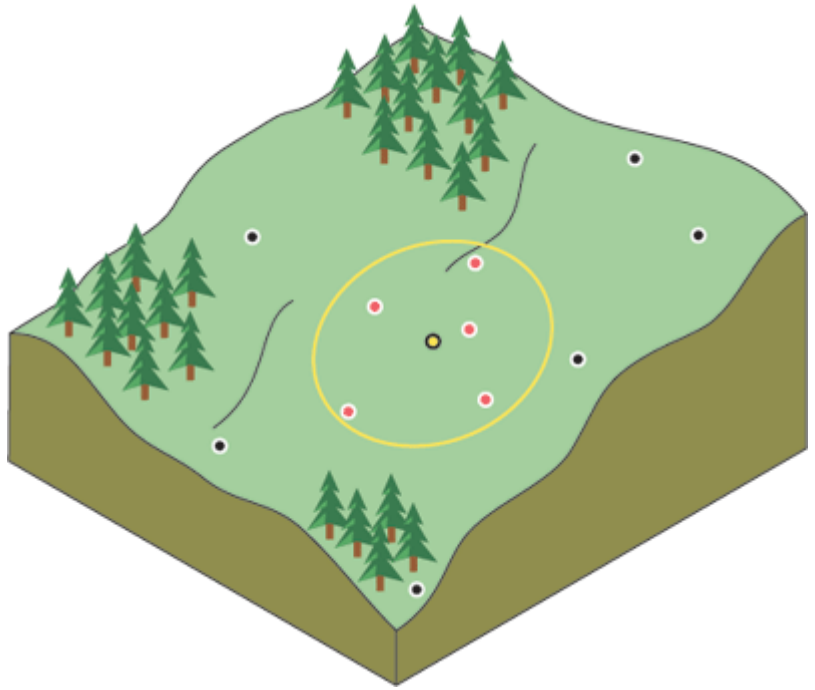
MUNI SCI

HR EXCELLENCE IN RESEARCH

ESDA v GeoDa 5.3.2025

(Opakování z minule) IDW

- Určuje hodnoty buněk pomocí lineárně vážené kombinace vzorkovacích bodů
- Váha je funkcí inverzní vzdálenosti
- Interpolovaný povrch by měl odpovídat prostorově závislé proměnné
- Metoda předpokládá, že vliv proměnné klesá s rostoucí vzdáleností od vzorku
- Např. při interpolaci teploty má vzdálenější meteostanice menší vliv, protože teplota se mění s lokálními podmínkami

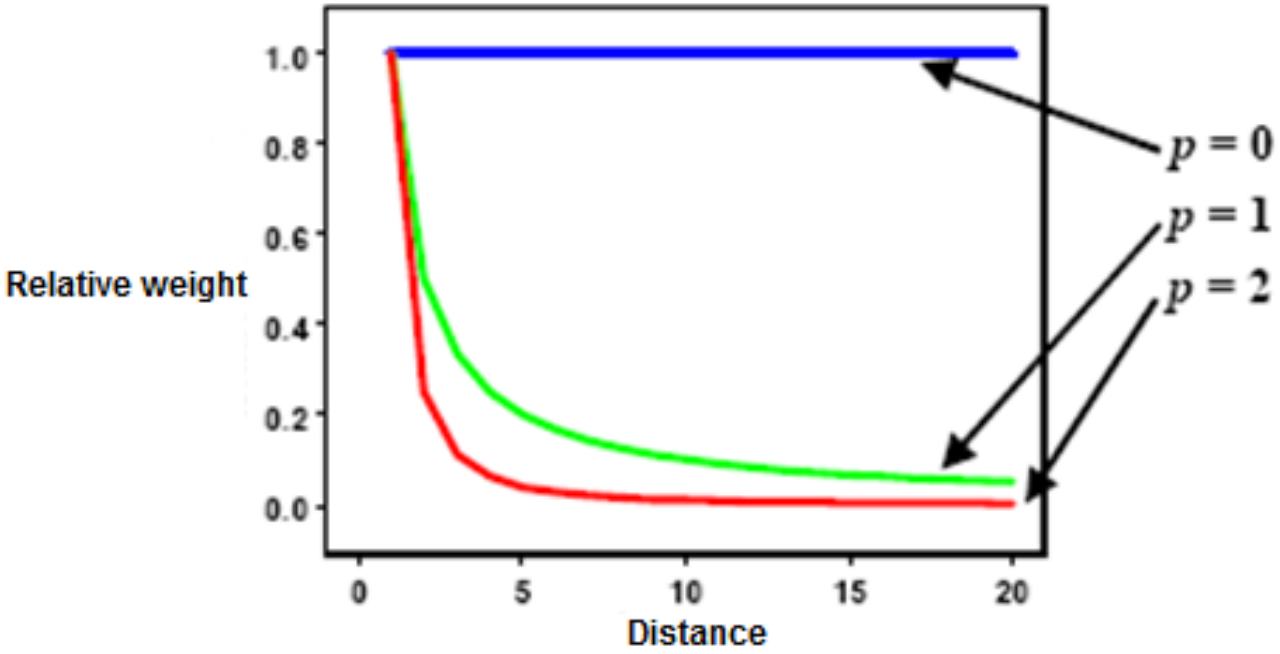


IDW neighborhood for selected point

Ovlivnění interpolace parametrem Power

- **Power** určuje vliv známých bodů na interpolované hodnoty podle jejich vzdálenosti
- Vyšší hodnota **P** zvýrazní vliv nejbližších bodů → detailnější (méně hladký) povrch
- Nižší hodnota **P** dá větší váhu vzdálenějším bodům → hladší povrch
- IDW není založeno na fyzikálním procesu, proto nelze určit „správnou“ hodnotu
- Optimální hodnotu lze najít minimalizací střední absolutní chyby

- 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



Omezení bodů pro interpolaci

- Výběrem menšího počtu vstupních bodů lze ovlivnit interpolovaný povrch
- Omezení bodů zrychlí výpočet
- Vzdálené body mohou mít nízkou nebo žádnou prostorovou souvislost, takže je lze vynechat
- Lze nastavit pevný počet bodů nebo poloměr, v němž budou body zahrnuty do interpolace

Proměnlivý vyhledávající poloměr (Variable search radius)

- Pokud je zvolen, mění se poloměr pro vyhledávání bodů pro každou buňku, v závislosti na hustotě měřených bodů poblíž interpolované buňky
- Hustě osídlené oblasti mají menší poloměr, řídce osídlené větší
- Lze nastavit maximální vzdálenost, kterou poloměr nepřekročí
- **Obecně se menší oblasti nebo minimální počet bodů používají, pokud jev vykazuje velkou variabilitu**

Pevný vyhledávací poloměr (Fixed search radius)

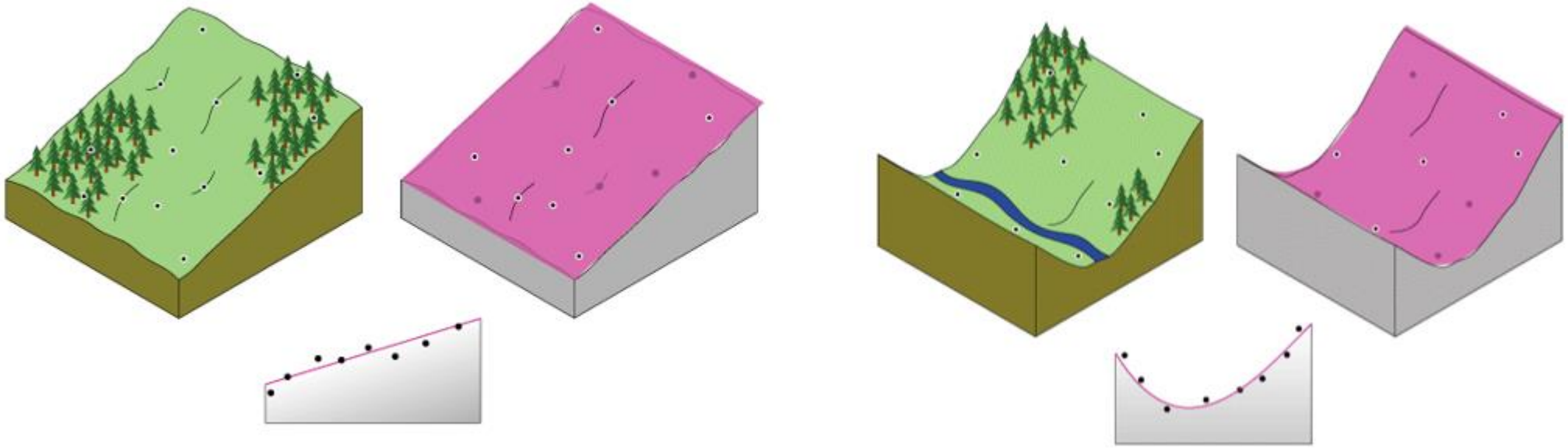
- Stejný poloměr se použije pro každou interpolovanou buňku
- Použijí se všechny body uvnitř tohoto poloměru
- Pokud je bodů méně než minimální požadovaný počet, poloměr se zvětší
- V různých částech mapy může být použitý jiný počet bodů, protože body nemusí být rozmístěné rovnoměrně

Použití bariér (Using barriers)

- Bariéra je liniová vrstva, která omezuje vyhledávání vstupních bodů
- Funguje jako hranice (např. útes, hřeben, řeka), která brání interpolaci přes ni
- Použijí se jen body na stejné straně bariéry jako interpolovaná buňka
- Pomáhá lépe modelovat přirozené překážky v krajině a zvyšuje přesnost interpolace

GPI

- Globální - deterministická - aproximující – spojitá
- Rovinná plocha (bez „ohybu papíru“) je polynom prvního řádu (lineární)
- Jeden ohyb odpovídá polynomu druhého řádu (kvadratickému)
- Dva ohyby třetímu řádu (kubickému) a tak dále; je možné použít až 10 ohybů.



(navázání z minule) 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 ke stanicím, které obsahují souřadnice. Tento výsledný soubor načtěte 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 Wieghted.
- **Vyzkoušejte:**
  - Při tvorbě povrchů zkuste experimentovat s nastavením jednotlivých parametrů (to, co tyto parametry ovlivňují je vždy popsáno při zakliknutí daného parametru dole .

IDW

- Input features: ***Location_TRAIN***
- Z value fiels: ***TMAX***
- Výstupy: ***GL a Raster***
- Power: Vyšší hodnota **p**, Blízké body mají větší vliv na interpolovanou hodnotu => Výsledná interpolace je ostřejší a lokálně detailnější.

- **Search neighborhood**: Určuje, které okolní body budou použity pro výpočet interpolované hodnoty v určitém místě.
- Tento parametr určuje, z jakého okolí budou body vybrány a jak budou rozděleny do sektorů.
- One sector:
  - Všechny body jsou brány najednou bez rozdělení na sektory.
  - Použije se pouze maximální vzdálenost a počet bodů v okolí.
  - Pokud jsou body nerovnoměrně rozmístěny, může dojít k převaze bodů z určité oblasti, což může vést k nesymetrické interpolaci.
  - Vhodné pro rovnoměrně rozmístěná data.

4 Sectors

- Prostor je rozdělen na čtyři sektory (kvadranty) ve tvaru elipsy.
- Z každého sektoru je vybrán určitý počet nejbližších bodů (obvykle alespoň jeden).
- Výhoda: Minimalizuje dominanci bodů z jednoho směru a vede k rovnoměrnější interpolaci.
- Vhodné pro případy, kdy data mají preferovaný směr variability (např. podél řeky nebo horského hřebene).

4 Sectors with 45 offsetem

- Funguje stejně jako předchozí možnost, ale sektory jsou otočeny o 45 stupňů.
- Výhoda: Pomáhá, pokud hlavní směr variability není zarovnán se standardním rozdělením sektorů.
- Vhodné pro případy, kdy jsou vzory prostorových dat diagonální nebo mají šikmou orientaci.

- Weight field:
  - Umožňuje dát některým bodům větší důležitost než jiným.
  - Čím větší hodnota váhy, tím větší vliv daného bodu na výslednou interpolaci.
  - Hodí se, když jsou některé měření spolehlivější než jiná nebo když chceme reflektovat hustotu vzorkování.

- 5) FInální spojité povrchy validujte pomocí testovacích dat pomocí nástroje GA Layer To Points.

- **GA Layer To Points** (Geostatistical Analyst)
  - (Co je to Geostatistical Layer?)
  - Pro odhad hodnot tam, kde nejsou změřené, nebo pro ověření předpovědí tam, kde už nějaká měření máme
  - Vstupy:
    - (GA layer) ***Locations_IDW_TRAIN_Tmax***
    - (Body) ***Locations_TEST***
  - Fields to validate on: ***Tmax***

- **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).

GPI

- Input features: ***Locations_TRAIN***
- Z value field: ***Tmax***
- Outputs: ***GL a raster***
- Order of polynomial: (vizualizujte si rezidua)
- Weight field:
  - Umožňuje dát některým bodům větší důležitost než jiným.
  - Čím větší hodnota váhy, tím větší vliv daného bodu na výslednou interpolaci.
  - Hodí se, když jsou některé měření spolehlivější než jiná nebo když chceme reflektovat hustotu vzorkování.

- **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.

- **GA Layer To Points** (Geostatistical Analyst)
  - Vstupy:
    - (GA layer) ***Locations_GPI_TRAIN_Tmax***
    - (Body) ***Locations_TEST***
  - Fields to validate on: ***Tmax***

ESDA v GeoDa

- Stáhněte si data „Syr“
- Je to dataset z oblasti v New Yorku, kde je zaznamenán výskyt leukémie (závislá proměnná) a několika nezávislými proměnnými (prediktory)
- Budeme zkoumat jestli existuje závislost mezi výskytem leukémie a expozicí trichlorethylenu (TCE)

Postup řešení

- 1) Načtěte data („Syr.shp“) do GeoDa
- 2) Otevřete a prozkoumejte data pomocí tabulky
  - Máme dvě nezávislé proměnné:
    - **Cases** (Počet případů leukémie v letech 1978–1982)
    - **Z** (Logaritmicky transformovaná míra výskytu leukémie, normalizovaná podle počtu obyvatel v okrsku) – Proč ji provádíme?

- Nezávislé proměnné (prediktory)
  - **PEXPOSURE -** potenciální expozice TCE (čím vyšší hodnota, tím bližší kontakt s chemikálií)
  - **PCTAGE65P -** procento obyvatel nad 65 let (může naznačovat dlouhodobou expozici)
  - **PCTTOWNHOME -** procento vlastnického bydlení (možný sociálně-ekonomický faktor)
- Používají se k vysvětlení nebo předpovědi závislé proměnné
- Faktory, které můžou způsobit změnu v závilé proměnné

Linking and brushing

- Kliknutím do tabulky se zvýrazní oblast na mapě a obráceně
- Ctrl + Click – lze vybrat větší oblast = Brushing

- Jaká je AREAKEY oblasti s nevyšším výskytem leukemie? Kolik tam žije lidí? A kolik procent lidí v této oblasti žije ve vlastním?

- 3) Zobrazte vybrané tematické mapy z nabídky „Map“ pro jednu nebo více proměnných, např. **PCTAGE65P**
  - Porovnejte kvantilové, percentilové, box a přirozené intervaly (natural breaks) mapy.
  - Prozkoumejte, jakým způsobem zobrazují stejnou tematiku různými způsoby

Quantile map

- rozděluje data na stejné počty jednotek v jednotlivých třídách

  (každá třída obsahuje přibližně stejný počet prvků)
- Jak se to dělá?
  - Seřadí se data od nejmenší po největší hodnotu
  - Rozdělení dat do určitého počtu tříd

Percentile

- Rozdělení hodnot podle percentilů
  - 1. percentil → Nejnižší 1 % hodnot.
  - 10. percentil → 10 % okrsků má nižší hodnoty než tento bod.
  - 50. percentil (medián) → Polovina hodnot je pod a polovina nad.
  - 90. percentil → Pouze 10 % okrsků má vyšší hodnoty.
- Zvýrazňuje extrémní hodnoty
- Nerozděluje data rovnoměrně (např. jako kvantilová mapa)

Box

- Založená na box-plotu (vizualizace rozložení dat, kde je zvýrazněn medián, kvartily a odlehlé hodnoty)
- hodnoty se rozdělí do kategorií podle **kvartilů (Q1, Q2, Q3)** a **extrémních hodnot**
- Používá **mezikvartilové rozpětí (IQR = interquartile range)** k určení hranic běžných a extrémních hodnot (IQR=Q3−Q1)
- Hinge = násobek mezikvartilového rozpětí
- **Odlehlé hodnoty** jsou hodnoty mimo rozsah **Q1 - 1.5 × IQR** nebo **Q3 + 1.5 × IQR**

Natural breaks

- Rozděluje hodnoty do skupin podle přirozených „skoků” v datech
- Používá **metodu Jenks Natural Breaks**, která minimalizuje vnitřní rozptyl ve třídách a maximalizuje rozdíly mezi nimi

Zobrazení souvislostí mezi dvěma proměnnými (Bivariate)

- Histogram (Explore -> Histogram – proměnná PCTAGE65P)
- Scatter Plot (Explore -> Scatter Plot – Y PCTAGE65P, X PCTOWNHOME)
- Co nám ty grafy říkají?

Kartogram

- Map -> Cartogram (velikost kruhu - PCTAGE65P, barva kruhu – Z = transformovaný výskyt leukémie)

Vztahy mezi více faktory (Multivariate)

- Scatterplot Matrix (Explore ->Scatterplot Matrix)
- Vybereme (PCTOWNHOME, PCTAGE65P, PEXPOSURE, Z)
- Zobrazí se nám vztahy mezi zvolenými proměnnými

Clustering

- Chceme identifikovat **přirozené shluky oblastí** na základě vybraných faktorů (např. věk obyvatel, expozice chemikáliím, výskyt leukémie)

K-means

- Cluters -> K-Means
- PCTOWNHOME, Počet klusterů 4
  - nesleduje geografickou polohu, pouze atributy

- **Geographically-compact k-means**
  - prostorově souvislé shluky