))
  - Lidské vnímání může být jednoduše ošáleno (generalizace, měřítko, symboly, legenda, barvy, intervaly,…)

Reprezentování hodnoty

- Diskrétní – výběr intervalů
  - všechna data ve stejném intervalu mají stejnou barvu nebo odstín)
- Spojité – barevná škála
  - nefunguje příliš pro velké datasety

- ve většině případů se pracuje s diskrétními hodnotami (převedou se spojité hodnoty na diskrétní kategorie)

Barvy

- Velice důležité, ale záludné
- emoční vazby – co je špatné, co dobré?
  - ovlivňují vnímání patternů
  - červené – teplo, nebezpečí; modré – chlad

Tip: www.colorbrewer2.org

Legenda

- Sequential (Sekvenční) - data jsou seřazena (vysoké a nízké hodnoty)
- Diverging (Divergující)
  - barvy divergují od neutrální po dva extrémy
  - začíná se uprostřed a postupuje se k oběma okrajům
  - důraz není na seřazení dat, ale na ale na to, jak se pohybovat od středu
- Qualitative (Kvalitativní)
  - pro kategorie, ošemetné
  - barvy mají tendenci přiřazovat hodnotu, ale v kategoriích žádná hodnota není, všechny jsou stejně přijatelné
  - žádné řazení, žádné vysoké nízké hodnoty, jen kategorie A B C

Rozdělení hodnot

- Spojité hodnoty na diskrétní pomocí volby intervalů
- Používají se 3 hlavní rozdělení:
  - Quantile - často defaultní, ale ne vždy ideální
  - Natural breaks
  - Equal interval

Quantile – část distribuce, u kvantilu na 4 části, kvintil 5 částí, …

- data se seřadí, prvních 25% je první quartil, 25-50% je druhý, 50% je medián,.
- každý interval má stejný počet vzorků, nekontroluju rozsah hodnot v intervalech, což může dát dojem homogenity

- **Natural breaks** – založen na logice shlukování (podobně jako u K- means)

  - shlukuje na základě podobnosti hodnot, což vyústí do tzv. „breaks“ mezi skupinami, proto se tak nazává
  - je to dobré na nalezení atypických hodnot
  - rozdílný počet vzorků v kategoriích

- **Equal intervals –** analogie s histogramem

  - rozhodnu se kolik budu mít kategorií se stejnou šířkou
  - v quantile rozdělení mám čtyři kategorie, ale s totožným počtem vzorků ne se stejném rozsahem hodnot

Statistické mapy

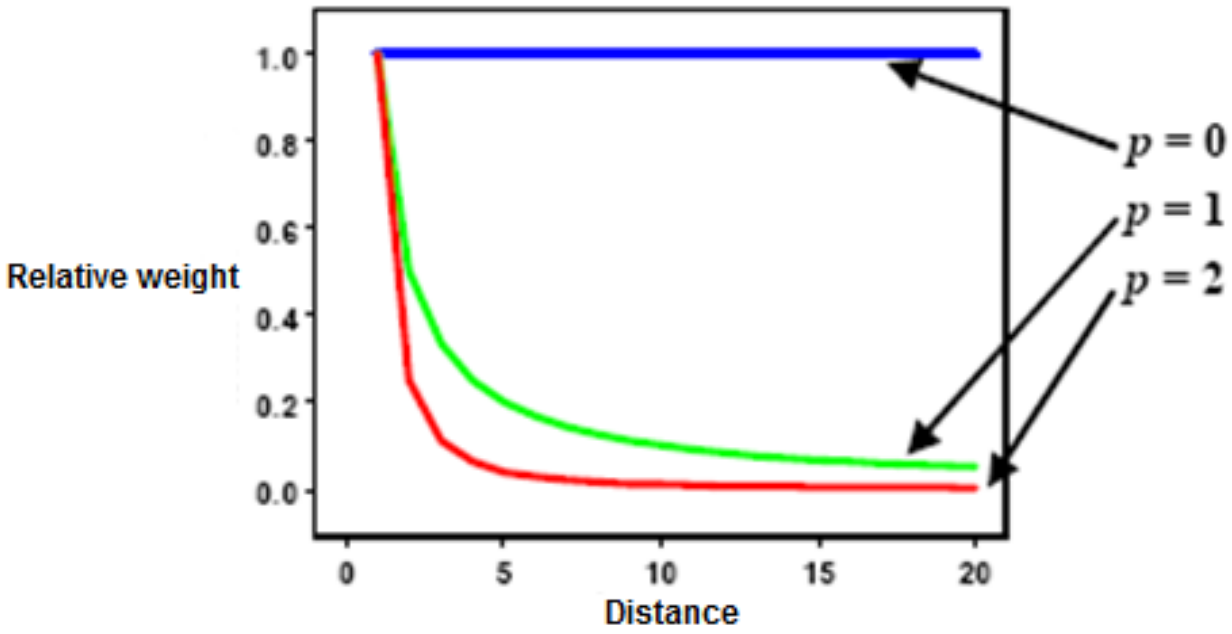
- Speciální mapy, visuální nástroje
- Outlier maps – extrémy v hodnotách
  - Percentile Map
  - Box plot/Map
  - Standard deviational map
  - Unique value maps
- Conditional Maps

- Histogram
  - Diskrétní reprezentace spojitých proměnných
- Box Plot
  - Reprezentace distribuce, kde jsou data seřazena od nízké k vysoké se zaměřením na medián, quartily a „ploty“
- Scatter plot

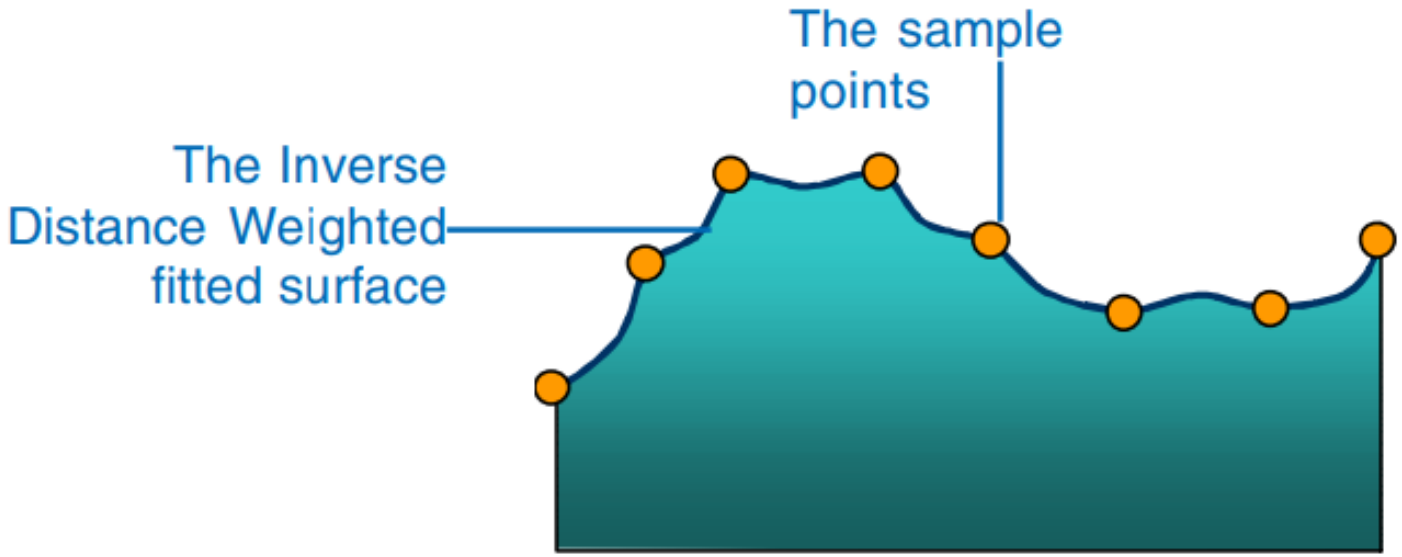
Inverse Distance Weighting (IDW)

- Lokální - deterministická - exaktní - spojitá
- Odhad nových hodnot na základě vzdálenosti. Tento odhad je založený na tom, že se usuzuje, že čím blíže si věci jsou, tím podobnější jsou. Tento předpoklad je potom promítnut do hodnoty p (power) v nastavení interpolace (přímá aplikace Toblerova zákonu – 1. zákon geografie, blízké věci k sobě mají větší vztah než vzdálené)
- Vzdálenostem nastavím váhy (jejich součet musí být roven 1, jinak over or under estimation)
- Poté roznásobím danou hodnotu udělenou vahou = interpolovaná hodnota (váha je inverzní vzdálenosti)
- Jak nastavit váhy? Jak rychle se mají váhy zmenšovat v závislosti na vzdálenosti?

- Pokud nastavíme p = 0 (stejná váha pro všechny), odhadnutá hodnota bude čistý průměr okolních vstupních hodnot. Čím vyšší p, tím větší váha bude přidělena nejbližším bodům. Příliš vysoká hodnota p však způsobuje tzv. "bulls eyes". Nejčastěji se používá p = 2

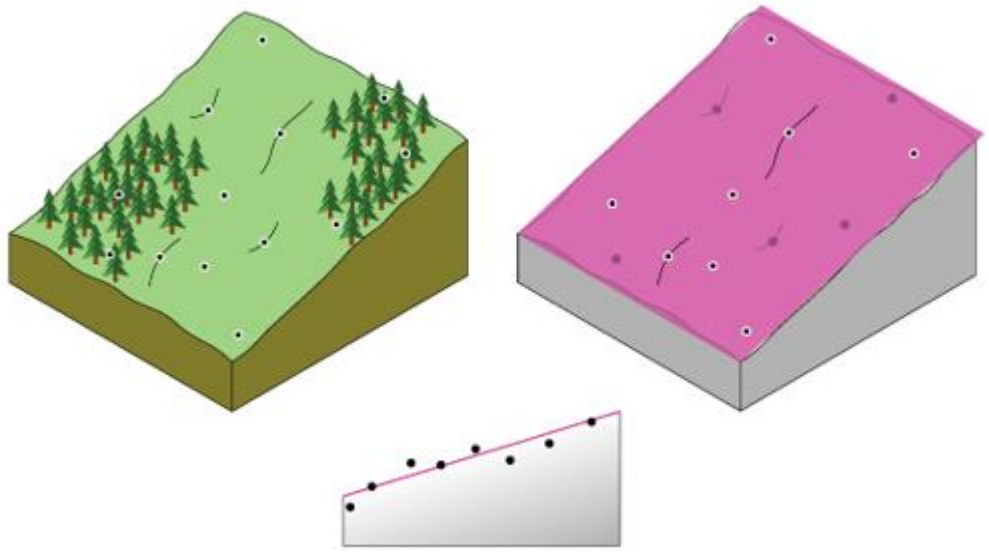


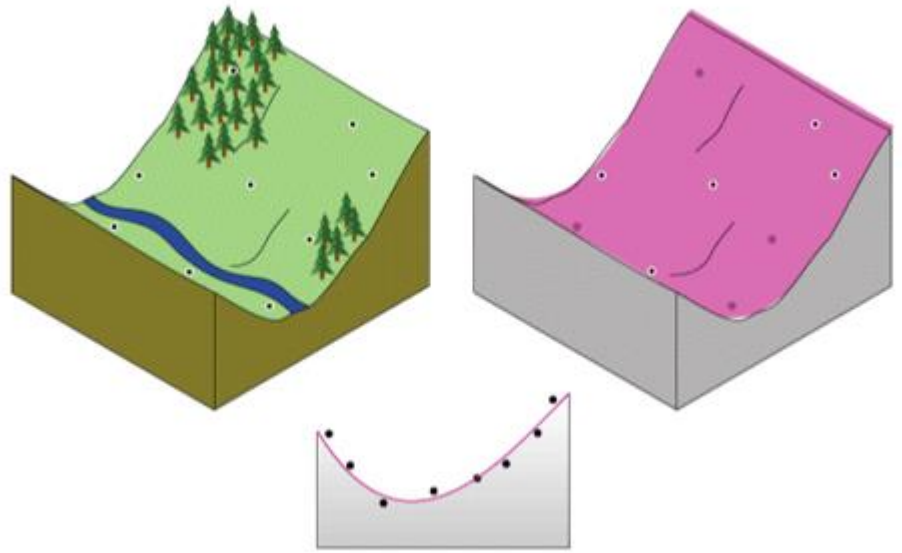
- Nevýhoda: IDW nedokáže vypočítat hodnoty vyšší nebo nižší, než jsou hodnoty vstupních
- Výhoda: Nastavuje se málo parametrů.
- Použití: Vhodné, pokud jsou vstupní data hustě a rovnoměrně rozmístěné.



GPI

- Globální - deterministická - aproximující – spojitá
- Do výpočtu odhadovaných hodnot jsou zahrnuty všechny vstupní body. Povrch je tak hladký (jako když položíme kus papíru a vložíme jej mezi vyvýšené body -zvýšený do výše hodnoty
- Nevýhoda: Nezachycuje lokální změny.
- Využití: Pokud se oblast jen velmi málo mění. Znečištění nad určitou oblastí či směr větru.





Cvičení

- Zadání: Pro měsíc srpen (8) roku 2003 vypočtěte průměrnou minimální a maximální měsíční teplotu pro všechny stanice. Vytvořte soubor typu ESRI Shapefile, nezapomeňte na souřadný systém. Zpracování proveďte v libovolném programu (MySQL, ArcMap, Open Office Calc, Google Docs Tabulky, MS Excel, QGIS, atd.). Vytvořená data prověřte pomocí průzkumové analýzy.

Postup při zpracování

- 1) Přiložená data si otevřete v MS Excel a spojte hodnoty pro měsíc srpen roku 2003 ku stanicím, které obsahují souřadnice. Tento výsledný soubor načtěte do ArcMap a vytvořte vektorovou vrstvu. Pozor na 0 hodnoty v souborech u MIN a MAX - ručně odstranit v MS Excel například.
- 2) Prozkoumejte vytvořenou vrstvu pomocí průzkumové analýzy ESDA. Projděte si všechny nástroje (histogram, qq graf...).
- 3) Nově vzniklý soubor ve formátu shp si rozdělte na dvě nové vrstvy - na trénovací a testovací. Trénovací (80 % dat) bude sloužit k výpočtu interpolace a testovací (20 % dat) poté k validaci výsledku. K tomuto využijte nástroj Subset Features (přes nabídku Geostatistical Analyst, nebo přes Toolbox).

- 4) Na trénovacích datech vytvořte pomocí metody IDW mapu prostorového rozložení minimální a maximální teploty vzduchu v měsíci srpnu roku 2003. Spatial Analyst -> Geostatistical Wizard -> Inverse Distance Wieghted.
- **Vyzkoušejte:**
  - Při tvorbě povrchů zkuste experimentovat s nastavením jednotlivých parametrů (to, co tyto parametry ovlivňují je vždy popsáno při zakliknutí daného parametru dole .

- 5) FInální spojité povrchy validujte pomocí testovacích dat pomocí nástroje GA Layer To Points.
- **Vyzkoušejte:**
  - Zkuste si chybu nějakým způsobem vizualizovat - například přes symobology dané vrstvy nebo třeba opětovnou interpolací (ale tentokrát interpolujeme atribut "error"). Dobrým způsobem může být například "graduated symbols" v symobology - zde si ale nějakým způsobem musíme poradit s tím, že defaultně nám to záporné chyby bude ukazovat nejmenšími symboly a kladné největšími - to však nebude příliš intuitivní. Můžeme si tedy například tuto vrstvu duplikovat, vypočítat si absolutní hodnoty chyb a k nim poté přidat labels z původní vrstvy (tak abychom viděli jestli se jedná o zápornou nebo kladnou chybu).