GEOSTATISTIKA - cv. 8: Hot-spot analýza a prostorová regrese

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.

- 911_calls.shp – tísňová volání
- response_stations.shp – výjezdová místa
- ObsData911Calls.shp – data ze sčítání, demografická a socioekonomická data

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**). K vypracování využijte program ArcMap.

Poznámky:

1) Nejprve je nutné volání z bodové vrstvy „agregovat“ do prostorových útvarů, které budou reprezentovat intenzitu jevu. Pokud jsou vstupní vrstvou polygony (správní/administrativní jednotky), je možné použít nástroj **Spatial Join**. Pokud používáme vlastní pravidelnou síť, lze využít nástroje **Create Fishnet** v kombinaci s nástrojem **Spatial Join**. Pro bodovou vrstvu je možnost agregování pomocí nástroje **Integrate** v kombinaci s nástrojem **Collect Events**. Příkaz **Integrate** nevratně změní vstupní data, takže si je před použitím příkazu **zkopírujte nebo zálohujte**! Poslední nástroj ještě neznáte, takže použijeme příkaz **Integrate** (**XY tolerance** bude **30 stop**) a poté spusťte **Collect Events**. Výsledkem bude agregovaná vrstva s počty bodů.
2) Vypočtěte míru prostorové autokorelace podle vzdálenosti pomocí nástroje **Incremental Spatial Autocorrelation** (vstupní atribut bude **ICount** z vrstvy generované příkazem **Collect Events**). Nastavení ponechte defaultní, zprávu uložte do pdf a prohlédněte si ji.
3) 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 první lokální maximum zjištěné v bodu 2). Peak by měl mít hodnotu okolo **4650** (zkoumáme lokální vliv, proto používáme první peak; kdybychom řešili regionální závislost, použila by se hodnota druhého / třetího peaku, viz ArcGIS help). Výslednou hodnotu **GiZScore** interpolujte pomocí libovolné metody.
4) 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.
5) 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.

6) 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**).
7) Pomocí nástroje **Spatial Autocorrelation** ověřte, jestli jsou hodnoty reziduí náhodně rozmístěny. V případě, že by nebyly, by byla 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
    - Distance Method: Euclidean Distance
    - Standardization: ROW (pro polygony so používá jen ROW, viz help)
8) Kontrola výstupů OLS:
    - **COEFFICIENT**: vypovídá v závislosti mezi závislou a vysvětlovanou proměnnou; v našem případě je pozitivní závislost u většiny atributů, nejvýraznější u **LowEduc**. Prakticky to znamená, že čím více **LowEduc**, tím více hovorů. Opačným případem je atribut **Dst2UrbCen** (vzdálenost od centra města). Zde je závislost negativní.
    - **VIF (variance inflation factor)**: řeší tzv. kolinearitu. Je-li hodnota vysoká (větší než 10), jsou atributy výrazně závislé. Kritická hodnota pro většinu testů v programu ArcMap uvažuje již hodnoty vyšší než **7,5**. Obecně je lepší **mít hodnotu VIF co nejnižší**.
    - Statisticky významné parametry: pro ověření se používají značky (*) u parametrů **Probability** a **Robust_Pr**. Pokud proměnná není významná, měla by být odstraněna. Dále se používá test **Koenker (BP) statistic**. Pokud je test signifikantní, lze modelu důvěřovat. Viz dále.
    - **Jarque-Bera test**: test **NESMÍ** být signifikantní. Testuje náhodné prostorové rozdělení reziduí (podobně jako krok 7). Pokud je signifikantní, tak jsme pravděpodobně zapomněli jeden nebo více závislých atributů.
    - Přesnost modelu: Vysvětlují jej hodnoty **Akaike information criterion (AIC)** a **Adjusted R-Squared**. Hodnota Akaike information criterion (AIC) by měla být co nejmenší. Pokud je pro vyšší počet atributů stejně vysoká, jako pro nižší, tak se přidáním atributů přesnost nezlepšila. Hodnota Adjusted R-Squared je v intervalu od 0 do 1. Čím blíže 1, tím je model lepší.

Koenker (BP) statistics: test mimo jiné řeší stacionaritu dat. Díky tomu, že je test signifikantní, víme, že jsou data nestacionární. Existuje zde tedy nějaká závislost, vzhledem k povaze dat půjde o závislost prostorovou. To znamená, že některé atributy mohou mít v určitých oblastech velký význam při vysvětlování proměnné, v jiných oblastech bude význam menší. Jelikož **Ordinary Least Squares** uvažuje globální prostorové rozdělení hodnot, vyzkoušíme metodu **Geographically Weighted Regression**, která prostorovou variabilitu řeší i na lokální úrovni.

9) Spusťte nástroj **Geographically Weighted Regression** a použijte následující nastavení: proměnná stejné jako v kroku 6), kernel typu bude ADAPTIVE. Všimněte si rozdílů mezi hodnotami AIC a R2Adjusted (ekvivalent **Adjusted R-Squared**). Vyšší přesnost je způsobena menším počtem bodů, pro které se fituje lineární model (46).
10) Ověřte prostorové rozložení atributu **StdResid** z kroku 9)

11) Zvizualizujte koeficienty pro jednotlivé proměnné (**Pop, Jobs, LowEduc** a **Dst2UrbCen**) formou jednoduché mapy.