MORAN a GEARY indexy pro hodnocení prostorové autokorelace plošných jevů

Metoda Joint count statistics (JCS) má značná omezení z hlediska typu dat. Pro intervalová a poměrová data jsou stejně jak v případě jevů vztažených k bodům nejvyužívanějšími měrami prostorové autokorelace plošných jevů indexy Moranův (I) a Gearyho (C)

Oba indexy mají některé společné charakteristiky, jejich statistické vlastnosti však jsou rozdílné. Vhodnější vlastnosti vzhledem k rozdělení hodnot má index I. Oba indexy jsou založeny na porovnávání hodnot atributů sousedních ploch. Mají-li tyto sousední plochy v celé studované oblasti podobné hodnoty, potom obě statistiky budou svědčit o silné pozitivní prostorové autokorelaci a naopak. Obě statistiky využívají odlišný přístup k porovnávání hodnot sousedních ploch.

Moranův index I

Index se vypočte podle následujícího vzorce:

$$I = \frac{n\sum\sum w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{W\sum(x_i - \bar{x})^2}$$

kde x_i je hodnota proměnné v ploše i
w_{ij} jsou váhy, W matice vah

Hodnota indexu kolísá od -1 pro negativní prostorovou autokorelaci do +1 pro pozitivní prostorovou autokorelaci. Očekávaná hodnota indexu je v případě nulové prostorové autokorelace je rovna

$$E_I = -\frac{1}{(n-1)}$$

Váhy se v případě tohoto indexu počítají z matic binární či stochastické (viz výše). Je-li použita binární matice, potom W ve jmenovateli je rovno dvojnásobku počtu hranic ve zpracovávané oblasti (2J).

Pokud jsou plochy s indexem *i* a *j* sousedé bude v čitateli $w_{ij} = 1$, pokud nesousedí bude 0. Pokud sousedí, vyjádří se součin odchylek hodnot *i* a *j* od průměru. Tyto součiny se sumují pro všechny sousedy. Jestliže **obě** sousední hodnoty budou nadprůměrné (ale i podprůměrné) dostaneme velké kladné číslo. Obě tyto situace ukazují na pozitivní autokorelaci – tedy podobné hodnoty jsou vedle sebe (sousedí spolu). Naopak, pokud hodnota v jedné ploše bude nadprůměrná a ve druhé podprůměrná – potom to indikuje negativní autokorelaci. **Budou-li ve zpracovávané oblasti převažovat sousedé s obdobnými hodnotami, Moranův index I bude kladný.**

Čitatel obsahuje výraz pro kovarianci $(x_i - \bar{x})(x_j - \bar{x})$, která je také základem pro definování Pearsonova korelačního koeficientu *r*. Na rozdíl od korelačního koeficientu, kovariance v případě Moran's *I* je kovariancí dvou ploch v prostoru a ve výše uvedeném vztahu pro *I* je vypočtena pouze pro případy, kdy plochy spolu sousedí. Jmenovatel vzorce je suma čtverců odchylek vážená maticí sousedství *W*.

Interpretace Moran's I:
Vypočteme hodnoty *I* a *E(I)* a následně musíme zjistit, zda rozdíl mezi nimi je statisticky významný. Tento rozdíl je opět nutné vztáhnout k míře rozptylu (např. směrodatné chybě - SE - viz. výklad k bodům) a pomocí ní odvodit standardizovanou hodnotu z-skóre

Odhady rozptylu resp. směrodatné chyby se budou lišit podle způsobu, jakým mohou být hodnoty vyšetřovaného atributu přiřazeny k jednotlivým plochám („sampling assumption").

Za **předpokladu normality** jsou hodnoty atributu x_i nezávislé a pocházejí ze základního souboru s normálním rozdělením, nejsou nijak omezeny daným prostorovým uspořádáním ve studované oblasti. Z tohoto předpokladu se rozptyl vypočte:

$$\sigma^2(I) = \frac{n^2 S_1 - nS_2 + 3(W)^2}{(W)^2(n^2-1)}$$

Za **předpokladu náhodnosti** je množina hodnot fixní. Konstantní není poloha spojená s určitou hodnotou atributu. Jinými slovy – existuje mnoho způsobů, jak je v prostoru rozmístěna daná množina hodnot. Naše rozmístění je jen jedno z možných.

Určení hodnoty rozptylu:

$$\sigma^2(I) = \frac{n\left[(n^2-3n+3)S_1 - nS_2 + 3W^2\right] - \left[\frac{1/n\sum(x_i-\bar{x})^4}{[1/n\sum(x_i-\bar{x})^2]^2}\right]\left[S_1 - 2nS_2 + 6W^2\right]}{(n-1)(n-2)(n-3)(W^2)}$$

Získáme-li hodnotu rozptylu, potom můžeme vyčíslit standardizovanou hodnot $Z_n(I)$

$$Z_n = \frac{I - E(I)}{\sigma^2(I)}$$

Pokud je hodnota $Z_n(I)$ menší (resp. větší) než -1,96 (resp. 1,96) je hodnota indexu I statisticky významně negativní (resp. pozitivní) na hladině významnosti α=0,05.

Gearyho poměr C (Geary's Ratio, C index)

Tento index je definován obdobně:

$$C = \frac{(n-1)\sum\sum w_{ij}(x_i - x_j)^2}{2W\sum(x_i-\bar{x})^2}$$

Pro výpočet indexu se jako vah využívá jedné z výše uvedených typů matic prostorových vah, nejčastěji matice binární či stochastické. Ve srovnání se vzorcem pro výpočet Moranova indexu je zřejmé, že Gearyho index se liší především v čitateli výrazu. Moranův index porovnává hodnoty atributů sousedních ploch prostřednictvím odchylek od průměru, naproti tomu Gearyho index porovnává hodnoty atributů přímo mezi sebou. Pro hodnotu indexu není rozhodující, která z hodnot x_i a x_j je větší či menší, ale jaký je jejich absolutní rozdíl – jejich nepodobnost (ve výrazu je druhá mocnina jejich rozdílu).

Gearyho index nabývá hodnot v intervalu 0 až 2. Hodnota nula indikuje dokonalou pozitivní autokorelaci (všechny sousední hodnoty atributů jsou stejné). Naopak hodnota 2 indikuje dokonalou negativní prostorovou autokorelaci. Na rozdíl od Moranova indexu, očekávaná hodnota Gearyho indexu nezávisí na počtu posuzovaných ploch *n*, ale má vždy hodnotu 1. Hodnota 1 znamená nulovou prostorovou autokorelaci.

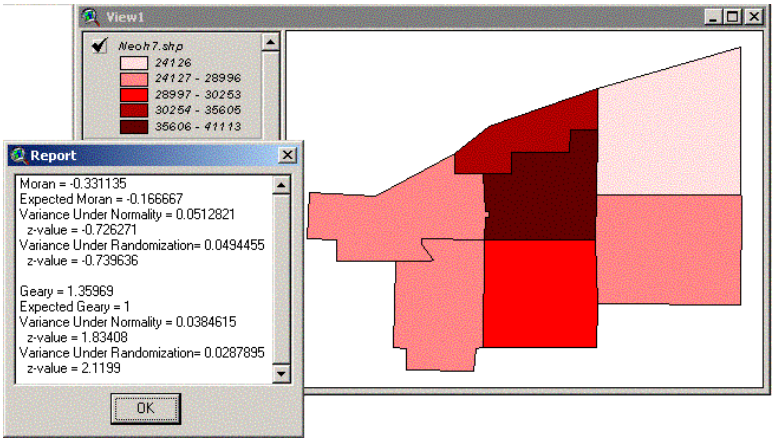
Vypočtené hodnoty indexu *C* lze porovnat s hodnotou jedna (očekávanou), pro prokázání statisticky významného rozdílu je však stejně jako v předchozích případech nutné vypočítat hodnotu *z*-skóre. Nejprve je nutné vypočítat rozptyl hodnoty indexu *C*. Hodnota rozptylu se opět vypočte rozdílně v závislosti na předpokladu normality či náhodnosti.

Například za předpokladu normality:

$$\sigma^2(c) = \frac{(2S_1 + S_2)(n-1) - 4W^2}{2(n+1)W^2}$$

Za předpokladu náhodnosti: (vzorec viz. Lee a Wong, 2000, s. 162)

Hodnoty *z*-skóre jsou založené na rozdílu pozorovaných a očekávaných hodnot. Jestliže hodnota indexu *C = 0* značí perfektní pozitivní prostorovou autokorelaci a *C = 1* nulovou, potom negativní hodnota z-skóre značí pozitivní prostorovou autokorelaci a kladná hodnota z-skóre značí autokorelaci negativní.



Obr. 5.1 Vstupní data a výsledky prostorové autokorelace (I a C indexy) pro průměrný příjem sedmi států v Ohiu.

Příklad 1: Na obrázku 5.1 je kartogram průměrného příjmu pro sedm států Ohia. Z hodnot vypočtených indexů vyplývá, že hodnota Moranova indexu indikuje **negativní** prostorovou autokorelaci (státy s vysokou hodnotou studovaného atributu jsou blízko států s nízkými hodnotami). Tato tendence však není statisticky významná na hladině 5 %.

Naopak podle vypočtených hodnot Gearyho indexu existuje statisticky významná negativní prostorová autokorelace v hodnotách průměrného příjmu u sedmi studovaných států celého regionu.

Obecná G-statistika

Oba výše uvedené indexy *I* a *C* mají dobře definované statistické vlastnosti, které popisují prostorovou autokorelaci globálně (jednou hodnotou pro celou zpracovávanou oblast). Nejsou však efektivní k identifikaci rozdílných shluků prostorového uspořádání uvnitř oblasti. Oba indexy jsou sice citlivé k identifikaci oblastí s podobnými hodnotami atributů, nerozlišují však, zda tyto podobné hodnoty nabývají vysokých či nízkých hodnot. Shluky ploch (též. místa prostorové koncentrace - spatial concentration) vysokých hodnot vyšetřovaného atributu ve studované oblasti se označují jako „**hot spots**“, naopak místa se shluky nízkých hodnot jako „**cold spots**“.

Odlišit oby typy shluků lze pomocí tzv. **obecné *G*-statistiky (general G-statistics).** Stejně jako v případě Moranova a Gearyho indexu je i *G*-statistika založena na míře prostorové asociace, která dává v čitateli výrazu do vztahu hodnoty atributu v ploše (bodě, místě) *i* a *j*. Obecná *G*-statistika je definována takto:

$$G(d) = \frac{\sum\sum w_{ij}(d) x_i x_j}{\sum\sum x_i x_j}$$

pro *i* různá od *j*. *G*-statistika je definována vzdáleností *d* mezi plochou *i* a plochami sousedními. Váha $w_{ij}(d)$ má hodnotu 1, jestliže se plocha *j* nachází ve vzdálenosti menší či rovné *d* od plochy *i*, jinak má váha hodnotu 0. Matice vah je tedy maticí binární a

symetrickou, vztahy sousedství jsou však definovány vzdáleností *d*. Suma těchto vah matice se rovná:

$$W = \sum_i \sum_j w_{ij}(d)$$

pro *i* různá od *j*. V důsledku takovéhoto definování vah, páry x_i a x_j nebudou zahrnuty v čitateli, pokud *i* a *j* jsou od sebe dále než *d*. Naproti tomu ve jmenovateli jsou zahrnuty všechny páry x_i a x_j bez ohledu na jejich vzdálenost. Z toho plyne, že jmenovatel bude vždy větší, maximálně však roven (při velkém *d*) čitateli. Čitatel výrazu pro *G(d)* statistiku, bude mít velkou hodnotu pokud sousední hodnoty budou velké a naopak. Vysoké hodnoty *G(d)* potom indikují prostorovou asociaci vysokých hodnot (hot spots) zkoumaného atributu, nízké *G(d)* potom prostorovou asociaci nízkých hodnot (cold spots).

Před výpočtem *G(d)* je nutné určit vzdálenost *d*, která definuje plochy, které budou považovány za sousedy plochy posuzované. Musí být vhodně zvolena tak, aby posuzovaná plocha měla alespoň jednoho souseda.

K interpretaci a k hodnocení statistické významnosti *G(d)* je nutné jako u výše uvedených indexů *I* a *C* vyčíslit očekávanou hodnotu *G(d)*, tedy *E(G)* a následně standardizovanou hodnotu *z*-skóre a tedy i rozptyl hodnoty *G(d)*. Očekávaná hodnota *G(d)* bude:

$$E(G) = \frac{W}{n(n-1)}$$

Očekávaná hodnota statistiky odpovídá případu, kdy neexistuje žádná prostorová asociace. Např. je-li vypočtená hodnota *G(d)* větší než očekávaná, můžeme říci, že pozorované uspořádání vykazuje pozitivní prostorovou asociaci. Statistickou významnost tohoto tvrzení je opět nutné testovat výpočtem hodnoty rozptylu *Var(G)* (vzorec viz. Lee a Wong, 2000, s. 166) a následně *z*-skóre. Opět, hodnota *z*-skóre menší než 1,96 indikuje statisticky nevýznamný výsledek na hladině α=0,05.

Příklad 2: Jsou použita stejná vstupní data jako v případě *I* a *C* indexů. Výchozí matice vzdáleností centroidů (obr. 5.2) je převedena na matici binární na základě zvolené vzdálenosti *d (d=30 mil)*- obr. 5.3

distmatrix.dbf

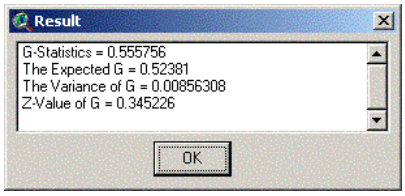
Id	Geauga	Cuyahoga	Trumbull	Summit	Portage	Ashtabula	Lake
Geauga	0.0000	25.1508	26.7057	32.7509	25.0389	26.5899	12.6265
Cuyahoga	25.1508	0.0000	47.8151	23.4834	31.6155	50.8064	28.2214
Trumbull	26.7057	47.8151	0.0000	41.8561	24.4759	29.5633	36.7535
Summit	32.7509	23.4834	41.8561	0.0000	17.8031	58.0869	42.7375
Portage	25.0389	31.6155	24.4759	17.8031	0.0000	45.5341	37.4962
Ashtabula	26.5899	50.8064	29.5633	58.0869	45.5341	0.0000	24.7490
Lake	12.6265	28.2214	36.7535	42.7375	37.4962	24.7490	0.0000

Obr. 5.2 Výchozí matice vzdáleností centroidů

distmatrix.dbf

Id	Geauga	Cuyahoga	Trumbull	Summit	Portage	Ashtabula	Lake
Geauga	0.0000	1.0000	1.0000	0.0000	1.0000	1.0000	1.0000
Cuyahoga	1.0000	0.0000	0.0000	1.0000	0.0000	0.0000	1.0000
Trumbull	1.0000	0.0000	0.0000	0.0000	1.0000	1.0000	0.0000
Summit	0.0000	1.0000	0.0000	0.0000	1.0000	0.0000	0.0000
Portage	1.0000	0.0000	1.0000	1.0000	0.0000	0.0000	0.0000
Ashtabula	1.0000	0.0000	1.0000	0.0000	0.0000	0.0000	1.0000
Lake	1.0000	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000

Obr. 5.3 Matice sousedství vypočtená pro d=30 z matice na obr. 5.2



Obr. 5.4 Výsledky výpočtu obecné G- statistiky pro vstupní data na obrázku 5.1 při použití matice vzdáleností centroidů a hodnotě definující vzdálenost d=30 mil.

Vypočtená hodnota *G(d)* vykazuje mírnou úroveň prostorové asociace, podle hodnoty *z*-skóre však výsledek není statisticky významný. Jinými slovy – dané uspořádání průměrného příjmu v sedmi státech Ohia je spíše výsledkem náhody než určitého systematického procesu.

Lokální statistiky prostorové autokorelace

Všechny tři uvedené indexy jsou příkladem indexů globálních. Jsou sumární hodnotou prostorové autokorelace pro celou zpracovávanou oblast. Je však pravděpodobné, že hodnoty prostorové autokorelace se budou v různých sub-oblastech měnit. Navíc můžeme očekávat, že pozitivní autokorelaci lze nalézt v jednom sub-regionu a negativní v jiném. Proměnlivost prostorové autokorelace v rámci studované oblasti lze vyšetřovat výše uvedenými indexy modifikovanými pro detekování prostorové autokorelace v lokálním měřítku.

LISA (Local Indicators of Spatial Association)

Jedná se o lokální verze Moranova a Gearyho indexu. Ke zjištění úrovně prostorové autokorelace na lokální úrovni je nutné vypočítat hodnotu indexu pro každou plochu zpracovávaného území. Lokální Moranův index pro jednotku *i* je definován takto:

$$I_i = z_i \sum_i w_{ij} z_j$$

kde z_i a z_j jsou odchylky od průměru nebo

$$z_i = \frac{(x_i - \bar{x})}{\sigma}$$

kde σ je směrodatná odchylka x_i. Podobně jako v případě globálního Moranova indexu znamenají vysoké hodnoty kumulaci podobných hodnot atributů (vysokých či nízkých) v sousedních plochách, nízké hodnoty potom kumulaci odlišných hodnot atributů. Obecně hodnoty w_{ij} mohou představovat po řadách standardizovanou matici vah, lze použít i jiných matic vah.

Zjištěné hodnoty lokálního Moranova indexu je nutné porovnat s očekávanými hodnotami a testovat statistickou významnost jejich rozdílu pomocí *z*-skóre.

Očekávané hodnoty při hypotéze náhodnosti:

$$E[I_i] = -w_{i.}/(n-1)$$

a hodnota rozptylu:

$$Var[I_i] = w_{i.}^2 \frac{(n - m_4/m_2^2)}{n-1} + 2w_{i(kh)} \frac{(2m_4/m_2^2 - n)}{(n-1)(n-2)} - \frac{w_{i.}^2}{(n-1)^2}$$

kde

$$w_{i.}^2 = \left(\sum_j w_{ij} \right)^2$$

$$w_{i.}^{(2)} = \sum_j w_{ij}^2 \qquad \text{pro} \qquad i \neq j$$

a výraz

$$2w_{i(kh)} = \sum_{k \neq i} \sum_{h \neq i} w_{ik} w_{ih}$$

Každá plocha ve zpracovávaném území má svoji *I* hodnotu a té přísluší hodnota očekávaná a také jistá hodnota rozptylu. Hodnoty *I* mohou být vynášeny do mapy v podobě kartogramu.

Lokální verze Gearyho poměru je definována následovně:

$$c_i = \sum_j w_{ij} (z_i - z_j)^2$$

Hodnoty rozdělení lokálního Gearyho indexu nemají tak vhodné vlastnosti jako v případě indexu Moranova. Jejich interpretace je však obdobná jako v případě globální verze indexu. Shlukování podobných hodnot atributů vede k nízkým hodnotám tohoto indexu a naopak.

Lokální G-statistika

Měří asociaci hodnot atributů v ploše *i* a v plochách okolních definovaných vzdáleností *d*:

$$G_i(d) = \frac{\sum w_{ij}(d) x_j}{\sum_j x_j} \text{ pro } \quad i \neq j$$

Obdobně jako v předchozích případech je nutné interpretovat hodnotu indexu pomocí, očekávaných hodnot, hodnot rozptylu a standardizovaných skóre. Očekávané hodnoty se vypočtou následovně:

$$E(G_i) = W_i / (n-1)$$

kde

$$W_i = \sum_j w_{ij}(d)$$

Definice rozptylu:

$$Var(G_i) = E(G_i^2) - [E(G_i)]^2$$

a

$$E(G_i^2) = \frac{1}{(\sum_j x_j)^2} \left[\frac{W_i (n-1-W_i) \sum_j x_j^2}{(n-1)(n-2)} \right] + \frac{W_i (W_i - 1)}{(n-1)(n-2)} \quad \text{pro} \quad i \neq j$$

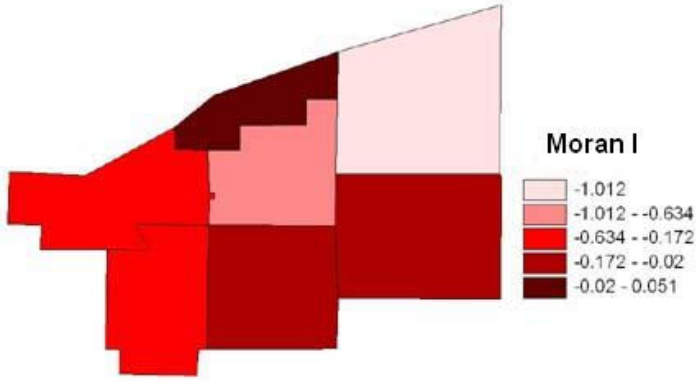
Vysoká hodnota *z*-skóre je spojena s výskytem shluků podobných a vysokých hodnot indexu. Jestliže je shluk tvořen nízkými hodnotami, *z*-skóre bude nabývat velkých záporných hodnot. Hodnoty *z*-skóre kolem nuly indikují neexistenci zřejmého prostorového uspořádání hodnot atributů v plochách studovaného území.

Příklad 3: Pro data z příkladu 1 byly vypočteny hodnoty lokálního Moranova indexu *I* (pro každý stát). Jako matice vah byla použita matice stochastická (obr. 5.5). Výsledky jsou prezentovány ve formě kartogramu na obr. 5.6 a 5.7.

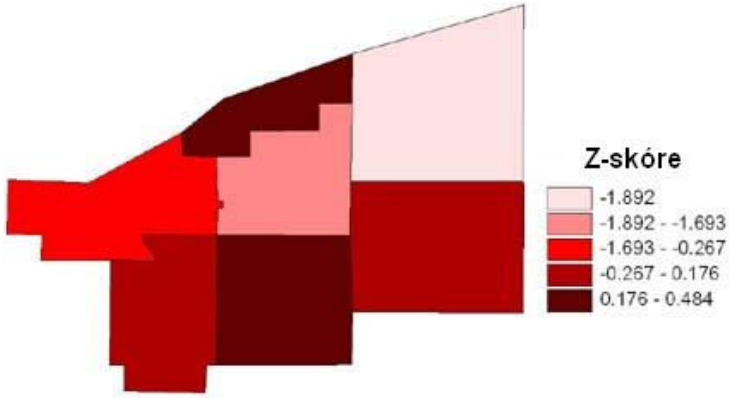
distmatrix.dbf

Id	Geauga	Cuyahoga	Trumbull	Summit	Portage	Ashtabula	Lake
Geauga	0.0000	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667
Cuyahoga	0.2500	0.0000	0.0000	0.2500	0.2500	0.0000	0.2500
Trumbull	0.3333	0.0000	0.0000	0.0000	0.3333	0.3333	0.0000
Summit	0.3333	0.3333	0.0000	0.0000	0.3333	0.0000	0.0000
Portage	0.2500	0.2500	0.2500	0.2500	0.0000	0.0000	0.0000
Ashtabula	0.3333	0.0000	0.3333	0.0000	0.0000	0.0000	0.3333
Lake	0.3333	0.3333	0.0000	0.0000	0.0000	0.3333	0.0000

Obr. 5.5 Stochastická matice vah k definování sousedství pro výpočet lokálního Moranova indexu I



Obr. 5.6 Kartogram hodnot lokálního Moranova indexu I



Obr. 5.7 Kartogram hodnot z-skóre pro lokální Moranův index I

Interpretace: Vysoké hodnoty indexu *I* mají ty státy, jejichž sousedé mají velmi podobné hodnoty studované charakteristiky. Podle *z*-skóre žádná z hodnot není statisticky významná a dané uspořádání průměrných příjmů v sedmi státech lze interpretovat jako náhodný proces.

Obdobným způsobem lze vizualizovat a hodnotit výsledky analýzy založené na lokálním indexu *C* a lokální *G*-statistice.

Moranovo korelační pole (Moran Scatterplot)

Lokální statistiky vystihují prostorovou heterogenitu v jednotlivých částech studovaného území. Pomocí nich je tedy možné jistým způsobem identifikovat oblasti s neobvyklými hodnotami měr prostorové autokorelace, které lze označit jako oblasti s odlehlými hodnotami (outliers). Efektivním nástrojem pro takovouto diagnostiku území je Moranovo korelační pole založené na regresním počtu.

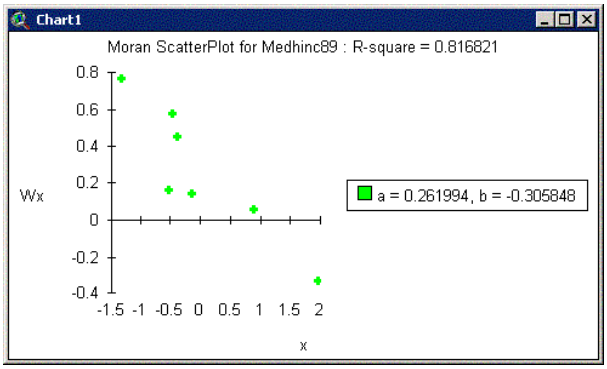
Předpokládejme, že x značí vektor hodnot x_i s odchylkami od průměru $(x_i - \bar{x})$ a dále W značí po řádcích standardizovanou matici vah. Potom můžeme sestavit regresní závislost hodnot Wx na x. Směrnice této regresní závislosti indikuje vzájemný vztah sousedních hodnot atributů. Tedy

$$x = a + IWx$$

kde a značí vektor koeficientů - (intercept). Hodnota I je regresní koeficient reprezentující směrnici a také hodnotou Moranova globálního indexu I. Vynesení regresní závislosti Wx na x umožňuje identifikovat odlehlé hodnoty. Pokud budou mít všechna pozorování podobné hodnoty prostorové autokorelace, v korelačním poli budou body blízko regresní čáry. Naopak pokud některá pozorování budou ukazovat lokálně výrazně vysoké či nízké hodnoty prostorové autokorelace ve vztahu k jejich sousedům, tato pozorování budou v grafu tvořit body výrazně nad či pod regresní čarou.

Regresní čára vyjadřuje obecný trend hodnot prostorové autokorelace v celém zpracovávaném území a parametr její směrnice je index I.

Příklad 4: Hodnota Moranova indexu (viz. Příklad 1) indikuje slabou **negativní** prostorovou autokorelaci (státy s vysokou hodnotou studovaného atributu jsou blízko států s nízkými hodnotami).



Obr. 5.8 Výsledek regresní analýzy a Moranovo korelační pole (Moran Scatterplot) pro průměrný příjem sedmi států Ohia (příklad 1). Parametr b představuje hodnotu Moranova indexu I

Z grafu je patrné že příjem (x) je nepřímo úměrný vážené hodnotě příjmu (Wx). Množinou bodů lze proložit přímku. Body, které se výrazně odchylují od přímky představují „outliers“ – představují oblasti s výrazně odlišnými hodnotami prostorové autokorelace.

Interpelace s ohledem na polohu bodů v jednotlivých kvadrantech

- high-high,low-low (2. nebo 3. kvadrant) = spatial clusters
- high-low,low-high (1. nebo 4. kvadrant) = spatial outliers