VYSOKÁ ŠKOLA BÁŇSKÁ – TECHNICKÁ UNIVERZITA OSTRAVA

Hornicko – geologická fakulta

Katedra geoinformatiky

PROSTOROVÁ ANALÝZA DAT

Inferenční statistika pro body

Lucie Orlíková

Ostrava, 2020

ÚVOD DO PROBLEMATIKY

Prostorová autokorelace

Prostorová autokorelace vyjadřuje míru, do jaké je výskyt určitého jevu v prostoru závislý na výskytu tohoto jevu v blízkém okolí, a je tak kvantitativním vyjádřením prostorové závislosti. Ta je ve své podstatě základním konceptem geografie a úzce souvisí s prvním zákonem geografie formulovaným Toblerem, který konstatuje, že „všechno souvisí se vším, ale bližší věci spolu souvisejí více než vzdálenější".

Prostorovou autokorelaci lze měřit několika odlišnými prostorovými auto-korelačními statistikami popisujícími podobnost blízkých pozorování v závislosti na skutečnosti, zda se jedná o diskrétní či spojitou proměnnou. Dostupné jsou také různé druhy testů prostorové autokorelace v prvotních datech a v regresních reziduích.

Všechny autokorelační statistiky tak závisí na nějaké definici prostorového vážení, která se pokouší kvantifikovat často subjektivní koncepty blízkosti, a vzájemně se liší vyjádřením atributové podobnosti c_{ij}.

Základním nástrojem pro posouzení prostorové autokorelace je funkce **Average Nearest Neighbor**.

Metoda vypočítá index nejbližšího souseda založený na průměrné vzdálenosti jednotlivých prvků od nejbližšího sousedního prvku. Výpočet je provedený z těžiště každého prvku. Poté je vypočtena průměrná vzdálenost všech nejbližších sousedů a následně dále porovnávána. V případě, že průměrná vzdálenost je menší než hypotetická hodnota náhodné distribuce prvků, vykazují analyzované prvky shlukování. V opačném případě, pokud je průměrná hodnota větší než hypotetická, jsou data považována za rozptýlená. Nulová hypotéza udává, že vrstva nebo hodnoty ve vrstvě vykazují statisticky náhodný vzorek.

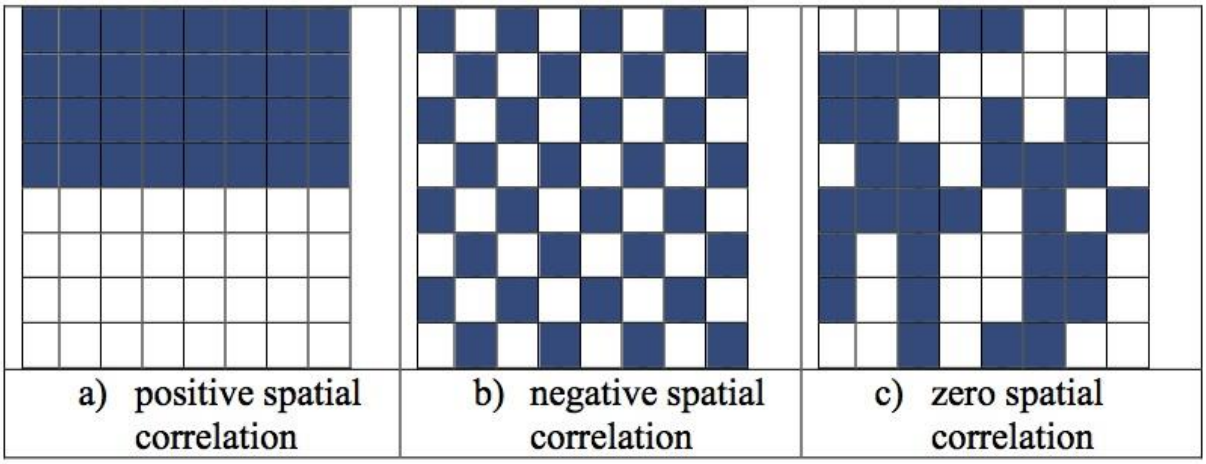
Poměr (ratio) průměrného nejbližšího souseda se vypočítá jako podíl pozorovaných průměrných vzdáleností s očekávanou průměrnou vzdáleností (ta je založena na hypotetickém náhodném rozdělení stejných hodnot prvků pokrývající stejnou plochu). V případě, že poměr je menší než 1, vzorek vykazuje shlukování, v opačném případě vykazuje disperzi.

Každý statistický nástroj provede v rámci svých výpočtů testy statistické významnosti, a kromě vlastní hodnoty, která reprezentuje výsledek toho, kterého nástroje, nám poskytne dvě velmi důležité hodnoty: tzv. p-hodnotu a Z-skóre.

P-hodnota znamená pravděpodobnost, že zadaný vzorek odpovídá nulové hypotéze, tj. v případě zkoumání prostorového uspořádání pravděpodobnost, že je jev rozložen v území náhodně.

Z-skóre je referenční hodnota pro standardní normální rozdělení pravděpodobnosti s průměrem rovným nule a směrodatnou odchylkou 1. Každá požadovaná hladina spolehlivosti má přiřazenu kritickou hodnotu Z-skóre. Když výsledek testu překročí kritickou hodnotu, říkáme, že výsledek je statisticky významný na dané hladině spolehlivosti, a můžeme odmítnout nulovou hypotézu.

Hodnocení prostorového rozložení hodnot atributů v prvcích můžeme aplikovat jak na diskrétní prvky, tak (a to zejména) při hodnocení plošných objektů souvisle pokrývajících území, kde nemá smysl hodnotit uspořádání poloh prvků (například chceme posoudit prostorové rozložení vysokého podílu seniorů v populaci obcí).



Ve statistice můžeme tento jev zkoumat pomocí prostorové autokorelace. Ta nám řekne, zda je nebo není prostorové rozložení hodnot závislé na prostorovém rozložení prvků – jinak řečeno, zda podobné hodnoty se pravděpodobně vyskytují v navzájem blízkých prvcích nebo zda mají stejnou pravděpodobnost se vyskytovat v kterémkoliv prvku v celém území. Pokud jsou si hodnoty blízkých prvků podobnější než hodnoty vzdálenějších prvků, hovoříme o pozitivní autokorelaci, pokud jsou si hodnoty blízkých prvků nepodobné, hovoříme o negativní autokorelaci, a pokud je podobnost/nepodobnost hodnot náhodná, říkáme, že zde není žádná prostorová autokorelace.

Výstupem nástroje Prostorová autokorelace (kritérium Moran I) (Spatial Autocorrelation (Moran's I)) je jedna sada hodnot pro celou vstupní datovou sadu: Z-skóre, p-hodnota a tzv. Moranův index. Pokud vyjde Z-skóre statisticky významné a kladné, je pravděpodobnost, že podobné hodnoty tvoří shluky.

V případě pozitivní autokorelace budeme chtít vyzkoumat, kde shluky jsou. K tomu máme k dispozici nástroje v toolsetu Mapování shluků (Mapping Clusters). Tyto nástroje vypočítají z-skóre, p-hodnotu a příslušný index pro každý prvek s uvážením hodnot v okolních prvcích. Je tak možné prvkům nastavit symboliku podle Z-skóre a tím v mapě zvýraznit prvky, které příslušejí do statisticky významného shluku. Můžeme hledat, kde jsou shluky vysokých/nízkých hodnot (Hot spot analýza (Getis-Ord Gi*)) nebo kde statisticky významně sousedí vysoké a nízké (extrémní) hodnoty (Analýza homogenních a heterogenních shluků (Cluster and Outlier Analysis (Anselin Local Moran's I)).

POSTUP V PROSTŘEDÍ R

```
#Načtení CSV dat ze Sčítaní lidí, domů a bytů

Census.Data <-read.csv("practicaldata.csv")


#Výpočet korelačního koeficientu. Pearsonův korelační koeficient měří sílu lineární závislosti mezi dvěma veličinami.

cor(Census.Data$Unemployed, Census.Data$Qualification)


#Vhodnější je využít korelační test, který ke koeficientu doplní i statistiky a interval spolehlivosti. Interval spolehlivosti neboli konfidenční interval je ve statistice typ intervalového odhadu neznámého parametru. Pro jeho stanovení je potřeba předem určit konfidenční hladinu (nejčastěji se používá 95 %, což je doplněk běžně používané hladiny spolehlivosti 5 % do sta procent).

cor.test(Census.Data$Unemployed, Census.Data$Qualification)


#Spearmanův korelační koeficient. Jde o neparametrickou metodu, která využívá při výpočtu pořadí hodnot sledovaných veličin, nevyžaduje tedy normalitu dat. Výhodou je, že lze tuto metodu použít pro popis jakékoliv závislosti - lineární i nelineární.

cor.test(Census.Data$Unemployed, Census.Data$Qualification, method="spearman")


#Pro vytvoření korelační matice mezi všemi záznamy v datech je nutné odebrat první sloupec.

data1 <- Census.Data[,2:5]


#Tvorba korelační matice.

cor(data1)

round(cor(data1),2)


#Korelační matice v grafické podobě.

library(ggplot2)

library(reshape2)


corr <- cor(data1)
```

```
qplot(x=Var1, y=Var2, data=melt(corr), fill=value, geom="tile") +
  scale_fill_gradient2(limits=c(-1, 1))

library(corrplot)

corrplot(corr, type="lower", tl.col="black", tl.srt=45)

#Prostorová autokorelace - nutné knihovny.
library("sp")
library("rgdal")
library("rgeos")

Output.Areas <- readOGR(".", "Camden_oa11")
OA.Census <- merge(Output.Areas, Census.Data, by.x="OA11CD", by.y="OA")
House.Points <- readOGR(".", "Camden_house_sales")

#Rozložení hodnoty v prostoru
library("tmap")

tm_shape(OA.Census) + tm_fill("Qualification", palette = "Reds", style = "quantile", title = "% with a Qualification") + tm_borders(alpha=.4)

#Prostorová autokorelace. Prostorová autokorelační statistika nám opatřuje a analyzuje míru závislosti mezi pozorováními v geografickém prostoru. Klasické statistiky prostorové autokorelace jsou Moranova, Gearyho, Getisova a standardní odchylky elipsy. Tyto statistické údaje vyžadují měření, které závisí na prostorové matici, odráží intenzitu geografického vztahu pozorování v okolí, např. vzdálenosti mezi sousedy, jako je délka společné hranice, nebo spadají-li do zadané směrové třídy, jako třeba do třídy „západ“. Klasické statistiky prostorové autokorelace porovnávají váhu prostorového kovariančního vztahu v párových místech.
library(spdep)
```

#Definice sousedství - Označují se podle pohybu šachových figur (Rook's case – věž, Queen's case – Dáma).

```
neighbours <- poly2nb(OA.Census)

neighbours


plot(OA.Census, border = 'lightgrey')

plot(neighbours, coordinates(OA.Census), add=TRUE, col='blue')

plot(neighbours2, coordinates(OA.Census), add=TRUE, col='red')
```

#Po definování sousedních polygonů, je možné spustit model. Je nutné převést datový typ u sousedství.

```
listw <- nb2listw(neighbours2)

listw
```

#Moranův test - patří mezi lokální metody.

```
moran.test(OA.Census$Qualification, listw)

moran <- moran.plot(OA.Census$Qualification, listw = nb2listw(neighbours2, style = "W"))
```

#Výběr z rozptylogramu - pozitivní vztahy.

```
local <- localmoran(x = OA.Census$Qualification, listw = nb2listw(neighbours2, style = "W"))

moran.map <- cbind(OA.Census, local)

tm_shape(moran.map) + tm_fill(col = "Ii", style = "quantile", title = "local moran statistic")
```

#metoda LISA identifikuje shluky (bodů) s podobnými hodnotami a shluky s rozdílnými hodnotami, tj. nezabývá se velikostí hodnot. Určuje, zda se vyskytuje shluk prvků, nebo prostorový outlier.

```
quadrant <- vector(mode="numeric",length=nrow(local))

m.qualification <- OA.Census$Qualification - mean(OA.Census$Qualification)

m.local <- local[,1] - mean(local[,1])

signif <- 0.1

quadrant[m.qualification >0 & m.local>0] <- 4
```

```
quadrant[m.qualification <0 & m.local<0] <- 1
quadrant[m.qualification <0 & m.local>0] <- 2
quadrant[m.qualification >0 & m.local<0] <- 3
quadrant[local[,5]>signif] <- 0

brks <- c(0,1,2,3,4)
colors <- c("white","blue",rgb(0,0,1,alpha=0.4),rgb(1,0,0,alpha=0.4),"red")
plot(OA.Census,border="lightgray",col=colors[findInterval(quadrant,brks,all.inside=FALSE)])
box()
legend("bottomleft",legend=c("insignificant","low-low","low-high","high-low","high-high"),
      fill=colors,bty="n")
```