MUNI SCI

HR EXCELLENCE IN RESEARCH

Regresní Analýza

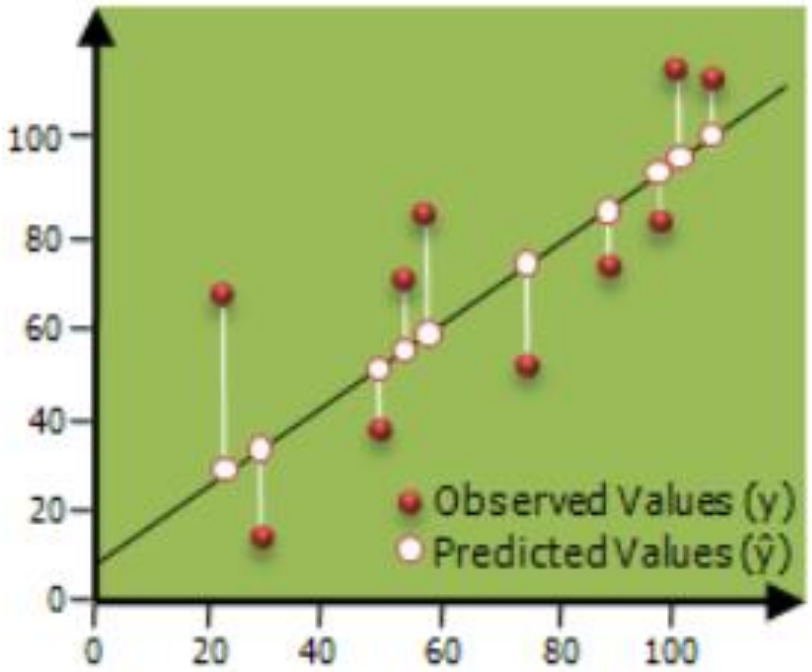
24.4.2024

Regresní analýza

- Cílem RA je zkoumat, jak jedna nebo více nezávislých proměnných ovlivňují závislou proměnnou (např. míru srážek, koncentraci znečišťujících látek, teplotu atd.)
- V geostatistice se používá k predikci hodnot na základě geografické polohy
- Umožňuje lépe porozumět a predikovat vzorce a procesy v prostorově strukturovaných datech, což je užitečné v mnoha aplikacích (environmentální studie, urbanistické plánování, geologie, a další…)

Ordinary Least Squares

- Metoda nejmenších čtverců
- Cílem OLS je najít lineární vztah mezi závislou proměnnou (např. teplotou) a jednou nebo více nezávislými proměnnými (např. nadmořskou výškou, vzdáleností od vodního zdroje atd.)



Jak funguje OLS?

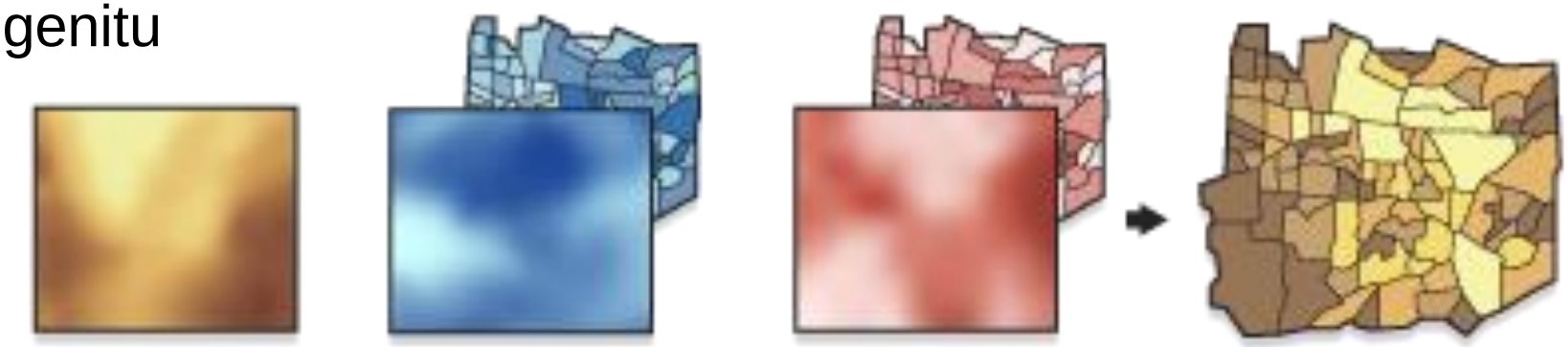
- Definice modelu $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots \beta_n X_n + \varepsilon$
  - *Y* je závislá proměnná
  - *X*1,*X*2,...,*Xn* jsou nezávislé proměnné
  - *β*0,*β*1,...,*βn* jsou parametry modelu, které chceme odhadnout
  - *ϵ* je náhodná chyba, předpokládaná jako normálně rozdělená s průměrem nula
- Definice modelu
  - OLS se snaží najít hodnoty β tak, aby minimalizovaly součet čtverců reziduí
  - Matematicky se toho dosahuje derivací a minimalizací funkce ztrát, která je součtem čtverců těchto rozdílů
- Vyhodnocení modelu
  - Po odhadnutí koeficientů se hodnotí kvalita modelu pomocí různých statistik, jako je *R*2 (koeficient determinace), který udává, jak dobře model vysvětluje variabilitu pozorovaných dat
  - Důležitá je také kontrola reziduí, zda jsou distribuována náhodně a nevykazují žádné vzory, které by naznačovaly problémy v modelu (např. neadekvátnost lineárního přístupu).

Specifika OLS

- Při použití OLS by jsme měli být opatrní, protože standardní předpoklady o nezávislosti reziduí často neplatí kvůli prostorové autokorelaci, to může vést k zkresleným odhadům standardních chyb koeficientů
- Mohou být potřeba speciální techniky, jako je prostorově vážená regrese nebo použití autoregresních modelů, aby se tyto problémy vyřešily a zlepšila přesnost a spolehlivost modelu

Geographically Weighted Regression

- Geograficky vážená regrese
- umožňuje lokalizovat regresní koeficienty pro každý bod v prostoru
- To znamená, že na rozdíl od tradiční regresní analýzy, kde se odhadují globální parametry modelu platné pro celou studovanou oblast, GWR přizpůsobuje regresní model pro každou lokaci zvlášť, což zohledňuje prostorovou heterogenitu



Jak GWR funguje?

- Lokalizované odhady parametrů
  - Každému místu v prostoru se přiřazuje vlastní sada regresních koeficientů. Namísto jednoho globálního modelu, který platí pro celou studovanou oblast, GWR vytváří mnoho "lokálních" modelů.
- Váhové funkce
  - GWR využívá váhové funkce k přiřazení většího významu bodům blízkým místu, pro které se vypočítává regrese, a menšího významu bodům, které jsou dále. Tento proces se nazývá prostorové vážení.
- Šířka pásma (bandwidth)
  - Šířka pásma je parametr, který určuje, jak daleko od středu lokace (bodu, pro který se odhadují koeficienty) mají data ještě významný vliv. Šířka pásma může být buď pevná pro celou studii, nebo se může pro každý bod adaptivně měnit.
- Odhad koeficientů:
  - Pro každou lokaci se použije regresní model s daty váženými podle vzdálenosti, a tím se odhadnou lokální regresní koeficienty.
- Interpretace, vizualizace a vyhodnocení modelu

Rozdíly mezi OLS a GWR

OLS

- **Globální** (jedna sada regresních koeficientů)
- Předpokládá **homogenitu** ve vztazích mezi proměnnými (efekt nezávislých proměnných na závislou proměnnou je konzistentní všude)
- **Nepoužívá váhovou funkci ani šířku pásma** (všechny datové body jsou zahrnuty do analýzy s rovnakým vážením)

GWR

- **Lokální** (různé regresní koeficienty různým místům v prostoru)
- Předpokládá **heterogenitu** (efekt nezávislých proměnných se může měnit v závislosti na geografické poloze)
- **Používá šířku pásma** (určí se tím, jak daleko z okolí bodu mají být data zahrnuta do lokálního odhadu, **a váhovou funkci**, která přiřazuje různé váhy bodům na základě jejich vzdálenosti od místa odhadu.

- **Koeficienty** získané OLS jsou interpretovány jako **průměrný efekt pro celou studovanou populaci**
- relativně **jednoduchá** a **málo výpočetně náročná metoda**
- **Ověření modelu** OLS je poměrně **přímočaré** s použitím standardních metrik jako **R^2, F-statistika a p-hodnoty**

- **Koeficienty** získané GWR jsou **specifické pro lokaci a musí být interpretovány v kontextu každého místa**, pro které byly odhadnuty
- **komplexnější** a **výpočetně náročnější** kvůli potřebě odhadnout mnoho sestav koeficientů pro různá místa
- vyžaduje **sofistikovanější přístupy k diagnostice a ověření**, protože je třeba hodnotit, jak dobře model funguje v různých lokalitách a zda nedochází k nadměrnému přizpůsobení (overfitting)

Cvičení

- Stáhnout data
  - **Data cv 09**
    - 911_calls.shp – tísňová volání
    - response_stations.shp – výjezdová místa
    - ObsData911Calls.shp – data ze sčítání, demografická a socioekonomická data

Zadání

- Na základě podkladových dat pro cvičení (bodová vrstva s telefonáty na tísňovou linku 911) rozhodněte, jestli jsou výjezdová centra záchranných složek v Portlandu (Oregon) správně rozmístěna. Pomocí prostorové závislosti zjistěte, proč jsou v hot-spotech tak četné hovory na tísňovou linku.

- *Používejte souřadný systém WGS 1984 UTM Zone 10N (s transformací WGS 1984 (ITRF00) To NAD 1983).*

Postup vypracování

OLS

- 1) První začneme využitím podobných nástrojů jako minulý týden, takže si je můžete připomenout při mrknutí do minulé kapitoly.
- Jako první použijte nástroj **Integrate** na bodovou vrstvu volání,t ento nástroj nám sjednotí body, které se nacházejí v blízkosti (ArcgGis Pro vyhodnotí X,Y toleranci během intergrace, na základě prostorové reference vstupních prvků)
- 2) Poté and bodovou vrstvou použijte nástroj **Collect Events** známý z minulé hodiny.

- 3) Vypočítejte Spatial Autocorrelation (Global Moran I)

  Calculate Distance Band from Neighbour Count -> Incremental Spatial Autocorelation -> Spatial Autocorrelation (Global Moran I)

Dobrovolný krok

- 4) Pomocí nástroje **Hot Spot Analysis (Getis-Ord Gi*)** vypočtěte centra zvýšeného výskytu volání na tísňovou linku. Vstupní vrstvou budou agregovaná volání (výstup příkazu Collect Events), vstupní atribut bude ICount. Distance Band or Threshold Distance bude nastavený na lokální maximum zjištěné v bodu 3). Výslednou hodnotu GiZScore interpolujte pomocí libovolné metody.

- 5) Použijte nástroje lokálních indexů pro zjištění, kde se vyskytují volání **- Hot Spot Analysis (Getis-Ord Gi*) a Cluster and Outlier Analysis**.
- 6) Spusťte **Ordinary Least Squares** nad vrstvou ObsData911Calls. Unique ID Field bude UniqID, Dependent Variable bude Calls a Explanatory Variables bude atribut Pop. Z výsledků je patrné, že počet obyvatel vysvětluje pouze z 39 % počet volání na tísňovou linku (hodnota Adjusted R-Squared). Tzn. víme, že vysoký počet obyvatel v městské části neznamená vysoký počet hovorů.

- 7) Pomocí nástroje Spatial Autocorrelation ověřte, jestli jsou hodnoty reziduí náhodně rozmístěny. V případě, že by nebyly, byla by chyba v tom, že jsme nenašli nějaký významný prostorový faktor (např. další závislý atribut).

  Nastavení bude následující:

  - Input Field: StdResid
  - Generate Report: ON
  - Conceptualization of Spatial Relationships: Inverse Distance/Contiguity edges corners
  - Distance Method: Euclidean Distance
  - Standardization: Row
  - Výsledek reziduí a Spatial Autocorre

- 8) Zjistěte závislost vybraných atributů na počtu volání anebo populaci (pomocí nástroje Create Scatterplot Matrix Graph v nabídce View/Graphs); použijte např. atributy Pop, Jobs, Renters, Bussiness, ForgnBorn, NotInLF, LowEduc, Dst2UrbCen, atd.
- 9) Znovu spusťte Ordinary Least Squares nad vrstvou ObsData911Calls. Unique ID Field bude UniqID, Dependent Variable bude Calls a Explanatory Variables budou atributy Pop, Jobs, LowEduc a Dst2UrbCen. Z výsledků je patrné, že zvolené atributy vysvětlují přes 83 % počet volání na tísňovou linku (hodnota Adjusted R-Squared).

- 10) Kontrola výstupů OLS
- 11) Spatial Autocorrelation - Bod 7
- 12) Pro automatický výběr a nalezení vysvětlujících proměnných slouží nástroj Exploratory Regression. V Search Criteria lze nastavit minimální a maximální počet proměnných, které vstoupí do modelu. Dívám se na ty modely, které mají data u Passing Models.

GWR

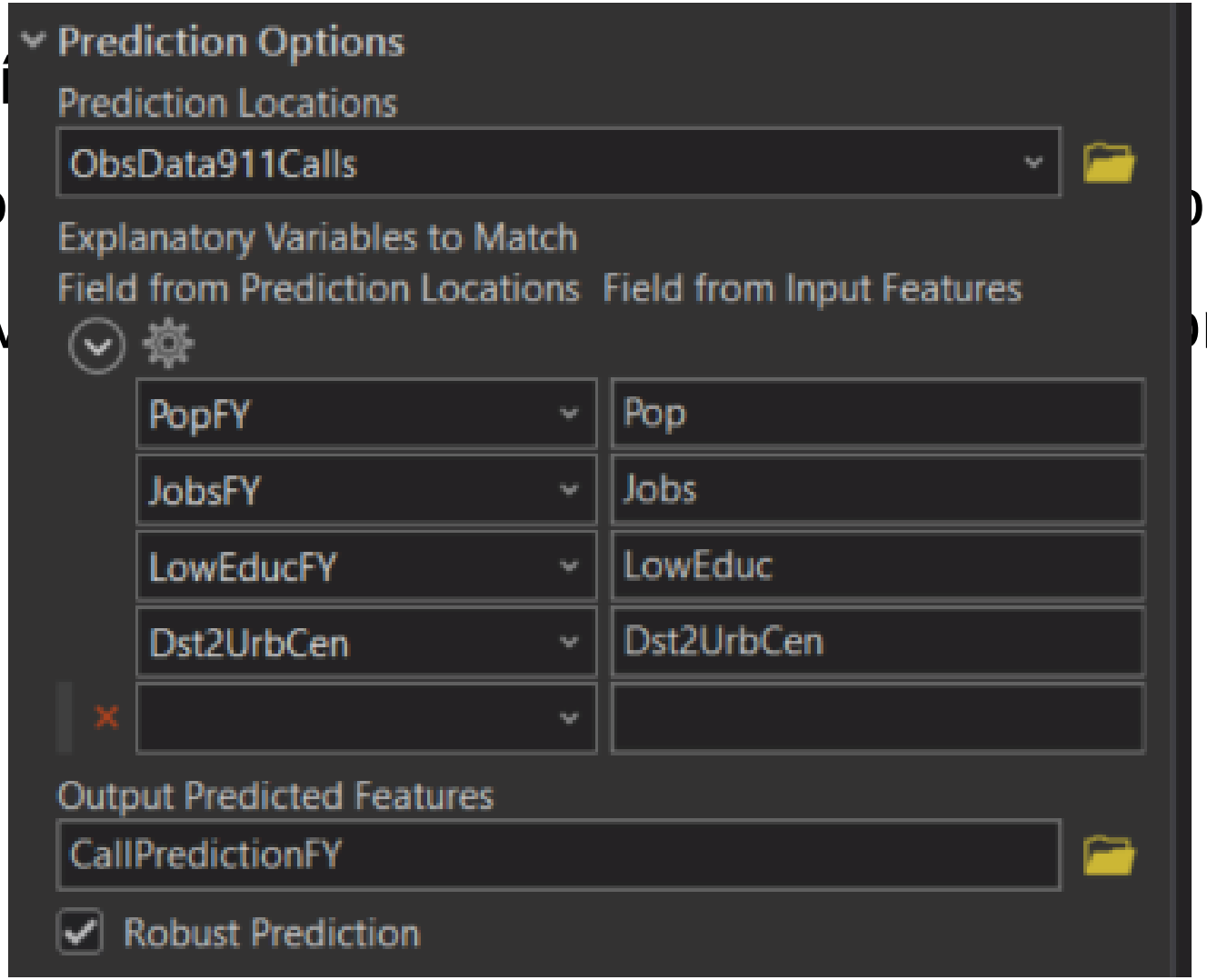
- GWR: lokální regresní model, který vypočítá rovnici pro každý polygon (feature), namísto pro všechny polygony dohromady. Velikost sousedství je však pro všechny vysvětlující proměnné stejné. GWR dovoluje vztahy mezi proměnnými, které se mění v prostoru.

- Kdy použít GWR?

- - Když je výsledek Koenker test statisticky významný (*) při použití OLS - data nejsou stacionární.

- 1) Spusťte nástroj Geographically Weighted Regression
  - Neighborhood type: Number of neighbors
  - Neighborhood Selection Method: Golden search
- 2) Porovnejte hodnoty AIC a R2Adjusted mezi metodami OLS (z minulého cvičení) a GWR.
- 3) Ověřte prostorové rozložení atributu StdResid. (nástroj Spatial Autocorrelation)
- 4) Jednoduše si vizualizujte koeficienty pro jednotlivé proměnné (Pop, Jobs, LowEduc a Dst2UrbCen) formou jednoduché mapy.

- GWR nám dovoluje i predikovat vysvětlovanou proměnnou (Calls), pokud máme pro naše vysvětlující proměnné hodnoty, které značí
- 5) Znovu o odu 1. Tentokrát v ons



Multiscale GWR

- MGWR je rozšíření GWR, které umožňuje, aby se okolí každého prostorového prvku lišilo mezi jednotlivými vysvětlujícími proměnnými. To znamená, že pro některé vysvětlující proměnné může být okolí větší nebo menší než pro jiné proměnné. Povolení různých sousedství pro různé vysvětlující proměnné je důležité, protože vztahy mezi vysvětlujícími proměnnými a závislou proměnnou mohou fungovat na různých prostorových škálách: koeficienty některých proměnných se mohou v rámci zkoumané oblasti měnit postupně, zatímco koeficienty jiných proměnných se mění rychle. Nastavení nástroje je zcela stejné jako u klasického GWR.
- 6) Ověřte prostorové rozložení atributu StdResid. (nástroj Spatial Autocorrelation)

- Příští týden 2. zápočtový test?

MUNI SCI

HR EXCELLENCE IN RESEARCH

Děkuji za pozornost!