MUNI SCI

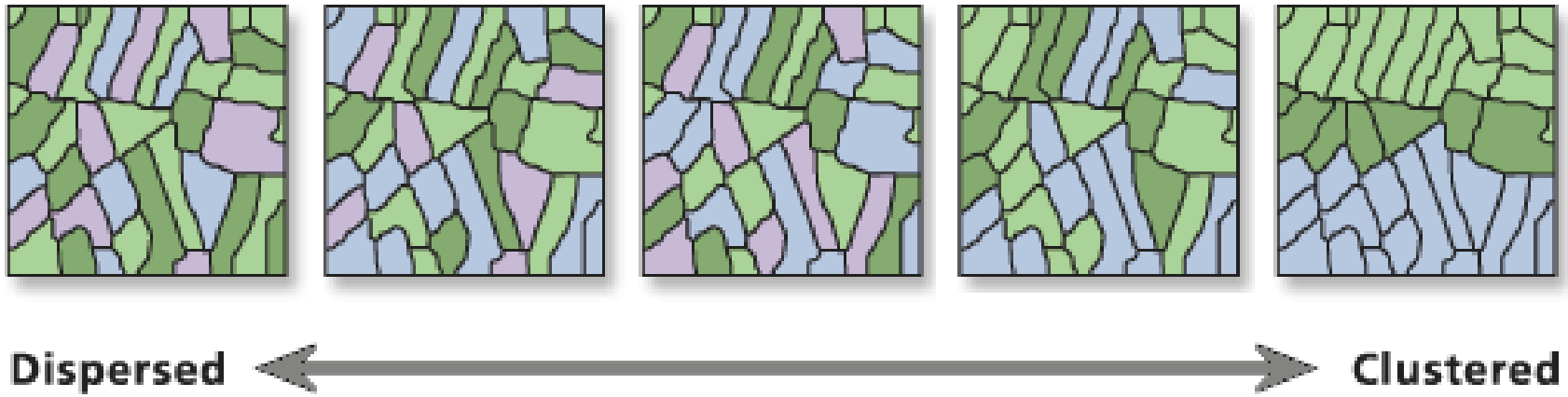
HR EXCELLENCE IN RESEARCH

Prostorová autokorelace bodových a plošných bodů

17.4.2024

Prostorová autokorelace

- Měří se na základě polohy bodů a hodnot atributů pomocí statistiky Globálního Moranova indexu, což je v podstatě ukazatel prostorové autokorelace
- Má hodnoty od -1, což značí perfektní rozptyl do +1, což naopak značí dokonalou korelaci, 0 značí náhodný prostorový vzor
- Na základě vlastností prvků a přidruženého atributu vyhodnotí, zda je vyjádřený vzor shlukovaný, rozptýlený nebo náhodný
- Nástroj v ArGis vypočítá hodnotu Moranova Indexu I a také z-score (velké záporné nebo kladné hodnoty značí pravděpodobnost přítomnosti významné prostorové autokorelace) a p-value (statistická významnost toho indexu, nízká hodnota znamená, že je málo pravděpodobné, že bysme pozorovali takovou míru prostorové podobnosti mezi hodnotami v datech, pokud by skutečně byly rozděleny náhodně)



Jak se to počítá?

- Nejprve se vypočítá průměr a rozptyl pro hodnocený atribut
- Poté se pro každou hodnotu prvku odečte střední hodnota, čímž se vytvoří odchylka od průměru.
- Vypočtou se tzv. Křížové produkty (cross-producs) odchylek (Hodnoty odchylek pro všechny sousední prvky, např. prvky v zadaném pásmu vzdálenosti, se násobí dohromady)
- Všechny cross-products se sečtou
- Vypočítá se součet druhých mocnin odchylek pro každou proměnnou
- A pak se vypočte Moranův index I pomocí tohoto vzorce

$$I = \frac{n}{W} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{X})(x_j - \bar{X})}{\sum_{i=1}^{n}(x_i - \bar{X})^2}$$

N je počet objektů v prostoru
Wji je prostorová váha pro pár objektů
W je součet všech prostorových vah
X je zájmový atribut
$\bar{X}$ je průměrná hodnota zájmového atributu

Příklad:

- Máme dvě proměnné A a B na konkrétních místech a průměr všech hodnot těchto proměnných je 10
- čím větší je odchylka od průměru, tím větší je křížový výsledek.

Feature values		Deviations		Cross-products
A=50	B=40	40	30	1200
A= 8	B=6	-2	-4	8
A=20	B=2	10	-8	-80

MUNI SCI
HR EXCELLENCE IN RESEARCH

- Pokud mají hodnoty z datasetu tendenci se prostorově shlukovat (tzn. že vysoké hodnoty se shlukují blízko jiných vysokých hodnot a nízké hodnoty se shlukují poblíž jiných nízkých hodnot), bude Moranův index kladný
- Pokud vysoké hodnoty odpuzují jiné vysoké hodnoty a mají tendenci být blízko nízkým hodnotám, bude Index záporný
- Pokud kladné hodnoty křížových výsledků vyvažují záporné hodnoty křížových výsledků, index se bude blížit nule
- Čitatel je normalizován rozptylem tak, aby hodnoty Indexu byly mezi -1,0 a +1,0

MUNI SCI
HR EXCELLENCE IN RESEARCH

- Poté, co nástroj Prostorová autokorelace (Global Moran's I) vypočítá hodnotu Indexu, vypočítá hodnotu očekávaného indexu
- Poté se porovnají očekávané a pozorované hodnoty indexu
- Vzhledem k počtu prvků v datové sadě a celkovému rozptylu hodnot dat, vypočítá nástroj z-skóre a p-hodnotu, které nám řeknou, zda je tento rozdíl statisticky významný či nikoli
- Hodnoty indexu nelze interpretovat přímo; lze je interpretovat pouze v kontextu nulové hypotézy (která zde tvrdí, že analyzovaný atribut je náhodně distribuován mezi prvky ve sledované oblasti)

MUNI SCI
HR EXCELLENCE IN RESEARCH

Interpretace výstupů:

- **P-value není statisticky významná**
  - nemůžeme zamítnout nulovou hypotézu, je totiž dost pravděpodobné, že prostorové rozložení hodnot prvků je výsledkem náhodných prostorových procesů
- **P-value je statisticky významná a Z-score je kladné**
  - můžeme zamítnout nulovou hypotézu, prostorové rozložení vysokých hodnot a/nebo nízkých hodnot v souboru dat je více prostorově seskupené, než by se očekávalo, kdyby základní prostorové procesy byly náhodné
- **P-value je statisticky významná a Z-score je záporné**
  - můžeme zamítnout nulovou hypotézu, prostorové rozložení vysokých hodnot a nízkých hodnot v souboru dat je více prostorově rozptýlené, než by se očekávalo, kdyby základní prostorové procesy byly náhodné. Rozptýlený prostorový vzor často odráží určitý typ konkurenčního procesu – prvek s vysokou hodnotou odpuzuje jiné prvky s vysokými hodnotami; podobně prvek s nízkou hodnotou odpuzuje ostatní prvky s nízkými hodnotami.

MUNI SCI
HR EXCELLENCE IN RESEARCH

- Výstupy z nástroje Spatial Autocorrelation:
  - Moranův Index I
  - Očekávaný index
  - Rozptyl
  - Z-score
  - P-value
- Ve výstupní zprávě nebo HTML reportu s grafickým shrnutím výsledků
- Osvědčené doporučení a postupy:
  - Výsledky nejsou spolehlivé s méně než 30 prvky
  - Zvolit vhodnou konceptualizaci prostorových vztahů: (ArcgisPro dokumentace)
    - ***Inverse distance and Inverse distance squared*** - nejvhodnější pro spojitá data nebo pro modelování procesů, kde čím blíže jsou dva prvky v prostoru, tím je pravděpodobnější, že se vzájemně ovlivňují nebo se navzájem ovlivňují, velká výpočetní náročnost pro velké datasety
    - Fixed distance band – vhodné pro bodová data, existuje spousta doporučení
    - Zone of indifference – pokud mám vhodnou fixed –distance, ale ostré hranice mezi sousedy nepředstavují přesnou reprezentaci dat
    - K nearest neighbors – když chci zajistit minimální počet sousedů pro analýzu
    - Contiguity edges only – pro polygony
    - Contiguity edges corners – pro polygony
    - Get spatial weights from file

MUNI SCI
HR EXCELLENCE IN RESEARCH

- Volba optimální fixní vzdálenosti (ArcGis pro dokumentace):
  - Všechny funkce by měly mít alespoň jednoho souseda
  - Žádná funkce by neměla mít všechny ostatní vlastnosti jako soused
  - Zejména pokud jsou hodnoty pro vstupní pole zkosené, chceme, aby každý prvek měl přibližně osm soused
- Měli bychom standardizovat řádky? (Standardizace)
  - Pro polygony to budeme chtít téměř vždy

MUNI SCI
HR EXCELLENCE IN RESEARCH

Cvičení

- Stáhněte a rozbalte si přiloženou geodatabázi
- Vrstva, ze které budeme vycházet je **LOND_drugs**
- **Zadání:**
  - Na základě podkladových dat pro cvičení (kriminalita spojená s drogami; Crime_type = ‘Drugs’) vypočtěte a interpretujte hodnoty indexů prostorové autokorelace.
  - **Globální indexy**
    - Incremental Spatial Autocorrelation
    - Spatial Autocorrelation (Morans I)
    - Multi-Distance Spatial Cluster Analysis (Ripleys K Function)
  - **Lokální indexy**
    - Cluster and Outlier Analysis (Anselin Local Moran's I)
    - Hot Spot Analysis tool calculates the Getis-Ord Gi*
    - Jejich optimized verze

MUNI SCI
HR EXCELLENCE IN RESEARCH

Postup vypracování

- Globální indexy
  - Před výpočtem jednotlivých indexů si první musíme data upravit
  - Použijeme nástroj **Collect Events**, který nám ze vstupní bodové vrstvy udělá novou bodovou vrstvu ve které nám body, jenž měly stejné souřadnice, spojí do jednoho a v atributové tabulce ve sloupečku ICOUNT vypíše kolik takových bodů se tam nacházelo (ICOUNT 1 znamená, že na daném místě byl pouze jeden bod - jeden drogový zločin, ICOUNT 5 pak znamená, že na daném místě zločinů proběhlo pět). Tento atribut pak budeme využívat jako vstupní atribut k výpočtu jednotlivých indexů

  - Jednotlivé nástroje se v ArcGis nachází v toolboxu **Spatial Statistics Tool - Analyzing Patterns** pro globální indexy a **Spatial Statistics Tool - Mapping Clusters** pro lokální indexy.
  - Každý nástroj v sobě zahrnuje užitečnou nápovědu a také vysvětlení jednotlivých vstupních parametrů při kliknutí na ně. Pro bližší informace a pomoc s interpretací indexů doporučuji stránky ArcGis, které se dopodrobna věnují jednotlivým nástrojům:

Arcgis dokumentace - Spatial statistics toolbox

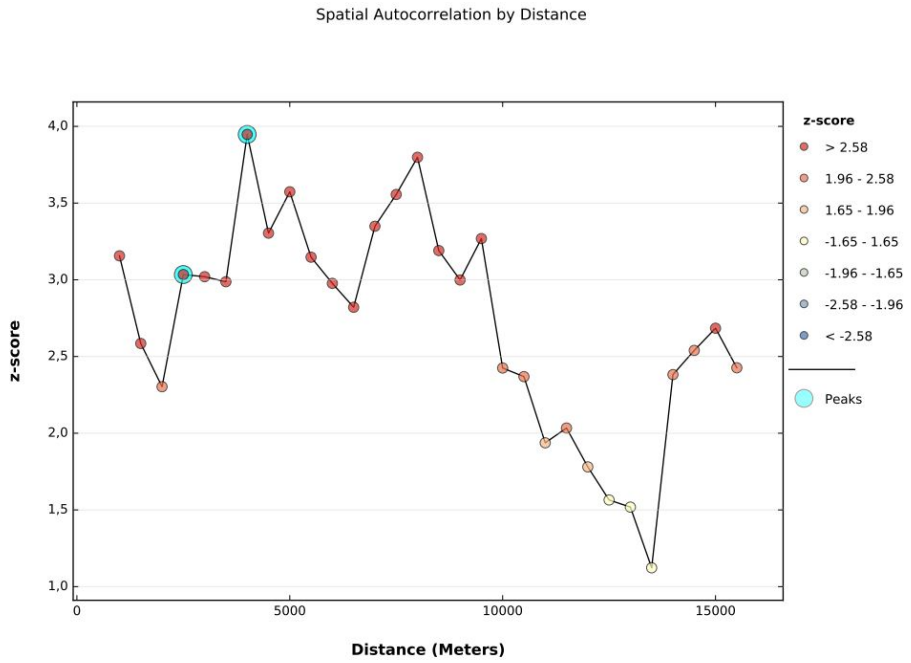
MUNI SCI
HR EXCELLENCE IN RESEARCH

- Tip:
  - Počítejte vždy s euklidovskými vzdálenostmi a tam kde je to možné si nechte vygenerovat report, který si potom můžete otevřít například v internetovém prohlížeči. Report je buď pdf nebo html soubor, který přehledně shrnuje výsledky daného nástroje, většinou i s grafem a interpretací výsledků. Je to tedy velice účinný nástroj k tomu, jak výsledkům porozumět.
- Krok 1: Spatial Autocorrelation (Morans I)
  - U toho nástroje věnujte pozornost parametru **Conceptualization of Spatial Relationships,** který mění váhy na různé vzdálenosti
  - V defaultním nastavení se sousedům postupně snižují váhy s rostoucí vzdáleností. Můžete si však například vyzkoušet možnost **Fixed Distance Band,** kdy si nastavíte fixní vzdálenost, za kterou už body nebudou mít na výpočet žádný vliv

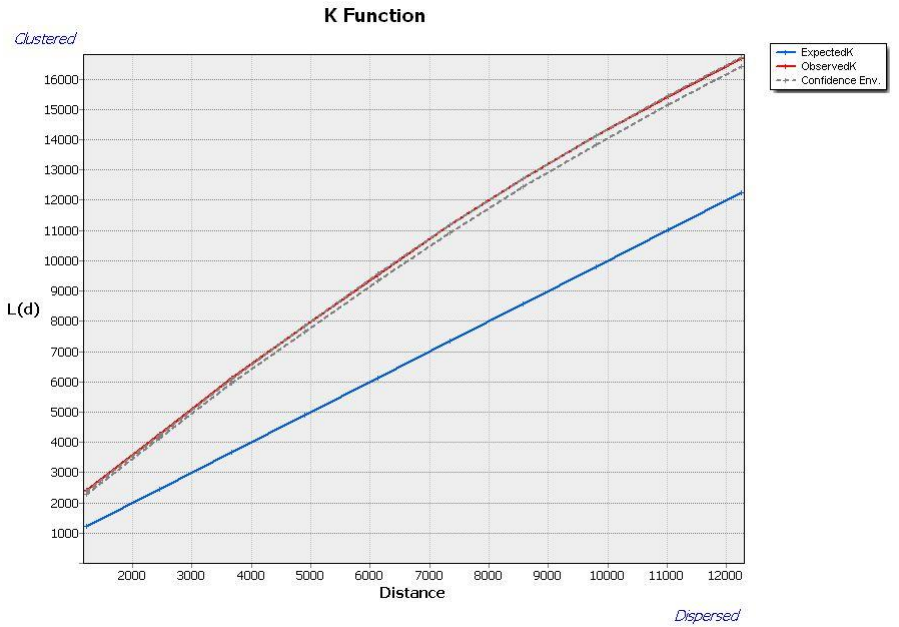
MUNI SCI
HR EXCELLENCE IN RESEARCH

Krok 2: Incremental Spatial Autocorrelation

- Tento nástroj počítá globální moranův index pro rostoucí vzdálenosti (Biginning Distance je úvodní vzdálenost - průměrná vzdálenost alespoň jednoho bodu v rámci sousedství, Distance Increment je pak nárust vzdálenosti po kterou je hledána prostorová autokorelace, ideálně polovina úvodní vzdálenosti). Tyto vzdálenosti si buď vypočte program sám nebo je můžete zadat ručně (použít lze nástroj Calculate Distance Band from Neighbor Count - zde se bude výsledek odvíjet od parametru sousedů). Ve výsledném grafu pak uvidíte peaky ve vzdálenostech na které je intenzita clustrování nejvyšší. Tuto vzdálenost lze zadat do nástroje Spatial Autocorelation (Moran I).



- Krok 3: Multi-Distance Spatial Cluster Analysis (Ripleys K Function)
  - Tento index zjišťuje jestli dané body a jejich hodnoty vykazují statisticky významné shlukování nebo rovnoměrné rozložení na různé vzdálenosti (viz studijní materiály)
  - Interpretace je taková, že pokud by se červená přímka přimykala k modré, znamenalo by to, že na danou vzdálenost soubor vykazuje náhodné rozdělení. Tam kde je červená přímka nad modrou je tendence ke shlukování a naopak
  - Tato metoda je však náchylná k různým vlastnostem datasetu jako je například k velikosti studované oblasti nebo pokud se body nacházejí v blízkosti hranic studované plochy. Ke zmírnění těchto problémů slouží parametry Boundary Correction Method a Study Area Method (více se k tomu můžete dočíst na [Multi-distance Spatial Cluster Analysis]())



Zhodnocení výsledku

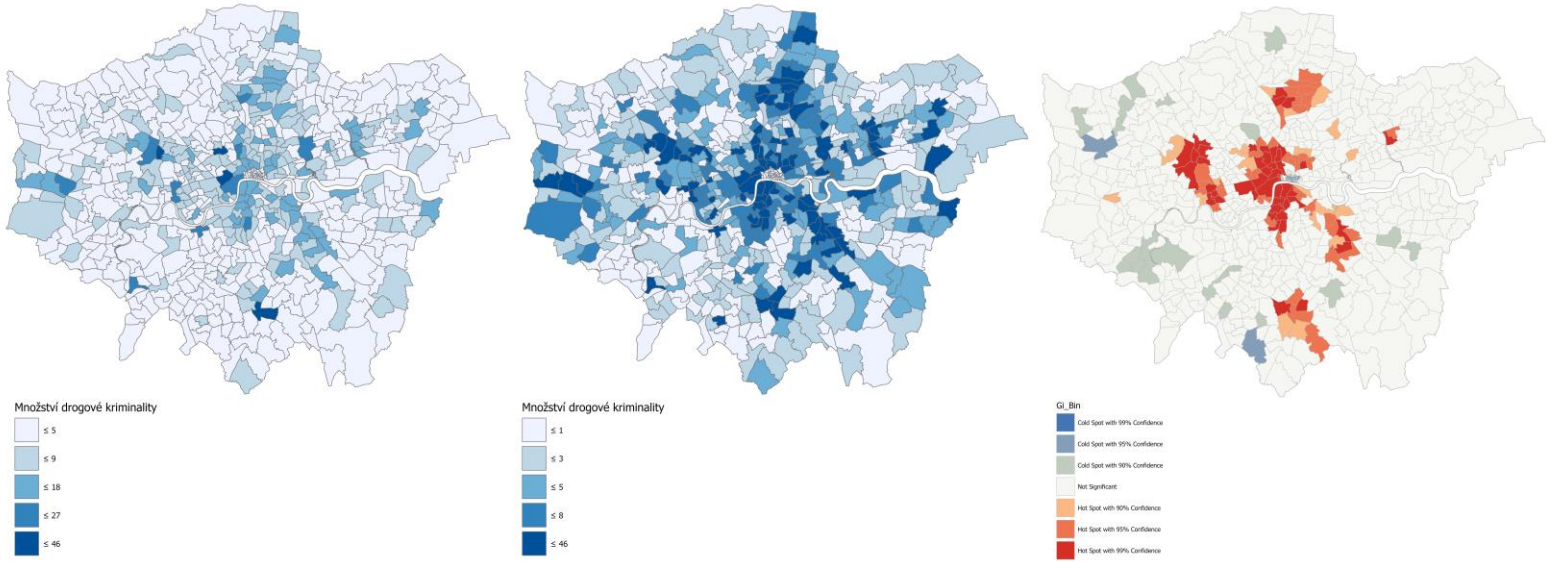
- I = Moranův index I
- E(I) = očekávaná hodnota indexu

Moranův index I	Prostorové uspořádání	H0
I > E(I)	pozitivní prostorová autokorelace, sousední body vykazují podobné hodnoty, shlukové uspořádání	Z > 1,96 statisticky významně pozitivní
I = E(I)	nulová prostorová autokorelace, body nevykazují znaky podobnosti, náhodné uspořádání	
I < E(I)	negativní prostorová autokorelace, sousední body vykazují rozdílné charakteristiky, pravidelné uspořádání	Z < -1,96 statisticky významně negativní

MUNI SCI
HR EXCELLENCE IN RESEARCH

Lokální Indexy

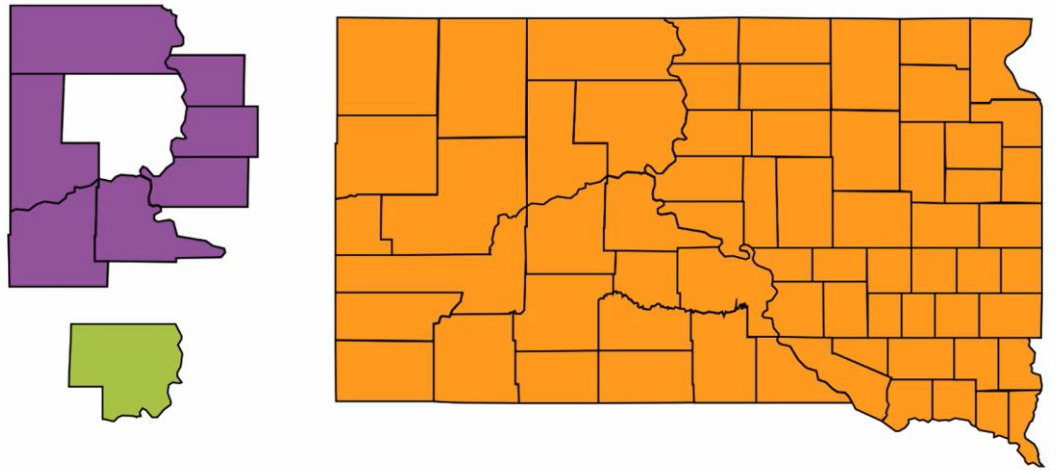
- Pro tyto indexy si vyzkoušíme jinou agregaci dat než v předešlém případě. Tentokrát využijeme nástroj **Spatial Join** na "spojení" záznamů z bodové vrstvy (původní vrstva drogové kriminality - NE výsledek collect events) do polygonové vrstvy londýnských čtvrtí. Do nástrojů potom bude vstupovat atribut ***Join_Count***



Drogová kriminalita v Londýně - Tyto tři mapy znázorňují drogovou kriminalitu v Londýně. První dvě používají rozdílné klasifikační metody pro zobrazení prostorového vzoru v datech. Třetí mapa využívá statistickou shlukovou analýzu. Mapa je založená na statistice, která slouží k určení, zda vzor, který vidíme, dává smysl. Nástroji Cluster and Outlier a Hot Spot se snažíme minimalizovat subjektivitu.

MUNI SCI
HR EXCELLENCE IN RESEARCH

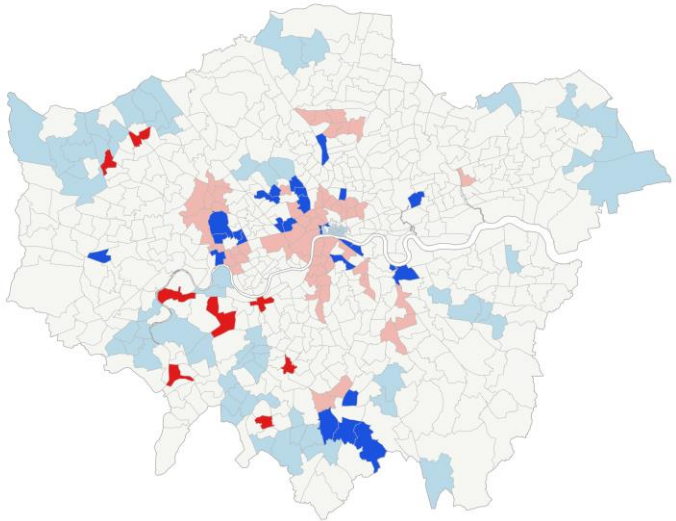
- Krok 1: Cluster and Outlier Analysis (Anselin Local Moran's I)
  - Vypočítá hodnotu Moranova indexu pro jednotlivé polygony. Vysoký index značí, že daný polygon má podobně vysoké či nízké hodnoty atribut a je tedy součástí clusteru. Zde se porovnává hodnota polygonu s průměrnou hodnotou ze všech ostatních polygonů v celé studované oblasti a hodnota sousedství s průměrnou hodnotou všech sousedství v celé studované oblasti - polygon (feature) a sousedství (neighborhood) se vyhodnocují zvlášť. Záporný index značí polygon lišící se od okolních a je outlierem. Aby byl tento vztah statisticky významný, musí mít příslušně vysokou p hodnotu



- Tip:
  - při počítání s polygony
  - Conceptualization of Spatial relationships: Contiguity edges corners
  - Standardization: Row (minimalizuje vliv, že každý polygon má jinou rozlohu a jiný počet sousedů)
  - FDR Correction (pokud zaškrtnu), používá konzervativnější přístup k identifikaci statisticky významných prvků. Konzervativní přístup znamená, že FDR korekce přistupuje k identifikaci statisticky významných prvků opatrně a spíše zahrnuje jen ty, u kterých je vysoká pravděpodobnost, že jsou skutečně významné. Tím se minimalizuje riziko falešně pozitivních výsledků a zvyšuje se spolehlivost výsledků statistické analýzy dat. **V našem případe Nezaškrtávat!!!**

MUNI SCI
HR EXCELLENCE IN RESEARCH

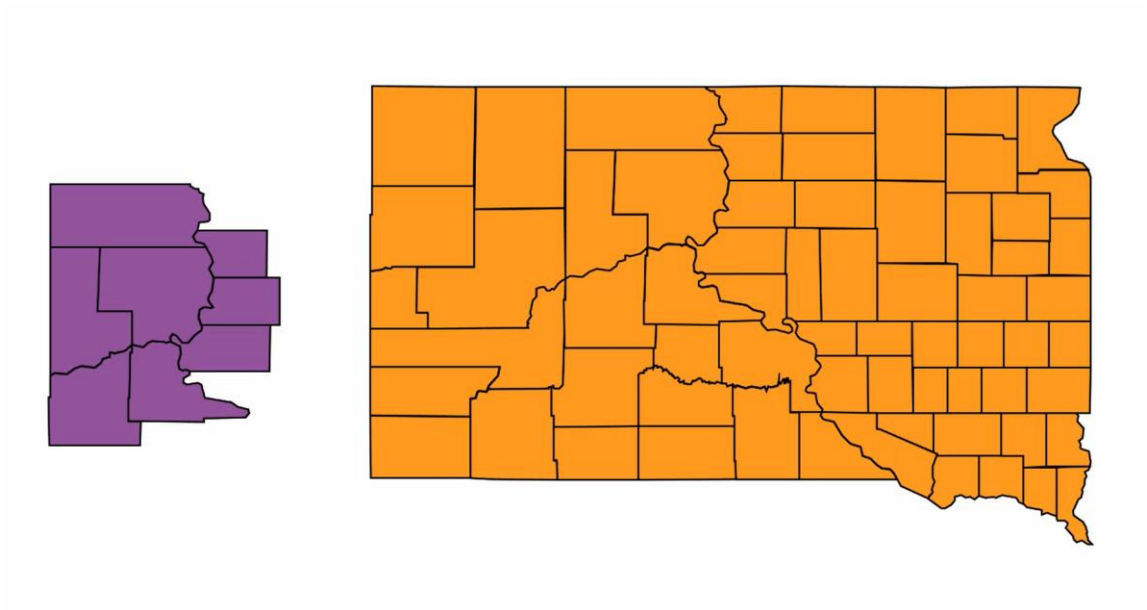
- HH: Polygon (feature) má vyšší hodnotu, než všechny ostatní polygony a spadá do sousedství, které má také vyšší hodnotu, než ostatní sousedství v celé oblasti. => High-High Cluster
- LL: Polygon má nižší hodnotu, než všechny ostatní polygony a spadá do sousedství, které má také nižší hodnotu, než ostatní sousedství v celé oblasti. => Low-Low Cluster
- HL: Polygon má vyšší hodnotu, než všechny ostatní polygony, ale spadá do sousedství, které má nižší hodnotu, než ostatní sousedství v celé oblasti. => High-Low Outlier
- LH: Polygon má nižší hodnotu, než všechny ostatní polygony, ale spadá do sousedství, které má vyšší hodnotu, než ostatní sousedství v celé oblasti. => Low-High Outlier



MUNI SCI
HR EXCELLENCE IN RESEARCH

Krok 2 : Hot Spot Analysis (Getis-Ord Gi*)

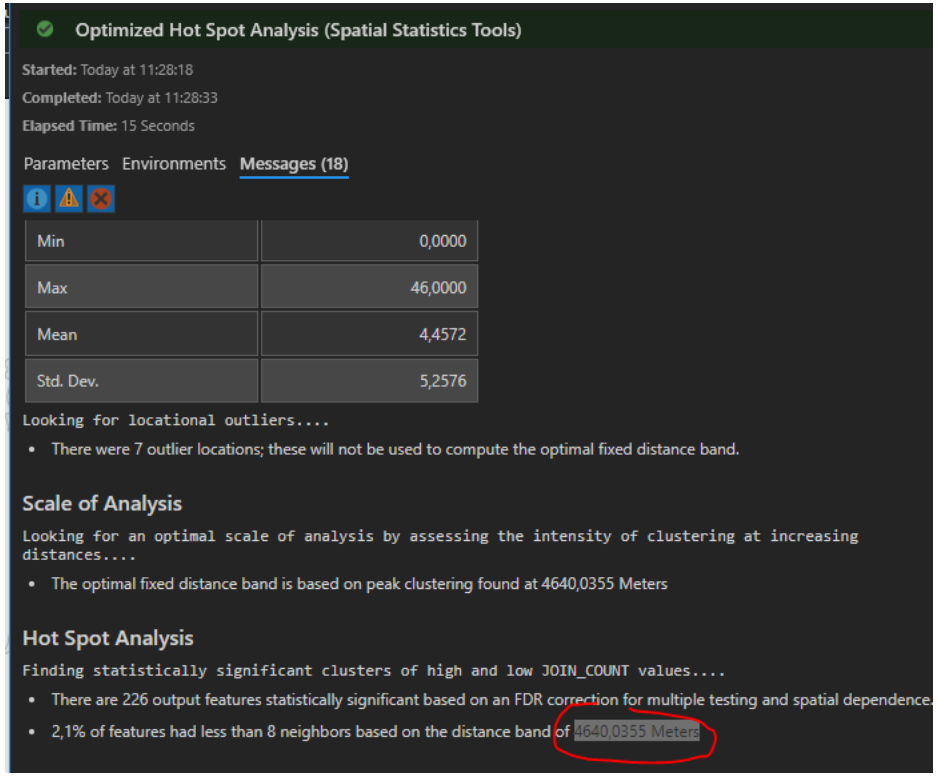
- Nástroj počítá signifikantní hot a cold spoty. Aby byl polygon přiřazen do hot-spotu musí mít vysokou hodnotu atributu a zároveň jeho sousedé také musí mít vysokou hodnotu daného atributu - polygon (feature) a sousedství (neighborhood) se vyhodnocují dohromady. Lokální suma polygonu a jeho sousedů je porovnávána se sumou celého souboru, pokud je tato suma výrazně vyšší, jedná se o hot-spot.

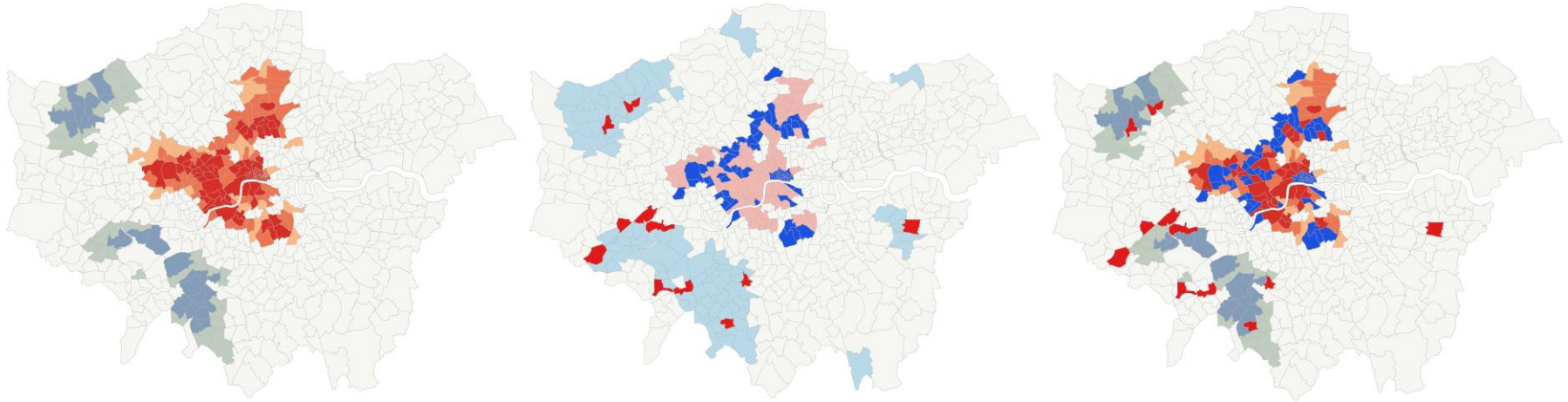


- Oba tyto nástroje mají také svou optimized verzi, kterou si také můžete vyzkoušet.

  Optimized metody za vás určitým výpočtem zoptimalizují nastavení

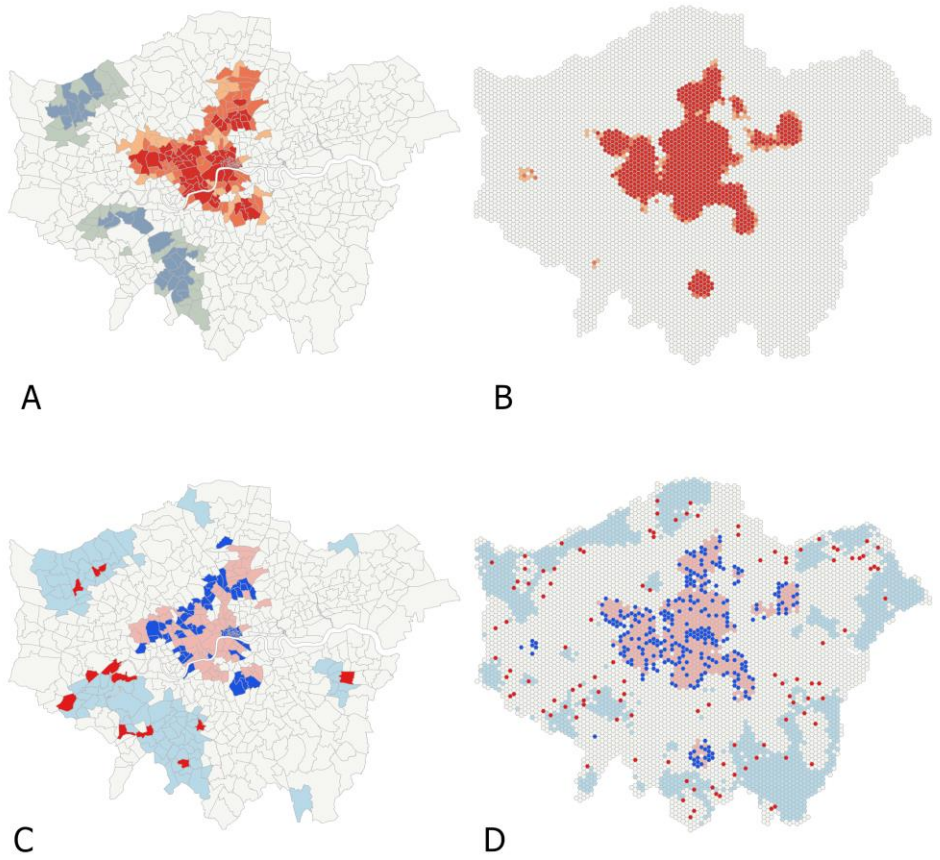
- Krok 3: Optimized Hot Spot Analysis / Optimized Outlier Analysis
  - Do nastavení nástroje vložíte pouze vstupní vrstvu a vyberete atribut. Jediný parametr, který lze u těchto optimalizovaných verzí nastavovat je **Distance Band** v Override Settings. Pokud není zadaná hodnota, je vypočtena automaticky a lze ji poté najít ve View Details po dokončení výpočtu (na obr). Je to vzdálenost, po kterou probíhá prostorová autokorelace, tedy stejný parametr, který se zjišťoval hned na začátku nástrojem **Inceremental Spatial Autoccorelation**.





A B C

Výsledky optimalizovaných verzí a jejich kombinaceA) Optimalized Hot Spot, B) Optimalized Outlier a C) Optimalized Hot Spot s Outlier



Poloha atributy porovnaniA) Optimized Hot Spot pro atributy B) Optimized Hot Spot pro polohu C) Optimized Outlier pro atributy D) Optimized Outlier pro polohu

Odkazy

- What is z-core and p-value?
- Modeling spatial relationships
- Video tutorial: Mapping clusters
- Video tutorial: Hot Spot and Cluster and Outlier Analysis
- Video tutorial: Optimized Hot Spot and Optimized Outlier Analysis

MUNI SCI
HR EXCELLENCE IN RESEARCH

MUNI SCI

HR EXCELLENCE IN RESEARCH

Děkuji za pozornost!