

Úvodní informace ESDA, IDW, GPI

Základy Geostatistiky - cvičení

26.2.2025

Podmínky zápočtu

— Absence

- cvičení povinná, povolena jedna neomluvená absence
- při absenci na cvičení - mail předem

— Vypracování a uznání protokolů

- celkem budou 3
- školní PC / vlastní počítač
- zadání postupně v průběhu semestru
- pozdní odevzdání = srážka bodů
- čas na vypracování 2 týdny, odevzdávací otevírací doba do 23:59
- pro udělení bodů za protokol je potřeba splnit minimálně zadání

— Dva zápočtové testy

- první cca v polovině semestru, druhý ke konci

— Bodování

- **Protokoly** - 5 bodů za každý
- za pozdní odevzdání do dvou dnů po limitu 2,5 bodů
- za pozdní odevzdání nad dva dny po limitu 0 bodů
- **Zápočtové testy** - 2,5 bodů za test (pravděpodobně 9. 4. 2025 a 7. 5. 2025)
- Celkem je tedy možné dosáhnout 20 bodů

Hodnocení	Body	Hodnocení
	20 - 18,5	A
	18 - 16,5	B
	16 - 14,5	C
	14 - 12,5	D
	12 - 10,5	E
	10 a méně	F

— Plánovaný harmonogram

Jaro	Z6101 Základy Geostatistiky
	Seminář
středa 19.2.	
středa 26.2	Úvodní informace, ESDA, IDW, GPI
středa 5.3.	ESDA in Geoda
středa 12.3.	LPI, RBF - zadání 1. protokolu
středa 19.3	Kriging
středa 26.3.	Cokriging - zadání 2. protokolu
středa 2.4.	Popisná statistika bodových objektů
středa 9.4.	1. zápočtový test , Statistický popis prostorového uspořádání bodů
středa 16.4.	Prostorová autokorelace bodových a plošných jevů
středa 23.4.	Hot-spot analýza a prostorová regrese - zadání 3. protokolu
středa 30.4.	
středa 7.5.	2. zápočtový test
středa 14.5.	
	odpadá

- **Software pro tento předmět**
 - ArcGis Pro
 - Geoda
- **Užitečné odkazy**
 - Geostatistika a prostorová interpolace – Ježek
 - [Geostatistical Analyst - ArcGis Pro](#)
 - [GeoDa](#)

Exploratory Data Analysis (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

Typické metody EDA

- Deskriptivní statistika (průměr, medián, směrodatná odchylka, rozptyl)
- Vizualizace dat (histogramy, boxploty, scatterploty)
- Korelační analýza
- Identifikace odlehlých hodnot

- 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 induktivní 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)

— Aktivita 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 použita grafika nástrojem k prozkoumání dat
 - hlavní koncepty dynamického mapování jsou „Linking“ a „Brushing“
- **Základní statistické grafy** – ESDA používá spoustu nápadů z EDA upraveným do vizuální analýzy a dynamické grafiky za použití hlavních konceptů (Linking, brushing)

Na co si dát pozor

- 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 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 homogeneity

- **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 nazývá
 - 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

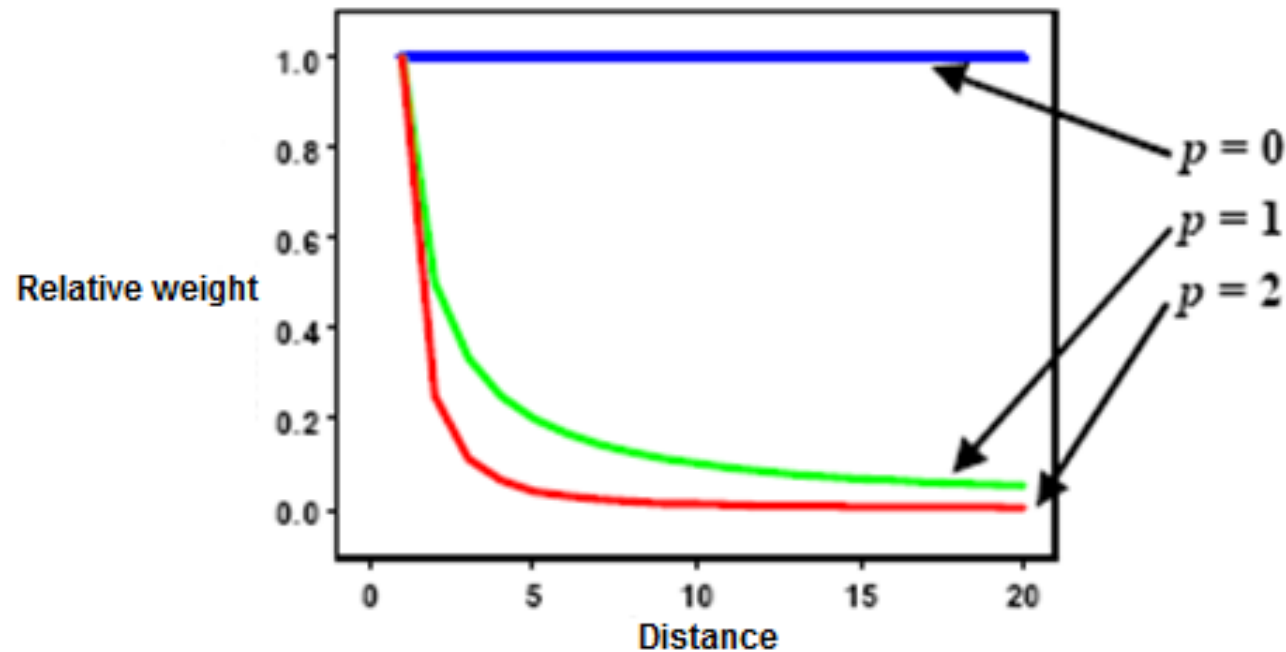
- Speciální mapy, vizuální nástroje
- Outlier maps – extrémní hodnoty
 - Percentile Map
 - Box plot/Map
 - Standard deviational map
 - Unique value maps
- Conditional Maps

- 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

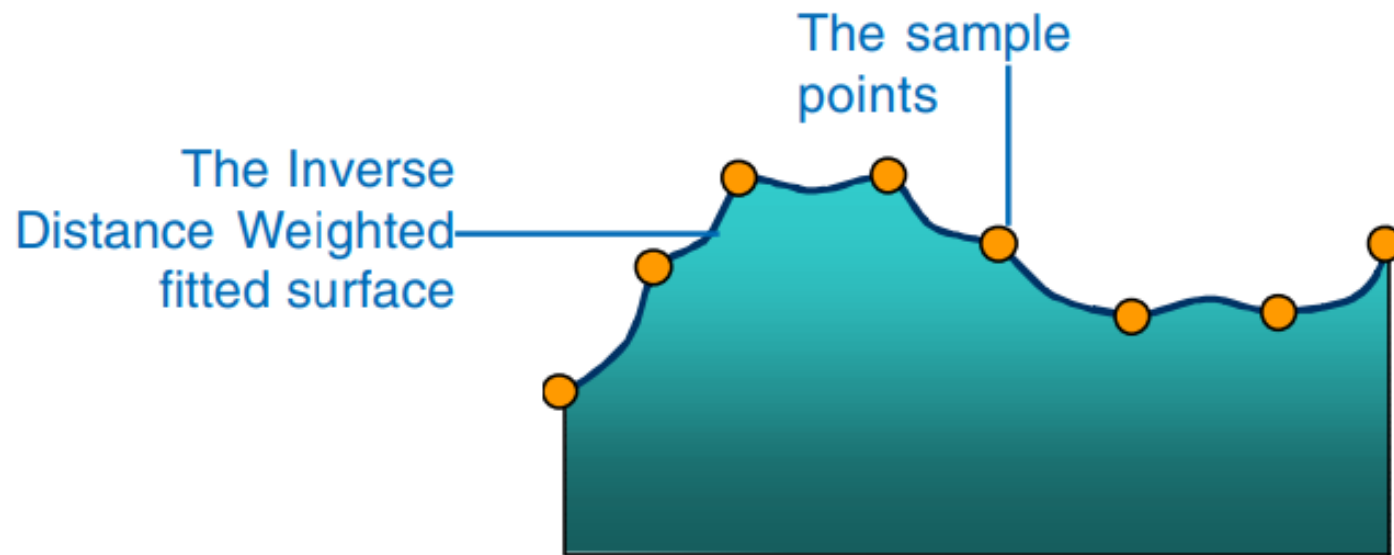
Inverse Distance Weighting (IDW)

- 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.



— Kdy použít IDW?

- Data mají lokální variabilitu (prostorová závislost klesá s rostoucí vzdáleností)
- Dostatek vzorkovacích bodů (Pokud je bodů málo nebo jsou nerovnoměrně rozmístěné, může interpolace vytvářet nerealistické výsledky)
- Data bez extrémních hodnot (hodnoty v interpolačním povrchu nemohou překročit minimum nebo maximum vstupních hodnot, extrémní hodnoty mohou silně ovlivnit interpolaci)
- Hladké povrchy bez výrazné směrové tendence
- Rovnoměrné rozmístění vzorků

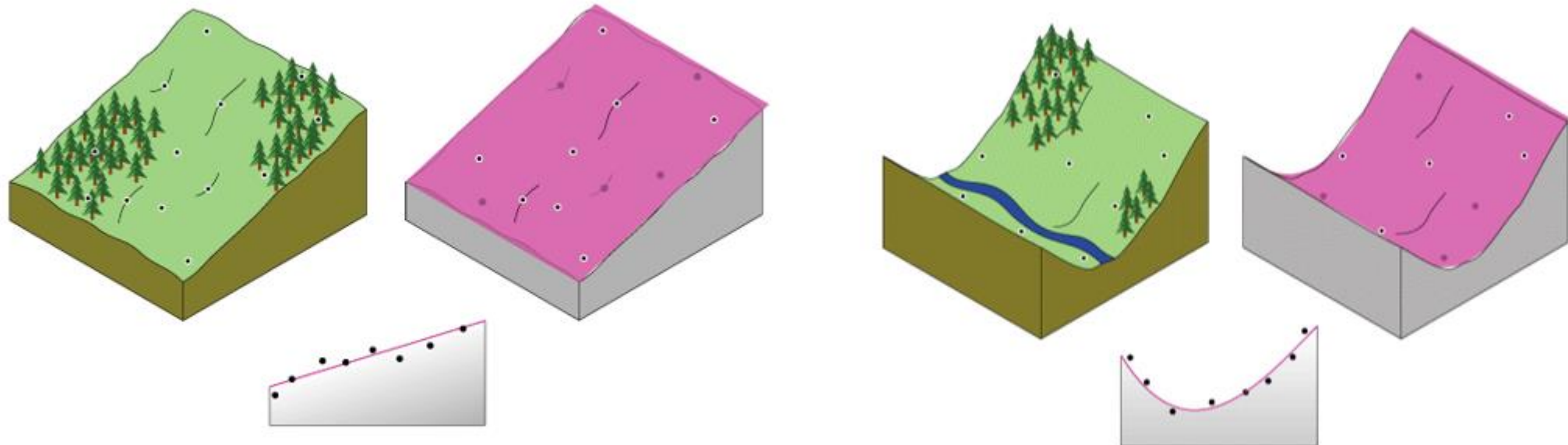
— Kdy IDW nepoužít?

- Data mají silný trend nebo anizotropii
- Nerovnoměrná hustota vzorků
- Povrchy s výraznou variabilitou (pokud očekáváte povrch s ostrými přechody nebo diskontinuitami (například výškový model s terasami nebo srázy)
- Velké oblasti s komplexními prostorovými vztahy

- Příklady použití:
 - Interpolace srážek nebo teploty v malé oblasti
 - Předpověď půdní vlhkosti
 - Interpolace hladiny podzemní vody

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.



– Kdy použít GPI?

- vhodná pro modelování globálních trendů v datech ((např. velkoplošné změny v nadmořské výšce, teplotě, znečištění)
- rychlá, ale nevhodná pro lokální variabilitu
- funguje pro hladká data, ale může být ovlivněna okrajovými hodnotami

– Kde se GPI používá?

- Geologie a geomorfologie – modelování velkoplošných změn terénu nebo geologických formací
- Meteorologie a klimatologie – interpolace teplot, tlaku nebo srážek v širokých oblastech
- Ekologie a environmentální vědy – modelování rozložení znečištění nebo změn v půdním pokryvu

Cvičení

- **Zadání:** Pro měsíc srpen (8) roku 2003 vypočtete průměrnou minimální a maximální měsíční teplotu pro všechny stanice. Vytvořte soubor typu ESRI Shapefile, nezapomeňte na souřadný systém. Zpracování proved'te 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 ke stanicím, které obsahují souřadnice. Tento výsledný soubor načtete do ArcGis 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 Wiegthed.
- **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 symbology dané vrstvy nebo třeba opětovnou interpolací (ale tentokrát interpolujeme atribut "error"). Dobrým způsobem může být například "graduated symbols" v symbology - 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).