Konzultace

Geostatistika jako prostorové modelování[1] statistických jevů

Jaroslav Kraus
Český statistický úřad, Praha

Úvod

Statistika jako vědní disciplína prošla v minulosti velkým rozvojem a je zajímavé, že tento metodický rozvoj nejrůznějších stránek statistiky přetrvává až do dnešních dní. Jedna z oblastí, která se minulých desetiletích intenzivně rozvíjela, se týká prostorové analýzy, což se obecně označuje jako geostatistika. Vyjdeme-li ze základní statistické úvahy – tedy, že každý statistický jev má své věcné, časové a prostorové vymezení, jsou možnosti geostatistiky zjevné.

Geostatistika byla v počátcích svého zrodu spojena s přírodními vědami (geologií). Název geostatistika byl poprvé použit francouzským matematikem G. Matheronem v roce 1962 a je dodnes celosvětově užíván jako označení disciplíny zahrnující specifické metody zpracování dat měřených v prostoru či v ploše. Původně se jednalo o odhad vydatnosti ložisek (např. vzácných kovů, železné rudy), později se však objevily aplikace v dalších oborech, i tam, kde vedle prostorové variability hraje roli také časová variabilita studovaných jevů (hydrogeologie, geofyzika, zemědělství a rybářství, ochrana životního prostředí, meteorologie a až nyní se přidávají i společenské vědy).

Základní přístup spočívá v tom, zda je zkoumaný jev spojitý nebo nespojitý, tomu pak odpovídá příslušný metodický aparát. Spojitost samozřejmě nespočívá v tom, že existují hodnoty měření pro každý bod plochy, ale v tom, zda můžeme usuzovat, zda daný jev je reprezentován celou zájmovou plochou. Tento článek je věnován jevům, u kterých se předpokládá jejich spojitost.

Principy prostorové analýzy spojitých jevů

Prostorová analýza dat je založena na využití měření daného jevu v určitých lokalitách a následném odhadu daného jevu v celé ploše diskursu universa. Toho se dosáhne prostřednictvím interpolace. V prostorové analýze se tedy odvozuje hodnota jevu ve všech místech plochy[2] z hodnot měření ve vybraných místech. Každé místo má pak hodnotu danou buď měřením nebo odhadem.

Z hlediska matematicko-statistického řešení existují dvě základní skupiny interpolačních metod: *deterministické a stochastické*. Oboje patří mezi *metody geostatistické*. Všech-

[1] Tato práce byla řešena v rámci grantu Ministerstva školství, mládeže a tělovýchovy č. MSM0021620831.

[2] Pojem *plocha* je zde používán jako obecný pojem, vymezený z hlediska jevu, u kterého se předpokládá prostorová variabilita.

ny tyto metody mají jedno společné východisko: jsou založeny na podobnosti zkoumaného jevu mezi blízko ležícími místy (v ploše se jedná o místa, která mají x, y souřadnice a jim odpovídající hodnoty měření označované jako z). V případě stochastických metod je v odhadu dále zabudován prvek náhodnosti, který bude v dalším textu přesněji vymezen. Tím se liší stochastické metody od deterministických – podobně jako tomu je v jiných typech úloh (např. analýze časových řad).

Použití deterministického nebo stochastického přístupu k řešení konkrétní úlohy úzce souvisí s povahou zkoumaného jevu. Obecně lze konstatovat, že přírodní procesy se prostorově analyzují a modelují spíše metodami deterministickými, zatímco sociální spíše metodami stochastickými. Je tomu tak proto, že míra komplexní složitosti systémů je u přírodních procesů jednodušší než u procesů společenských. Toto tvrzení však není možné absolutizovat. Například nadmořskou výšku (v neznámých místech) je možné odhadovat pomocí deterministických metod (polynomickou interpolací), u výskytu určité horniny nebo chemického prvku v daném místě je již nutné zabudovat prvek neurčitosti. U prostorového modelování společenských procesů lze za určitých (zjednodušujících) podmínek od prvku náhodnosti abstrahovat. U demografického jevu (který je předmětem této práce) lze předpoklad určité náhodnosti očekávat. Z metodického hlediska se často doporučuje, začít zkoumat prostorový vztah daného jevu za předpokladu *přípustného zjednodušení úlohy*, což vede k deterministickému přístupu. Proto je nutné vysvětlit oba způsoby řešení.

Jevy, které jsou v prostoru blíže k sobě, mají tendenci se sobě více podobat než jevy, které jsou prostorově vzdálenější. To je základní geostatistický princip [Tobler, 1970]. Při zvětšující se vzdálenosti od místa predikce vliv těchto míst na predikci klesá a od určité vzdálenosti je vliv těchto vzdálených míst na predikci v daném místě nulový.

Obecným úkolem každé geostatistické úlohy je zajistit dostatek měřených hodnot ve vymezené oblasti zkoumání (ať již se jedná o dílčí nebo celkovou oblast). To je pouze obecné konstatování. Jaký je dostatečný počet měření, závisí na frekvenčním rozložení zkoumané proměnné a složitosti plochy. Jestliže hodnoty jednotlivých měření nejsou zásadně odlišné a struktura plochy není zasažena podstatnými změnami, je možné interpolovat plochu především z bodů ležících blízko sebe. Z tohoto přístupu vychází *metoda inverzního vážení vzdálenosti* (anglicky Inverse Distance Weighting – dále v textu IDW), která patří mezi deterministické metody.

Metoda IDW má rovněž svoje nevýhody – např. že nevytváří „hladkou“ plochu, protože neexistuje derivace funkce ve všech jejích bodech. Proto se používají i jiná řešení prostorového modelu, založená např. na využití polynomických funkcí. Tento přístup má smysl tehdy, když prostorová struktura jevu má jednoznačný trend klesání nebo růstu v určitém směru. Ze statistického pohledu se pak úloha řeší jako minimalizace chyby při použití metody nejmenších čtverců. Pokud se měří tato chyba u polynomu prvního řádu (tedy roviny), pak se měří odchylka změřené hodnoty bodu od hodnoty interpolované rovnicí přímky (umocněné na druhou a sumarizované). Podobně jako polynomy prvního řádu, je možné použít i další – vždy s ohledem na charakteristiku zkoumané plochy.

Další krok pak spočívá v rozhodnutí, zda dochází v rámci celkového charakteru plochy k místním „zlomům“ či nikoli. Pokud tomu tak je, pak má smysl rozdělit celou plochu na dílčí oblasti, pro ty pak ze zjištěných pozorování vypočítat hodnoty a odhad celé plochy pak počítat z těchto dílčích reprezentantů. Tato metoda se nazývá *lokální polynomická interpo-*

lace. Postup bez zastoupení dílčích reprezentantů se nazývá *globální polynomická interpolace*. Vedle metody IDW a metod založených na polynomických funkcích, existuje i řada dalších.

Techniky, které byly diskutovány v předchozích příkladech, se označují jako deterministické, protože interpolace neznámé hodnoty je založena na měřeních v okolí neznámé hodnoty pomocí vhodné deterministické funkce. Druhá skupina interpolačních metod je tvořena geostatistickými metodami, jež jsou založeny na *statistickém modelu, který v sobě obsahuje společně s (prostorovou) autokorelací jevu rovněž apriorní předpoklad nejistoty (neurčitosti) mezi měřenými místy*. Tyto techniky vedou nejen k vytvoření prostorové predikce, ale umožňují rovněž určit přesnost této předpovědi.

Metoda odhadu (interpolace) – Kriging

Nejčastěji používanou metodou odhadu je metoda *Kriging*, která je podobná deterministické metodě IDW. Pro stanovení odhadu pomocí funkce Z v bodě s se souřadnicemi *x*,*y* se použijí váhy, které se označují λ. Tyto váhy jsou závislé nejen na vzdálenosti mezi měřenými body, ale také na prostorovém *vztahu* (uspořádání) mezi měřenými body. Vzájemný vztah (odhad), tak lze zapsat rovnicí $estZ(s) = \Sigma\lambda_i Z(s_i)$, kde $Z(s_i)$ hodnota funkce i-tého pozorování.

Kvantifikace prostorového vztahu je dána *(auto)kovarianční funkcí*, která se rovněž někdy označuje jako *(auto)korelační funkce*, *(auto)korelace* a zapisuje se jako

$c_k = 1/(n-k)\sum_{i-1}^{n-k}(z_i - \bar{z})(z_{i+k} - \bar{z})$, kde z_i je hodnota funkce v bodech *i*, *K* až do pozorování *n*.

S pojmem autokovariance je úzce spjat pojem *variogram*. Ten vyjadřuje autokorelaci jevu vzhledem ke vzdálenosti a směru působení autokorelace. Zapisuje se ve tvaru

$\gamma_k = 1/2(n-k)\sum_{i-1}^{n-k}(z_i - z_{i+k})$.

V případě konstantní střední hodnoty, lze základní model metody Kriging zapsat ve tvaru: $Z(s) = \mu+\varepsilon(s)$, kde s označuje (*x*, *y*) souřadnice bodu a $Z(s)$ je hodnota funkce (tj. sledovaného jevu) v daném bodě, μ je konstantní střední hodnota a $\varepsilon(s)$ je náhodná chyba. U náhodné chyby $\varepsilon(s)$ se předpokládá *stacionarita*, což znamená, že její *velikost nezávisí na místě měření ale pouze na vzdálenosti míst měření*. Pak lze odhad funkce *est*(*Z*) v předpovědním místě s_0 definovat jako: $estZ(s_0) = \sum_{i=1}^{N}\lambda_i Z(s_i)$, kde $Z(s_i)$ je zjištěná hodnota (pozorování) v i-tém místě, λ_i je neznámá váha měřené hodnoty v i-tém místě a N je počet pozorování.

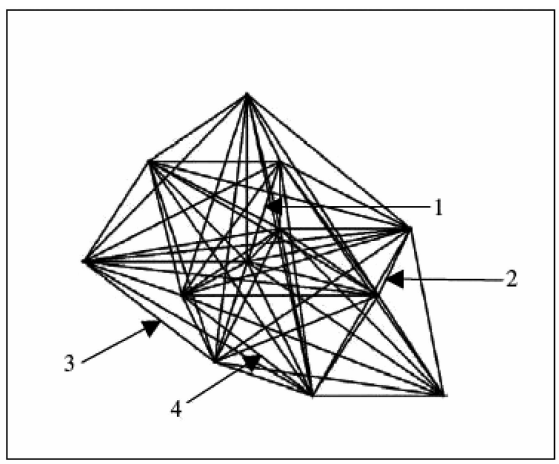
Jedná se o stejný typ odhadu jako v případě metody IDW s tím rozdílem, že v metodě IDW váha λ_i závisí výlučně na vzdálenosti od předpovědního místa. V metodě Kriging váha λ_i závisí na hodnotě dané semivariogramem, vzdálenosti od předpovědního místa a prostorových vztazích okolo místa předpovídané hodnoty.

Metoda vážení vychází z předpokladu, že suma vah je rovna jedné: $\Sigma\lambda_i = 1$. Proto je nutné dosáhnout toho, aby rozdíl mezi skutečnou hodnotou $Z(s_o)$ a předpovídanou hodnotou $\Sigma\lambda_i Z(s_i)$ byl v úhrnu minimalizovaný. To znamená minimalizovat výraz $\left(Z(s_0) - \sum_{i=1}^{N}\lambda_i Z(s_i)\right)^2$, na kterém jsou výpočtové rovnice metody Kriging založeny. Řešení

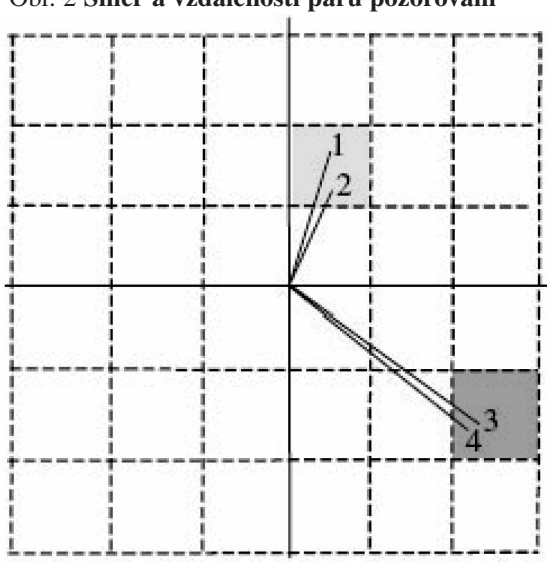
této minimalizace, za předpokladu nevychýlenosti výběru, lze zapsat jako Γ*λ = g, kde λ je vektorem pro všechny váhy $\lambda_i s$, Γ je matice modelových hodnot semivariogramu mezi všemi páry pozorování a vektor g obsahuje modelové hodnoty semivariogramu mezi místy pozorování i a j. Prvky matice Γ je tedy možné vypočítat, pokud je známa hodnota semivariogramu. Výsledný vektor *g* pak obsahuje modelové hodnoty semivariogramu mezi předpovědním místem a místem se známou hodnotou.

Aby bylo možné vypočítat hodnotu matice Γ, je nutné prozkoumat strukturu dat empirického semivariogramu. Prvním krokem je výpočet vzdáleností každého páru (viz Obr. 1 a Obr. 2, příklady vzdáleností jsou očíslovány od 1 do 4). Tato vzdálenost je založena na Euklidovské metrice a počítá se jako $d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$.

Obr. 1 **Vzdálenosti všech párů pozorování**



Obr. 2 **Směr a vzdálenosti párů pozorování**



Je zřejmé, že s rostoucím počtem pozorování počet lokalizací párových bodů prudce narůstá a výpočet se tak stává nezvládnutelným. Řešením tohoto problému je *metoda zvaná seskupování* (anglicky „binning"). Tato metoda znamená vytvoření tříd vzdáleností a následující výpočty se pak provádí nad reprezentanty (průměry) těchto vzdáleností.

V dalším kroku dochází k vytvoření modelu pomocí empirického semivariogramu. Ten je založen na rozložení skupinových reprezentantů (vytvořených metodou binning) z hlediska jejich rozptylu a vzdálenosti. Při výpočtu matice Γ se nepoužívají přímo hodnoty empirického semivariogramu ale hodnoty získané výpočtem modelového řešení. V nejjednodušším případě se použije metoda nejmenších čtverců a regresní přímka (tj. lineární regrese). Modelů zachycujících průběh prostorové závislosti je celá řada a není jednoduché určit, který z nich vybrat[3].

Závěrem se počítá statistická chyba predikce. Za předpokladu normality rozložení výběrové chyby, 95procentní interval spolehlivosti předpovědi se vypočte podle vztahu: $Předpověď\_Kriging \pm 1.96 * \sqrt{Kriging\_variance}$. Proto se předpovědní mapy zkoumaného jevu vytváří intervalově.

Při prostorové analýze jevu se předpokládá, že z*měřené hodnoty jsou výsledkem na sobě nezávislých měření*. To ale *neznamená, že události, které se měří v terénu, jsou na sobě nezávislé*. Tato závislost se nazývá autokorelace. Metoda Kriging je založena na analýze semivariogramu, kovarianční funkci (prostorové autokorelaci) a odhadu neznámých (prostorových) hodnot. Je tedy možné konstatovat, že v geostatistice se data použijí dvakrát: nejprve pro odhad prostorové autokorelace a potom pro vytvoření modelového řešení.

Celkové řešení prostorového modelu vyžaduje několik návazných kroků:

- Výpočet empirického semivariogramu řešícího kvantifikaci prostorových vztahů.
- Vytvoření modelu slouží k optimalizovanému průchodu semivariogramu (viz dále). Tento model předpokládá prostorovou korelaci dat.
- Výpočet prostorového modelu metodou Kriging. Rovnice řešící prostorový model v sobě obsahují datové matice a vektory, které závisí na prostorové autokorelaci v měřených prostorových místech. Tyto rovnice vedou k výpočtu vah vzdáleností.
- Výpočet prostorového modelu založeného na vypočtených vahách.

Stacionarita

Stacionarita v případě prostorových dat znamená, že případná závislost prostorové proměnné, sledované ve dvou místech, vyplývá ze *vzdálenosti těchto míst a nikoliv místa*, kde k měření došlo. V případě statistických pozorování se pracuje s přístupem, že jednotlivá pozorování jsou na sobě nezávislá. V případě prostorových dat je tomu jinak: jednotlivá pozorování jsou sice nezávislá, avšak hodnota pozorování v určité lokalitě souvisí s pozorováními v lokalitách sousedních, takže jednotlivá pozorování (v případě prostorové závislosti) jsou na sobě z tohoto pohledu závislá. Z toho vyplývá závěr, že *z opakovaných pozo-*

[3] "Interpolation is a black art, and interpolation methods should never be trusted. Careful researchers always interpolate several times using different methods, and choose the result they dislike least." Roger Bivand, Department of Geography, Norwegian School of Economics and Business Administration.

rování (tj. v prostoru) je možné učinit odhad společně s variabilitou sledovaného jevu a chybou zatěžující tento odhad.

V případě prostorových dat je idea stacionarity použita pro získání hodnot prostorových údajů. Existují dva typy stacionarity. První se nazývá průměrná stacionarita a předpokládá se, že průměr je konstantní mezi (prostorově) výběrovými soubory a nezávislý na jejich lokalizaci. Druhý typ stacionarity se nazývá stacionarita pro kovarianci a vnitřní stacionaritu v semivariogramu. Ve stacionaritě druhého typu platí předpoklad, že kovariance je stejná mezi jakýmikoli dvěma místy, která mají od sebe stejnou vzdálenost a směr – bez ohledu, kde jsou tato dvě místa vybrána. Vnitřní stacionarita semivariogramu je předpokladem, že rozptyl rozdílů je stejný mezi jakýmikoli dvěma místy, která mají stejnou vzdálenost a směr, bez ohledu na to, kde jsou ty dva body vybrány. Stacionarita druhého typu a vnitřní stacionarita jsou nezbytným předpokladem k odhadu prostorových závislostí, což umožňuje výpočet modelu a odhadnout chybu predikce.

Jestliže prostorový model v sobě obsahuje stacionaritu dat, je možné začít zkoumat, jak jsou data korelovaná. Tento proces se nazývá *prostorové modelování* nebo *strukturální analýza dat*, případně *variografie*. Nejprve se začíná s grafem empirického semivariogramu pro všechny páry bodů rozdělené do skupin podle vzdálenosti *h*. Empirický semivariogram je grafem průměrných hodnot vynášených na osu *y* a vzdáleností párů bodů (podle skupin) vynášených na osu *x*.

V dalším kroku se empirické hodnoty prokládají funkcí, která vytváří modelové řešení (podobně jako v regresní analýze dat), nejčastěji použitím metody nejmenších čtverců. Důležitou otázkou je volba modelu – tedy funkce, která nejlépe vystihuje průběh prostorové korelace. Často se používá model, který má tu vlastnost, že zesiluje vliv blízko u sebe ležících míst. S růstem vzdálenosti klesá význam odlehlých míst na situaci v predikované lokalitě. Základním cílem je vypočítat parametry tak, aby byly minimalizovány odchylky od jednotlivých empirických hodnot semivariogramu, ve shodě se zvoleným kritériem. Existuje řada řešení, která jsou vždy založena na nějakých vstupních předpokladech (viz dále). Empirické hodnoty vykreslené v semivariogramu se nazývají shluk (mračno) semivariogramu.

Hlavním cílem variografie je tedy prozkoumat a kvantifikovat prostorovou závislost jevů (tj. prostorovou autokorelaci). Ta vychází z obecného předpokladu, že jevy, které jsou blíže k sobě, si jsou podobnější než jevy, které jsou prostorově vzdálenější. Tudíž hodnoty bodů ležících blíže k sobě by měly být podobnější, než hodnoty u bodů vzdálenějších.

Seskupování hodnot

S rostoucím počtem pozorování prudce narůstá počet kombinací a vytvoření empirického semiovariogramu se stává prakticky nemožné. Proto jsou vzdálenosti jednotlivých míst pozorování transformovány do skupin na základě jejich *vzdálenosti a směru*. Vlastní seskupování hodnot se provádí v příslušném souřadnicovém systému, ze kterého data pochází. Tím vzniká síť, jejíž jednotlivé buňky tvoří základ pro výpočet hodnot jednotlivých skupin.

Dalším prvkem, který je nutné brát do úvahy, je to, co by se dalo nazvat „směrové zatížení“, jež může být statisticky prokazatelné a věcně vysvětlitelné nebo prokazatelné, ale

věcně nevysvětlitelné. Pokud směrové zatížení existuje, pak se hovoří o *anizotropii*, opakem pak je *izotropie*. Směrové zatížení je spjato s šířkou pásma, ve kterém se počítá.

Počet a způsob vytvoření skupin, do kterých jsou data přetříděna, má významný vliv na tvorbu empirického semivariogramu. Jestliže je při seskupování vzdálenost příliš veliká, místní působení autokorelace může zůstat skryto. Jestliže je seskupovací vzdálenost naopak příliš malá, mohou existovat třídy bez zastoupení a empirický semivariogram je zkreslen. Nejjednodušší je situace, když jsou body v prostoru rovnoměrně rozloženy. Pokud tomu tak není, je nutné velikost buňky odhadovat na základě empirického vztahu mezi vzdáleností, velikostí buňky a počtem buněk.

Semivariogram, resp. modelování kovariance, je klíčovou spojnicí mezi prostorovým popisem určitého jevu a jeho modelovým vyjádřením. Cílem modelu je prognóza hodnot v neměřených místech pomocí některé geostatistické metody. Empirický soubor poskytl informace o hodnotách v místech měření, to ale není dostatečné pro určení hodnot ve všech směrech a všech vzdálenostech. K tomu jsou určeny různé modely prostorového řešení.

Existuje celá řada funkcí, které je možné použít pro modelování empirického semivariogramu (tj. jeho nahrazení funkčním předpisem): kruhové, sférické, tetrasférické, exponenciální, kvadratické, Gaussovy, Besselovy, aj. Zvolený typ funkce ovlivňuje výsledné modelové řešení. Obecně platí, čím lépe vystihuje funkce empirický průběh, tím přesnější prostorová predikce. Ve *sférickém modelu* se s rostoucí vzdáleností vliv autokorelace snižuje a od určité vzdálenosti prakticky zaniká. V *exponenciálním modelu* se autokorelace s rostoucí vzdáleností zmenšuje a její působení zaniká až v nekonečnu.

Pro tvorbu semivariogramu jsou důležité následující charakteristiky: vzdálenost, ve které přestává autokorelace působit (anglicky „range“), a prahová hodnota autokorelační funkce („sill“), kdy autokorelace přestává působit. Počáteční hodnota vlivu se označuje jako „nugget effect“ a často se nepřekládá (je to spjato se vznikem geostatistiky v geografii).

Při pohledu na obecný model je patrné, že od určité vzdálenosti a na určité prahové úrovni je hodnota autokorelace neměnná. Teoreticky by mělo platit, že křivka by měla začínat v bodě nula, tedy že v nulové vzdálenosti bodů by měla být hodnota autokorelace nulová. Je dáno empirickou zkušeností, že rozdíl hodnot bodů ležících v těsné blízkosti vykazuje nenulovou hodnotu. To se nazývá efekt počáteční hodnoty. Ten může mít dva obecné důvody: chyba měření nebo variabilita jevu v rozmezí menším než měřeném, případně obojí.

Existují dva typy komponent, které ovlivňují kvalitu prostorové predikce: celkový trend obsažený v datech a směrový vliv, tj. to, že jev působí určitým směrem, ať je důvod (známý či neznámý) tohoto působení jakýkoli. Celkový trend je základní proces, který má vliv na odhad. Lze jej *popsat deterministickým způsobem, vhodnou matematickou funkcí a extrahovat. To se nazývá odstranění trendu predikce („detrending“).* Prostorové vazby se tak mohou po odstranění vlivu trendu lišit. Anizotropie (směrové působení proměnné) se tedy liší v případě, že se vliv trendu připouští od situace, kdy se žádný prostorový trend v datech nepředpokládá. Příčina anizotropie (směrového působení) v datech je apriori pokládána za neznámou, a proto je modelována jako náhodná chyba.

Anizotropie není obvykle definována deterministicky, protože neexistuje jediná příčina prostorové závislosti. Anizotropie je charakteristikou náhodného procesu, který vykazuje vyšší autokorelaci jevu v jednom směru proti ostatním. Autokorelační proces může mít více nestejně velkých lokálních ohnisek. Anizotropie se tak směrově liší, o izotropii se hovoří tehdy, pokud se semivariogram neliší z hlediska směrového působení jevu. Velmi často se stává, že prostorovou závislost jevů není možné vyjádřit jedním modelovým řešení, ale více – zpravidla z hlediska toho, co jsme schopni popsat.

V řadě metod Kriging (např. Jednoduchá metoda Kriging nebo Universální metoda Kriging) se pracuje s apriorním předpokladem, že data pocházejí z normálního rozložení. V těch případech, kdy tomu tak není, je nutné data transformovat. Příkladem používané transformace je Box-Coxova metoda, kterou lze zapsat jako $Y(s) = Z(s)^{\lambda} - 1/\lambda$, pro $\lambda \neq 0$, kde $Z(s)$ je původní funkce v bodě s(x,y) a $Y(s)$ její transformace. V případě použití této transformace se předpokládá, že lze data seskupit do určitého množství skupin sledovaného jevu. Jestliže tedy existuje datový soubor v určité malé oblasti, pak variabilita jevu v této oblasti může být výrazně odlišná, než variabilita v jiné větší oblasti. Po použití Box-Coxovy transformace se data více přimykají k normálnímu rozložení.

Další možnou transformační metodou je logaritmická transformace, tj. když $\lambda = 0$, pak $Y(s) = 1nZ(s)$pro $Z(s) > 0$. Tento typ transformace se používá, když jsou soubory kladně zešikmené a mají vysokou špičatost.

Prvotním zdrojem informace o rozložení dat je histogram rozložení četností, případně další speciální pohledy na data – např. graf testování normality rozložení jevu, anglicky často označovaný jako QQ graf. V QQ grafu se zkoumá rozložení teoretické proměnné (pocházející z normálního rozložení) a empirické proměnné daného jevu. V případě normality empirické proměnné je výsledkem QQ grafu přímka. Čím více se empirická křivka vzdaluje tvaru přímky, tím menší pravděpodobnost, že proměnná pochází z normálního rozložení. Z toho je patrná struktura rozložení, přesnou informaci je vhodné získat statistickým testováním normality rozložení dat.

Problém při tvorbě modelu rovněž vzniká při existenci tzv. odlehlých pozorování (anglicky outliers). Ta mohou být buď globální nebo lokální. V obou případech jde o to, že hodnota měřeného pozorování leží zcela mimo rámec ostatních okolních pozorování.

Tato odlehlá pozorování je nutné identifikovat ze dvou důvodů: buď se jedná o případ skutečné datové abnormality[4], kterou by měl model respektovat, nebo se jedná o nekorektní údaj a ten by měl být opraven. Pokud se tento problém neřeší, pak je tím ovlivněna tvorba semivariogramu a výsledná predikce jevu.

Modelová řešení pomocí metod Kriging

Jak bylo řečeno na začátku této kapitoly, stochastické modely pracují se střední hodnotou jevu $\mu(s)$ a náhodnou chybou $\varepsilon(s)$ při interpolaci prostorových dat. To lze obecně vyjádřit rovnicí ve tvaru $Z(s) = \mu(s) + \varepsilon(s)$. Apriorní omezující požadavek na druh sledovaných jevů (datových souborů) neexistuje, lze pracovat s ordinálními, kardinálními i nomi-

[4] Může jít i o náhodné kolísání v případě malých datových souborů.

nálními hodnotami bez omezení z toho důvodu, že před výpočtem modelu se stejně provede vnitřní datová transformace.

Geostatistické metody Kriging jsou založeny na autokorelaci dat – *míra prostorové závislosti je výlučnou funkcí vzdálenosti*. Bez ohledu na to, jaká prostorová variabilita v datech existuje, není možné od střední hodnoty $\mu(s)$ očekávat úplnou predikci jevu v neznámém bodě s (s x,y souřadnicemi), ale vždy je nutné do modelu zahrnout předpoklad náhodné chyby $\varepsilon(s)$. Základní požadavek zní, aby byla v průměru rovna nule. Další požadavek je, aby chyba autokorelace mezi dvěma místy, vyjádřená jako $\varepsilon(s)$ a $\varepsilon(s+h)$, nezávisela na lokalitě bodů, ale na vzdálenosti h mezi nimi.

V případě analýzy trendu, mohou nastat různé situace. Nejjednodušší je stav, kdy trend je konstantní. To lze zapsat jako $\mu(s) = \mu$, pro všechny body s. Řešení se využívá v Jednoduché metodě Kriging, kterou lze zapsat ve tvaru $Z(s) = \mu + \varepsilon(s)$. Jiná situace nastává, když trend není konstantní a když regresní koeficienty jsou neznámé.

V *Univerzální metodě Kriging* se předpokládá model ve tvaru $Z(s) = \mu(s) + \varepsilon(s)$, kde $\mu(s)$ je nějaká deterministická funkce. Zde se jedná se o polynom druhého řádu, který vystihuje trendovou složku $\mu(s)$ prostorového jevu. Předpokladem složky $\varepsilon(s)$ je její náhodnost, a tedy střední hodnota chyby $\varepsilon(s)$ by měla být rovna nule. Autokorelace prostorových dat je počítána ze složky $\varepsilon(s)$ polynomickou regresní funkcí s předpokladem, že jde o autokorelaci prostorových dat.

Existuje celá řada dalších modelů pro konkrétní situace (např. pro binární proměnnou), není cílem představovat vyčerpávajícím způsobem všechny metody.

Jiným případem jsou modely, kdy je prostorová závislost dána více než jednou proměnnou, což lze zapsat ve tvaru $Z_j(s) = \mu_j(s) + \varepsilon_j(s)$ pro j-tou proměnnou. V tom případě lze očekávat různý trend každé proměnné a kromě autokorelace $\varepsilon_j(s)$, resp. $\varepsilon_j(s)$ je nutné předpokládat i autokorelaci mezi $\varepsilon_j(s)$ a $\varepsilon_j(s)$. To vede k řešení pomocí *metod Cokriging*, kdy existuje prostorová proměnná Z_1, která je autokorelována a další prostorové proměnné Z_n, které mají rovněž své prostorové vyjádření. Tento stav je nepochybně blíže realitě, protože většina jevů neexistuje sama o sobě, ale ve vazbě na jevy jiné.

Problém spočívá ve zmnožení počtu odhadů: pro jednotlivé proměnné a pro korelaci mezi proměnnými (teoreticky lze rovněž předpokládat situaci, že proměnné nejsou mezi sebou korelovány, pak se situace dostává k předchozím metodám Kriging).

V metodě Cokriging lze zapsat model pro dvě proměnné ve tvaru $Z_1(s) = \mu_1(s) + \varepsilon_1(s)$, $Z_2(s) = \mu_2(s) + \varepsilon_2(s)$, kde μ_1, resp. μ_2 jsou neznámé konstanty a $\varepsilon_1(s)$ a $\varepsilon_2(s)$ jejich náhodné chyby, přičemž mezi Z_1 a Z_2 se nepředpokládá korelace. Stejným způsobem lze situace dále dělit na případy popsané již pro jednu proměnnou – jednotlivé metody Kriging. Požadavek na normalitu dat, tj. transformační metody, pak platí i pro tyto případy dvou a více proměnných.

Analýza modelového řešení

Jak je patrné z předchozího textu, modelové řešení je výslednicí postupně na sebe navazujících kroků, na jejichž konci je model, jehož kvalita by měla být ověřena. Východiskem

ověřování je empirický datový soubor, jehož zkoumání vede k potvrzení (případně k zpřesnění) parametrů modelového řešení.

Semivariogram a kovariance kvantifikují základní úvahu o podobnosti jevů ležících blízko sebe. V obou případech se měří intenzita statistické korelace (tj. prostorové závislosti) jako funkce závislosti. Semivariogram je definován jako $\gamma(s_i,s_j)=0{,}5\mathrm{var}(Z(s_i)-Z(s_i))$ a sleduje se v něm vnitřní korelovanost veličiny v různých místech. Jestliže tedy dva body s_i a s_j leží blíže k sobě (tj. ve smyslu $d(s_i,s_j)$), pak diference změřených hodnost $Z(s_i)-Z(s_i)$ bude menší, než když jsou tyto body od sebe více vzdálené.

Kovarianční funkce je definována jako $C(\boldsymbol{s}_i,\boldsymbol{s}_j)$, kde $\mathrm{cov}(Z(s_i)-Z(s_i))$ znamená kovarianci. Mezi semivariogramem a kovarianční funkcí existuje vztah, který je možné zapsat následujícím způsobem: $\gamma(s_i,s_j) = h - C(s_i,s_j)$. Z toho vyplývá, že pro prostorovou predikci lze využít obou forem zápisu prostorové variability jevu. Kovarianci (semivariogram) není možné postihnout libovolnou funkcí ale takovou, kde chyba ε nabývá nezáporných hodnot.

Kriging metody umožňují vypočítat náhodnou chybu. Výpočet pomocí jednotlivých metod vede k tomu, že pro jednu lokalitu (stejné x, y souřadnice) se získají rozdílné výsledky[5] s rozdílnou složkou $\varepsilon(s)$. Odhad je pak možné zapsat jako $Z(s) = \mu(s) + \varepsilon(s) + \delta(s)$, kde $\delta(s)$ je právě chyba měření. Počáteční efekt je složen z vlastní variability, která se označuje jako mikrovariabilita prostorového jevu a variability $\delta(s)$, neboli chyby měření. Protože se pracuje s proměnnou definovanou v ploše, lze očekávat, že semivariogram (tedy i kovarianční funkce) bude závislý jednak na vzdálenosti, jednak na *směru působení*. Předpokládá se, že existují dva body s_i a s_j a vektor, který je odděluje, se označí jako $s_i - s_j$. Existuje-li v datech anizotropie, pak je velikost tohoto vektoru v jiném (např. kolmém) směru jiná. Izotropický model je ve všech směrech stejný. Po vyhodnocení vlivu anizotropie je možné použít empirický semivariogram a kovarianční funkci k odhadu parametrů konkrétní metody Kriging.

V případě, že se jedná o datový soubor více proměnných, je třeba vytvořit kovarianční model všech proměnných. Kovarianční funkce mezi k-tým a m-tým datovým souborem je definována jako $C_{km}(s_i,s_j) = \mathrm{cov}(Z_k(s_i),Z_m(s_i))$. Problém spočívá v tom, že model bývá zpravidla asymetrický: $C_{km}(s_i,s_j) \neq C_{mk}(s_i,s_j)$.

Různé modely vedou k různé kovarianci a vliv anizotropie pak modelové řešení ještě více ztíží. Při vědomí těchto skutečností je třeba zadávat parametry modelového řešení tak, aby se empirické řešení v maximálně možné míře přibližovalo teoretickému modelu.

Hodnocení modelu

Neexistuje obecný předpis na zkoumání prostorové variability dat. Jestliže pochází data ze souboru, ve kterém není zjištěn převažující směr (anizotropie), pak se do modelu zahrnují data ze všech směrů bez rozdílu. Není-li tomu tak, pak je možné do datového modelu zahrnovat data pouze z určitého prostorově vymezeného okruhu. Další možností je vytvoření datových podmnožin – je-li pro tento přístup nějaké odůvodnění. To znamená, že je možné omezit horní i dolní hranici pozorování na určité oblasti.

[5] S tím souvisí i problém malých čísel.

Odhad parametrů prostorového modelu vychází ze všech pozorování – byť s nimi může pracovat určitým vymezeným způsobem (viz předchozí odstavec). Jednou z možností, jak kvalitu navrženého modelu ověřit, je zkoumání příspěvku jednotlivých pozorování k celkové predikci. To je možné zajistit tak, že predikční model postupně vynechává pro predikci jedno pozorování za druhým, a poté se hodnotí jejich příspěvek pro predikci. V případě dobré predikce platí, že takto získané body se zobrazují v grafu na ose *x* a *y* ve stejné hodnotě (viz měřítko os *x* a *y*). Standardizovaná chyba pak pochází z normálního rozložení.

Zjišťování intenzity prostorové závislosti

Základním nástrojem pro zjišťování intenzity prostorové závislosti jsou statistiky, které analyzují a vyhodnocují proces a výsledek *prostorového shlukování dat* (clustering). Samotný princip prostorové autokorelace vytváří předpoklady pro tvorbu shluků – ve statistikách se řeší otázka intenzity shlukovacích mechanismů a prostorová variabilita jevu. Ta může být buď uniformní, postupně vedoucí k jednomu místu, která se dá dobře vyjádřit jednou statistikou, nebo lokální, kdy vzniká více shluků, a pak je nutné vyhodnocovat místní vztah vůči dalším blízkým lokalitám.

Statistiky, které měří intenzitu tohoto jevu, vycházejí z věcné hodnoty jevu v daném místě a z indikátoru prostorové podobnosti. Prostorová podobnost je formulována prostřednictvím tzv. prostorové matice vah, s prvky matice w_{ij}, které jsou nenulové pro sousední hodnoty (pojem „sousední“ hodnota je dán věcným vymezením úlohy). Věcnou hodnotu jevu je možné vyjádřit pomocí x_i a x_j, kde x_i a x_j jsou pozorování v místech *i* a *j*.

Pravděpodobně nejčastěji používanou statistikou, která měří prostorovou intenzitu jevu je tzv. *Moranovo I*, kterou lze zapsat ve tvaru $\mathrm{I} = \left[\sum_i \sum_j z_i z_j w_{ij} / S_0\right] / \left[\sum_i z_i^2 / \mathrm{N}\right]$, kde z_i je i-tá odchylka od průměru proměnné z_i, N je počet pozorování a S_o je normalizační faktor rovný sumě všech vah $S_0 = \sum_i \sum_j w_{ij}$. Vedle globální statistiky (tzv. globální Moranovo I), existuje i její lokální verze (lokální Moranovo I), která umožňuje určit místní shluky a případná odlehlá pozorování (outliers). Pro každé místo pozorování je vypočtena statistika $\mathrm{I} = 1/\mathrm{m}\left[\sum_j z_j w_{ij}\right]$, kde m je tzv. konstantní škálovací faktor. Výpočet umožňuje definovat statistickou signifikaci jevu v dané lokalitě, společně s typem jevu: vazba nízká-vysoká hodnota, vysoká-vysoká, resp. nízká-nízká hodnota jevu – v závislosti na okolní situaci.

Výpočet této statistiky je založen na standardním algebraickém počítání průměrů a variability – ve smyčce pro všechny hodnoty, takže výsledkem je vážený průměr počítaný z hodnost v okolních lokalitách. Z uvedeného postupu řešení je patrné, že objem výpočtu je velký – i v dnešní době výkonné výpočetní techniky. V případě, že počet pozorování je větší ($\mathrm{N} > 1000$), je obtížné se dopočítat výsledku. V těch případech se úloha řeší použitím náhodné permutace dat. Výpočet je založen na empirické distribuci statistiky, která se získá po prostorovém náhodném výběru a výpočtu z původních pozorování. Efektivní implementace náhodné permutace daného jevu je klíčová pro vlastní výpočet lokální i globální statistiky.

Co dodat závěrem. Metodické postupy jsou jedna věc, jejich aplikace pak vede k dalšímu porozumění zkoumané problematiky. Nezbývá tedy, než se těšit na budoucí aplikace geostatistiky, které jistě v průběhu doby na statistickém úřadě vzniknou.

Literatura

[1] Armstrong, M. Basic Linear Geostatistics, Springer, 1998.

[2] Bolstad, P. GIS Fundamentals, Eider Press, s. 395 – 463, 2005.

[3] Gribov, A. et al. Modeling the Semivariogram: New Approach, Methods Comparison and Case Study, ESRI white paper, 2001.

[4] Gribov, A., Krivoruchko, K. Geostatistical Interpolation and Simulation with non-euclidean distances, ESRI white paper, 2004.

[5] Gribov, A., Krivoruchko K. Geostatistical Mapping with Continuous Moving Neighborhood, ESRI white paper, 2004.

[6] Jones, H. Population geography, Paul Chapman Publishing, 1990.

[7] Kolektiv autorů. GIS, UsingArcGis Geostatistical Analysis, ESRI press, s. 49 – 239, 2005.

[8] Korčák, J. Geografie obyvatelstva ve statistické syntéze, Universita Karlova, 1973.

[9] Kraus, J., Rychtaříková, J. (Ed). Atlas sčítání 2001, PřFUK, s. 40, 2005.

[10] Kraus, J. Regionální diferenciace plodnosti, Demografie 45, s. 263 – 268, 2004.

[11] Krivoruchko, K. et al. Creating Exposure Maps Using Kriging, ESRI white paper, 2004.

[12] Krivoruchko, K. et al. A New Method for Handling the Nugget Effect in Kriging, ESRI white paper, 2005.

[13] Krivoruchko, K. Assessing the Uncertainty Resulting from Geoprocessing Operations, ESRI white paper, 2004.

[14] Longley, P. A. et al. Geographic Information Systems and Science, John Wiley&Sons, s. 363 – 382, 2005.

[15] Longley, P. A., Batty, M. Advanced Spatial Analysis, CASA, 2005.

[16] Martin, D. Geographic Information Systems: socio-economic applications, Routledge, 1996.

[17] Tobler W. A computer simulating urban growth in the Detrit region, Economic Geography 46, s. 234 – 240, 1970.

[18] Tuček, J. Geografické informační systémy, Computer Press, s. 53 – 157, 1998.

Jaroslav Kraus, Český statistický úřad, Na padesátém 81, 100 82 Praha 10, e-mail: jaroslav.kraus@czso.cz

Abstract

One of the methods of statistical data analysis, which has been intensively developing in the last decade, is geostatistics. The main solution of the spatial (geostatistical) model is based on that the phenomenons, which lie closer together are more similar than phenomenons that are more distant, which is the basic principle of geostatistics. Another basis in application of geostatistics is that territorial

changes on the level of examined phenomenon exist and it does not concern accidental events. The aim of each geostatistical task is to ensure enough measured values in the specified area of observation (weather it concerns a partial or whole area). If the values of individual measurements are not fundamentally different and the structure of the surface is not affected by significant changes, it is possible to interpolate the area particularly from points lying close together. The method of inversive distance weighing (IDW) is based on this approach; it is often used as an approximation of the final spatial prediction. The most often used method of prediction is the Kriging method. In order to determine the estimate by function, weights, which are dependent, not only the distance between measured points but also on the spatial relationship (organisation) between measured points are used. During the spatial analysis of a phenomenon, it is expected that measured values are a result of mutually independent measurements. This dependency is called autocorrelation. The Kriging method is based on the semivariogram analyses, co variation function (spatial autocorrelation) and an estimate of unknown (spatial) values. The basic instrument for finding out the intensity of spatial dependence are statistics, which analyse and evaluate the process and results of spatial clustering of data for e.g. Morane I, where the question of intensity of the clustered mechanisms and spatial phenomenon variability is dealt with.

Key words: geostatistics, spatial modelling, Kriging method.