UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA
KATEDRA MATEMATICKÉ ANALÝZY A APLIKACÍ MATEMATIKY

BAKALÁŘSKÁ PRÁCE

Geostatistika pro kompoziční data

Vedoucí bakalářské práce:
doc. RNDr. Karel Hron, Ph.D.
Rok odevzdání: 2014

Vypracovala:
Kristýna Kovalová
ME, III. ročník

Prohlášení

Prohlašuji, že jsem bakalářskou práci zpracovala samostatně pod vedením pana doc. RNDr. Karla Hrona, Ph.D. s použitím uvedené literatury.

V Olomouci dne 27. března 2014

Poděkování

Ráda bych zde poděkovala vedoucímu bakalářské práce doc. RNDr. Karlu Hronovi, Ph.D. za odborné vedení, obrovskou ochotu a nekonečnou trpělivost. Dále bych ráda poděkovala Mgr. Pavlu Holečkovi za pomoc s typografickým systémem TEX, ve kterém je práce napsána. V neposlední řadě si velký dík zaslouží moje rodina a přítel.

Obsah

Úvod

V životě se často setkáváme s informacemi, které mají prostorový nebo prostorově - časový charakter. Takováto data, jenž nesou tematické informace poutané k určitému místu nebo času, nazýváme též geodata. Vzhledem k tomu, že se s prostorovými daty setkáváme denně, byly vyvinuty metody, pomocí nichž je můžeme analyzovat, zkoumat, odhadovat jejich vývoj, atd.

Pro práci s výše definovanými daty byl vyvinut obor nazývaný geostatistika. Geostatistika umožňuje popis prostorové spojitosti přírodních jevů, k čemuž využívá upravené metody klasické statistiky. Cílem této práce bude přehledné zpracování základních pojmů geostatistiky a jejich následná aplikace na reálných datech s přihlédnutím k specifikám příslušného výběrového prostoru.

V první kapitole si uvedeme předpoklady pro práci s prostorovými daty a seznámíme se s základními pojmy, které se týkají geostatistiky. Zavedeme si pojem variogram, který patří mezi základní geostatistické nástroje pro zobrazení, modelování a prostorovou korelaci měřených dat. Dále si uvedeme nástroj pro určování lokálního odhadu, tzv. kriging.

Druhá kapitola čtenáře seznámí s konkrétním případem geostatistiky pro kladná data, která se v praxi často vyskytují (atmosférický tlak, denní srážky, ...) a uvedeme si přístup, který nám pomůže uvedená data analyzovat.

V závěrečné kapitole aplikujeme teoretické poznatky z prvních dvou kapitol na reálná data. Jedná se o měření získaná v rámci mezinárodního projetku geochemického mapování v oblasti poloostrovu Kola. Související výpočty budou provedeny pomocí volně dostupného statistického softwaru R.

1. Geostatistika

V této kapitole si představíme základní pojmy, které budou nezbytně nutné pro následnou práci s kladnými daty. Geostatistika je intenzivně se rozvíjející část aplikované matematiky, která má původ v těžebním průmyslu padesátých let 20. století. Rozvinula se, aby pomohla zlepšit, popř. nahradit detailní propočty, vycházející z deterministických geologických modelů. První krůčky v této oblasti pocházejí z Jižní Afriky a jsou spojené s prací inženýra těžebního průmyslu Daniela G. Krige a statistika H. S. Sichel [11].

Původně byl tento obor vyvinutý pro řešení názorových problémů při odhadu zásob železné rudy. V sedmdesátých letech 20. století se díky rozvoji výpočetní techniky rozšířil i do jiných částí vědních oborů zabývajících se prostorovými daty, jsou to například hydrogeologie, meteorologie, oceánografie, geochemie, geografie, ochrana životního prostředí, krajinná ekologie, zemědělství a další [3]. Při analýze dat pomocí geostatistiky musíme mít na paměti tři věci:

Popis dat - je zapotřebí prozkoumat, zda se jedná o data prostorově nebo časově určená, o data s více proměnnými a zkontrolovat, zda se v datech vyskytují odlehlé hodnoty, které mohou zkreslovat skutečnou strukturu dat.

Interpretace - grafické výstupy získané z numerických výsledků jsou dále interpretovány pomocí minulých zkušeností s podobnými daty. Interpretace prostorové a časové struktury, asociace a běžné vztahy mezi proměnnými jsou zabudovány do modelu, který odpovídá analyzovaným datům.

Odhad - nalezme-li odpovídající model dat, dalším krokem je odhadnout hodnoty v různých měřítkách, v různých oblastech, které nejsou součástí měřených dat. Metody, které provádějí tyto odhady, jsou založeny na metodě nejmenších čtverců.

1.1. Regionalizované proměnné a náhodná veličina

Regionalizované proměnné je vhodné umístit do pravděpodobnostního rámce, jelikož v pravděpodobnostním modelu je regionalizovaná proměnná $z(\mathbf{x})$ považována za realizaci náhodné veličiny $Z(\mathbf{x})$ (tj. nekonečná třída náhodných pro-

měnných vytvořených ve všech bodech $\mathbf{x}$ daného definičního oboru D). Výhodou tohoto přístupu je, že se budeme snažit charakterizovat pouze prosté vlastnosti náhodné veličiny $Z(\mathbf{x})$ a nikoli jednotlivých realizací $z(\mathbf{x})$.

Z mnoha oblastí vědy vyplývá, že data jsou buď časově nebo prostorově závislá a nebo obojí. Takováto data jsou tedy často mnohorozměrná (s více proměnnými), tj. bylo naměřeno několik veličin a ty je třeba přezkoumat. Datový soubor pak může mít následující tvar,

$$\begin{bmatrix} t_1 & x_1^1 & x_1^2 & x_1^3 & z_1^1 \ldots z_1^i \ldots z_1^N \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ t_\alpha & x_\alpha^1 & x_\alpha^2 & x_\alpha^3 & z_\alpha^1 \ldots z_\alpha^i \ldots z_\alpha^N \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ t_n & x_n^1 & x_n^2 & x_n^3 & z_n^1 \ldots z_n^i \ldots z_n^N \end{bmatrix}.$$

V této datové matici je celkový rozsah výběru označen písmenem n, tedy $\alpha = 1, 2, 3, ..., n$. Rozdílné proměnné jsou označeny indexem i a počet proměnných je N, takže $i = 1, 2, 3, ..., N$. Pro představu můžeme říci, že vzorky (odpovídající řádkům datové matice) mohou být kusy skály stejného objemu pořízené geology a proměnné by pak mohly reprezentovat chemické nebo fyzikální měření provedené na skalách.

Jednotlivá pozorování mohou být pořízena v různých okamžicích t_α a na různých místech $\mathbf{x}_\alpha$.

1.1.1. Regionalizované proměnné

Nechť byla naměřena pouze jedna vlastnost v různých prostorových objektech (tedy $N = 1$) a doba, po kterou bylo provedeno měření, nebyla zaznamenána. V tomto případě můžeme index i a časové souřadnice t_α vynechat. Máme n pozorování označených

$$z(\mathbf{x}_\alpha), \quad \text{s} \quad \alpha = 1, \ldots, n,$$

získaných v místech $\mathbf{x}_\alpha$. Výběr objektů v definičním oboru D může být považován za část z většího souboru objetků. V případě, že objekty jsou body v prostoru

(popř. v rovině), je možné v tomto definičním oboru uvažovat i nekonečně mnoho pozorování. Při této možnosti, kdy máme potenciálně nekonečně mnoho pozorování stejného druhu, upustíme od indexů a definujeme jednotlivé regionalizované proměnné (zkratka REV = regionalized variable) jako

$$z(\mathbf{x}) \quad \text{pro všechna} \quad \mathbf{x} \in D.$$

Soubor dat $\{z(\mathbf{x}_\alpha), \alpha = 1, \ldots, n\}$ je vnímán jako soubor realizací několika regionalizovaných proměnných. Studie regionalizované proměnné obvykle zahrnuje alespoň dva geometrické aspekty: zaprvé definiční obor D, ve kterém jsou regionalizované proměnné definovány a zadruhé geometrický opěrný bod, pro který je každý výběr regionalizovaných proměnných definovaný. Tento opěrný bod může být například část půdy, která byla analyzována.

Každá naměřená hodnota v datovém souboru má umístění v definičním oboru D a nazýváme ji regionalizovaná hodnota.

1.1.2. Náhodná funkce

Vzhledem k regionalizovaným hodnotám ve všech bodech v daném definičním oboru je související funkce $z(\mathbf{x})$ pro $\mathbf{x} \in D$ regionálně variabilní. Soubor hodnot

$$\{z(\mathbf{x}), \mathbf{x} \in D\}$$

může být považován za jeden výběr z nekonečné řady náhodných proměnných (jedna náhodná proměnná v každém bodě definičního oboru). Skupina náhodných proměnných

$$\{Z(\mathbf{x}), \mathbf{x} \in D\}$$

se nazývá náhodná funkce (RAF = random function).

Na obrázku 1 vidíme, že REV $z(\mathbf{x})$ je jednou realizací RAF $Z(\mathbf{x})$. Regionalizovaná hodnota $z(\mathbf{x}_0)$ v určitém místě $\mathbf{x}_0$ je realizací náhodné proměnné $Z(\mathbf{x}_0)$, která je sama členem nekonečné řady náhodných proměnných RAF $Z(\mathbf{x})$. Velké písmeno „Z“ se používá k označení náhodné proměnné a malé „z“ se používá pro jejich realizace. Bod $\mathbf{x}_0$ je libovolný bod definičního oboru, který může, ale nemusí být součástí výběru.

$$\begin{array}{ccccc} \mathbf{RAF} & Z(\mathbf{x}) & \Rightarrow & Z(\mathbf{x}_0) \\ \Downarrow & \Downarrow & & \Downarrow \\ \mathbf{REV} & z(\mathbf{x}) & \Rightarrow & z(\mathbf{x}_0) \end{array}$$

Obrázek 1: Realizací náhodné funkce je regionalizovaná hodnota.

1.2. Rozdělení pravděpodobnosti

V tomto modelu náhodný mechanismus $Z(\mathbf{x}_0)$, působící v daném bodě $\mathbf{x}_0$ definičního oboru, vytváří realizace na základě rozdělení pravděpodobnosti F, charakterizovaného distribuční funkcí

$$P(Z(\mathbf{x}_0) \leq z) = F_{\mathbf{x}_0}(z),$$

kde P je pravděpodobnost, že hodnota Z v bodě $\mathbf{x}_0$ je menší nebo rovna pevné hodnotě z. Distribuční funkce dvou náhodných veličin $Z(\mathbf{x}_1)$ a $Z(\mathbf{x}_2)$ na dvou různých místech je potom

$$P(Z(\mathbf{x}_1) \leq z_1, Z(\mathbf{x}_2) \leq z_2) = F_{x_1,x_2}(z_1, z_2),$$

kde výraz na pravé straně rovnosti označuje pravděpodobnost, že $Z(\mathbf{x}_1)$ je menší nebo rovna z_1 a současně $Z(\mathbf{x}_2)$ je menší nebo rovna z_2. Stejným způsobem můžeme definovat distribuční funkci n proměnných umístěných na n různých místech,

$$P(Z(\mathbf{x}_1) \leq z_1, \ldots, Z(\mathbf{x}_n) \leq z_n) = F_{x_1,\ldots,x_n}(z_1, \ldots, z_n).$$

Takto jsme vybudovali velmi obecný model, který je schopen popsat jakýkoli proces v přírodě nebo v technologii. V praxi máme k dispozici jen málo dat z jedné nebo několika realizací RAF, a takto bude tedy pouze obtížně možné odvodit všechny jedno - a vícerozměrné distribuční funkce pro všechny množiny bodů. Zjednodušení je tedy nutné a je zajištěno myšlenkou stacionarity.

1.2.1. Striktní stacionarita

Pro úplnost uvedeme definici stacionarity. Stacionaritou rozumíme, že vlastnosti RAF zůstávají stejné, i když daná množina n bodů přechází z jedné části

definičního oboru do druhé. Nazývá se to *invariance vůči posunutí*. Konkrétně říkáme, že RAF $Z(\mathbf{x})$ je striktně stacionární, jestliže pro každý soubor n bodů $\mathbf{x}_1, \ldots, \mathbf{x}_n$ (kde n je libovolné přirozené číslo) a pro každý vektor $\mathbf{h}$ platí

$$F_{x_1,\ldots,x_n}(z_1, \ldots, z_n) = F_{x_1+h,\ldots,x_n+1}(z_1, \ldots, z_n),$$

tj. posun bodu v daném směru nemění mnohorozměrnou distribuci. Je třeba poznamenat, že stacionarita je vlastnost modelu náhodné funkce a nikoli regionalizované proměnné. Prakticky můžeme říci, že „REF je stacionární“, ale takto zkráceně říkáme, že „REV můžeme považovat za realizaci stacionární RAF“.

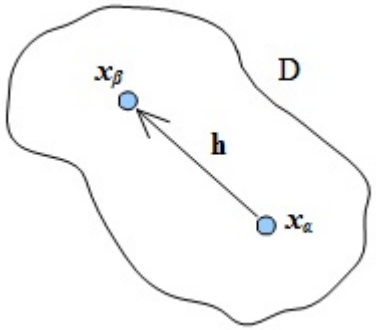
Striktní stacionarita vyžaduje specifikaci „vícebodové“ distribuce pro všechny množiny bodů $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. Lehčí bude uvažovat pouze dvojice bodů v definičním oboru a pokusit se charakterizovat první dva momenty, nikoli plnou distribuci. Samozřejmě taková strategie je ideální v případě normálního rozdělení, kde první dva momenty plně charakterizují rozdělení pravděpodobnosti. Potom tedy můžeme předpokládat stacionaritu prvních dvou momentů proměnné, nazývanou *stacionarita druhého řádu*, nebo předpokládáme stacionaritu prvních dvou momentů rozdílu dvojic hodnot ve dvou bodech, nazývanou *vnořená stacionarita*, která vede k pojmu variogram.

1.3. Variogramy

Variogram je základním geostatistickým nástrojem pro vizualizaci a prostorovou autokorelaci[1] naměřených hodnot. Udává míru rozptylu, jak se mění hodnoty proměnné při změně vzdálenosti z bodu $\mathbf{x}$ do bodu $\mathbf{x} + \mathbf{h}$ při neměnícím se $\mathbf{h}$.

Pro definici experimetálního variogramu si nejdříve musíme představit pojem *variogram cloud*, tzv. mračno bodů. Variogram cloud je sám o sobě silný nástroj pro zkoumání vlastností prostorových dat. Dvojice výběrových hodnot jsou hodnoceny výpočtem druhé mocniny rozdílu mezi hodnotami. Poté jsou dvojicím výběrových hodnot v prostoru přiřazeny výsledné rozdíly a vznikne variogram

[1] Autokorelace je statistická korelace mezi prostorově náhodnými proměnnými stejného typu a vlastností, kde korelace závisí na vzdálenosti a/nebo směru mezi lokacemi [2].



Obrázek 2: Vektor spojující dva body v dvourozměrném prostoru.

cloud. Cloud, neboli mrak, je rozdělen do tříd podle rozdělení v prostoru a průměr rozdílů v každé třídě tvoří posloupnost hodnot experimentálního variogramu.

Variabilitu regionalizované proměnné $z(\mathbf{x})$ měříme tak, že počítáme rozdíl dvojic datových hodnot z_α a z_β, které odpovídají pozicím bodů $\mathbf{x}_\alpha$ a $\mathbf{x}_\beta$ v prostorovém definičním oboru D. Míra týkající se nepodobnosti dvou hodnot, označená γ^*, je

$$\gamma^*_{\alpha\beta} = \frac{(z_\alpha - z_\beta)^2}{2}.$$

tj. polovina rozdílu mezi těmito dvěma hodnotami.

Dva body $\mathbf{x}_\alpha$, $\mathbf{x}_\beta$ v geografickém prostoru mohou být spojeny pomocí vektoru $\mathbf{h} = \mathbf{x}_\alpha - \mathbf{x}_\beta$, jak je znázorněno na obrázku 2.

Nepodobnost γ^* závisí na vzdálenosti a orientaci dvojice bodů popsané vektorem $\mathbf{h}$

$$\gamma^*(\mathbf{h}) = \frac{1}{2}(z(\mathbf{x}_\alpha + \mathbf{h}) - z(\mathbf{x}_\alpha))^2.$$

Vzhledem k tomu, že nepodobnost vzniká ze součtu čtverců, znaménko vektoru $\mathbf{h}$, tj. pořadí, ve kterém uvažujeme body $\mathbf{x}_\alpha$ a $\mathbf{x}_\beta$ nehraje roli. Nepodobnost je tedy symetrická vzhledem k $\mathbf{h}$,

$$\gamma^*(-\mathbf{h}) = \gamma^*(+\mathbf{h}).$$

Proto budou nepodobnosti v grafickém znázornění charakterizovány pomocí absolutních hodnot vektorů $\mathbf{h}$. S využitím všech výběrových párů v datovém souboru

(až do vzdálenosti poloviny průměru definičního oboru), se vytváří mapa nepodobnosti γ^* vzhledem k prostorové separaci prostřednictvím vektoru $\mathbf{h}$, která se nazývá variogram cloud.

1.3.1. Experimentální variogram

Průměrné nepodobnosti $\gamma^*(\mathbf{h})$ mohou být vytvořeny pro danou třídu vektorů η_k seskupením všech n_k bodových dvojic, které mohou být spojeny pomocí vektoru $\mathbf{h}$, patřícího do η_k. Hledáme třídu η_k skupiny vektorů, jejichž délky jsou v zadaném intervalu délek a jejichž orientace je stejná až do zadané tolerance úhlu. Průměrná nepodobnost vzhledem k vektoru třídy η_k je hodnotou experimentálního variogramu

$$\gamma^*(\eta_k) = \frac{1}{2n_k} \sum_{\alpha=1}^{n_k} (z(\mathbf{x}_\alpha + \mathbf{h}) - z(\mathbf{x}_\alpha))^2 \quad \text{s} \quad \mathbf{h} \in \eta_k.$$

Vypočítáme ji tedy jako součet čtverců eukleidovských vzdáleností mezi každou dvojicí bodů dělený dvojnásobkem počtu bodů. V praxi někdy bereme jako průměrnou funkci γ^* místo třídy η_k přímo vektory $\mathbf{h}$.

Na základě experimentálního variogramu hledáme teoretický variogram, což je funkce, která nejlépe aproximuje hodnoty vykreslené v experimentálním variogramu. Použití teoretického variogramu zajišťuje, že rozptyl kterékoliv lineární kombinace náhodně vybraných hodnot je kladný. Nahrazení teoretickým variogramem se většinou provádí na základě výstupu empirického variogramu.

1.3.2. Teoretický variogram

Změny v prostoru náhodné funkce $Z(\mathbf{x})$ lze popsat pomocí rozdílů mezi hodnotami v párech bodu $\mathbf{x}$ a $\mathbf{x}+\mathbf{h}$

$$Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x}),$$

které se nazývají přírůstky. Existují dva předpoklady o přírůstcích:

- průměr přírůstků $m(\mathbf{h})$, tzv. drift, je neměnný pro každou změnu daného vektoru $\mathbf{h}$ v rámci definičního oboru,

- rozptyl přírůstků má konečnou hodnotu $2\gamma(\mathbf{h})$ v závislosti na délce a orientaci daného vektoru $\mathbf{h}$, ale ne na jeho pozici v definičním oboru.

To znamená, že pro každou dvojici bodů $\mathbf{x}$, $\mathbf{x} + \mathbf{h} \in D$ máme

$$E[Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x})] = m(\mathbf{h}) = 0,$$

$$\text{var}[Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x})] = 2\gamma(\mathbf{h}).$$

Tyto dvě vlastnosti náhodné funkce poskytují definici pro teoretický variogram

$$\gamma(\mathbf{h}) = \frac{1}{2}E[(Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x}))^2].$$

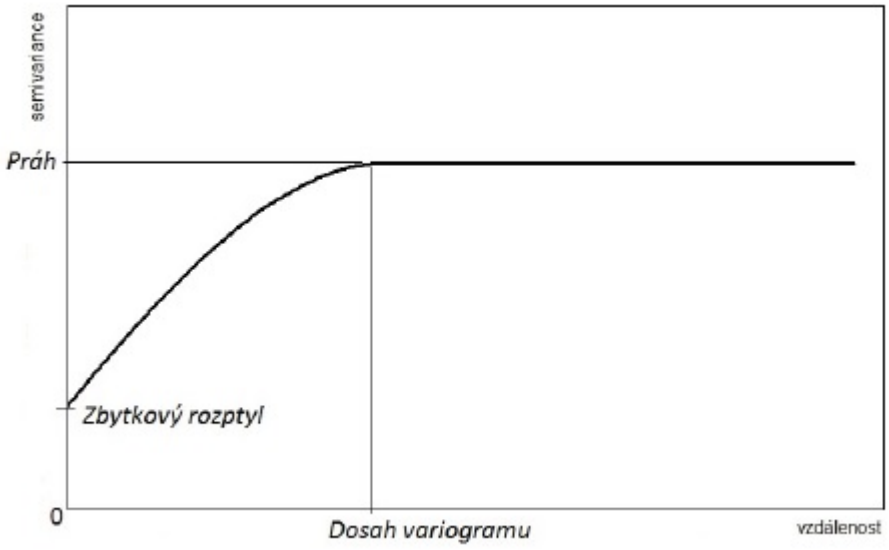
Teoretický variogram, viz obrázek 3, lze popsat pomocí tří základních parametrů: prahová hodnota (sill), zbytkový rozptyl (nugget) a dosah (range). Rozdíl mezi zbytkovým roztylem a prahovou hodnotou označujeme jako částečný práh (partial sill). Prahová hodnota je plochá část variogramu a stanovuje celkový rozptyl v datech. Dosah variogramu je vzdálenost, od které se hodnota semivariance[2] už nemění a určuje maximální vzdálenost, na které jsou data korelována [4]. Pro hledání nejvhodnějšího proložení experimentálního variogramu používáme metodu nejmenších čtverců. Nejčastějšími modely variogramu jsou sférický (předpokládá, že s rostoucí vzdáleností se vliv autokorelace zmenšuje a od určité vzdálenosti prakticky zaniká), exponenciální (působení autokorelace zaniká až v nekonečnu) nebo gaussovský, viz. obrázek 4.

Hodnota variogramu na počátku je podle definice nulová,

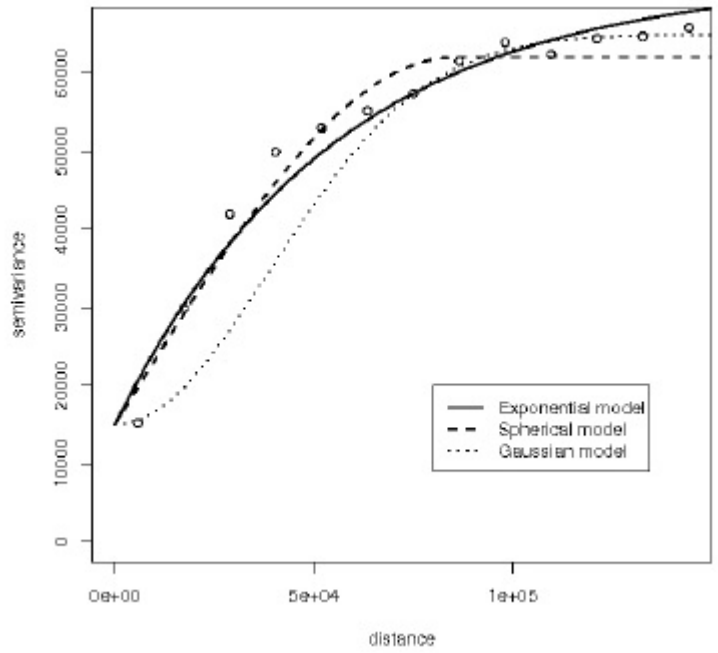
$$\gamma(\mathbf{0}) = 0.$$

Tato vlastnost říká, že na nulové vzdálenosti by měl být nulový rozptyl, ovšem v praxi proložená křivka protíná osu y ve vyšší hodnotě. Tato hodnota, nazývaná zbytkový rozptyl, odpovídá náhodné složce, která je složená z chyb měření a rozptylu na menší vzdálenosti než je námi použitá vzorkovací vzdálenost.

[2]Semivariance je polovina průměrné kvadratické odchylky sledovaných dat v jednotlivých třídách [2].



Obrázek 3: Parametry teoretického variogramu.



Obrázek 4: Modely teoretického variogramu.

Hodnoty variogramu jsou nezáporné

$$\gamma(\mathbf{h}) \geq 0,$$

a variogram je dokonce sudá funkce,

$$\gamma(-\mathbf{h}) = \gamma(\mathbf{h}).$$

Variogram roste pomaleji než $|\mathbf{h}|^2$, to znamená

$$\lim_{|\mathbf{h}|\to\infty} \frac{\gamma(\mathbf{h})}{|\mathbf{h}|^2} = 0.$$

Jinak bychom nemohli předpokládat, že drift je roven nule.

1.4. Kovarianční funkce

Kovarianční funkce $C(\mathbf{h})$ je definována na základě předpokladu stacionarity prvních dvou momentů (průměr a kovariance) náhodné funkce, tedy

$$\begin{aligned} &E[Z(\mathbf{x})] = m \quad \text{pro všechna} \quad \mathbf{x} \in D, \\ &E[Z(\mathbf{x}) \cdot Z(\mathbf{x}+\mathbf{h})] - m^2 = C(\mathbf{h}) \quad \text{pro všechna} \quad \mathbf{x}, \mathbf{x}+\mathbf{h} \in D. \end{aligned}$$

Vlastnosti kovarianční funkce:

1. Kovarianční funkce je omezená a její absolutní hodnota nepřesahuje rozptyl

   $$|C(\mathbf{h})| \leq C(\mathbf{0}) = \mathrm{var}(Z(\mathbf{x})).$$

2. Stejně jako variogram je sudá funkce, ale na rozdíl od variogramu může nabývat záporných hodnot.

   $$C(-\mathbf{h}) = C(+\mathbf{h}).$$

3. Kovarianční funkce dělená rozptylem se nazývá korelační funkce

   $$\rho(\mathbf{h}) = \frac{C(\mathbf{h})}{C(\mathbf{0})},$$

   která je samozřejmě omezená

   $$-1 \leq \rho(\mathbf{h}) \leq 1.$$

4. Kovarianční funkce je pozitivně definitní. To znamená, že použijeme-li kovarianční funkci C($\mathbf{h}$) pro výpočet rozptylu lineární kombinace $n+1$ náhodných veličin $Z(\mathbf{x}_\alpha)$, je tento rozptyl vždy kladný. To je nutně spojeno pozitivně semidefinitní maticí $\mathbf{C}$ kovariancí

$$\text{var}\left(\sum_{\alpha=0}^{n} w_\alpha Z(\mathbf{x}_\alpha)\right) = \sum_{\alpha=0}^{n}\sum_{\beta=0}^{n} w_\alpha w_\beta C(\mathbf{x}_\alpha - \mathbf{x}_\beta) = \mathbf{w}^\top \mathbf{C}\ \mathbf{w} \geq 0$$

pro každou sadu bodů $\mathbf{x}_\alpha$ a každou sadu vah w_α (sestavených do vektoru $\mathbf{w}$). Vztah mezi kovarianční funkcí a variogramem je poté následující:

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h}).$$

1.4.1. Příklady kovariančních funkcí

Představíme si několik modelů kovarianční funkce. Jsou definovány pro izotropní (tj. invariantní vůči rotaci) náhodné funkce [1].

1. **Nugget-effekt model** - kovarianční funkce $C(\mathbf{h})$, která modeluje nespojitost na počátku je tzv. nugget-effect model:

$$C_{nug}(\mathbf{h}) = \begin{cases} b, & \text{pro} \quad |\mathbf{h}| = 0 \\ 0, & \text{pro} \quad |\mathbf{h}| > 0 \end{cases}$$

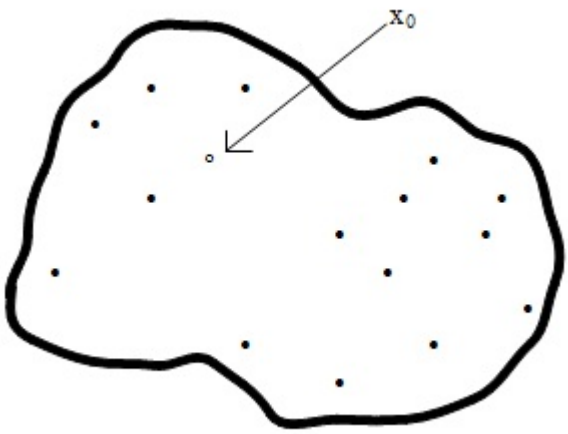
kde b je kladná hodnota a musí platit, že

$$\lim_{|\mathbf{h}|\to\infty} \gamma(\mathbf{h}) = b.$$

2. **Exponenciální kovarianční funkce** - v tomto modelu klesá kovarianční funkce exponenciálně s rostoucí vzdáleností bodů a je definován takto:

$$C_{exp}(\mathbf{h}) = b\,\exp\left(-\frac{|\mathbf{h}|}{a}\right) \quad \text{s} \quad a, b > 0.$$

Parametr a určuje, jak rychle kovarianční funkce klesá. Pro hodnotu $|\mathbf{h}| = 3a$ se kovarianční funkce sníží o 95% své původní hodnoty, takže tato vzdálenost označuje, v jak velkém okolí se naměřené hodnoty ještě ovlivňují. Exponenciální model je spojitý a asymptoticky se blíží k nule pro $|\mathbf{h}| \to \infty$.



Obrázek 5: Definiční obor s nepravidelně rozmístěnými výběrovými body (černé tečky) a umístění bodu $\mathbf{x}_0$, který nás zajímá.

3. **Sférický model** - tento model je definován následovně:

$$C_{sph}(\mathbf{h}) = \begin{cases} b\left(1 - \frac{3}{2}\frac{|\mathbf{h}|}{a} + \frac{1}{2}\frac{|\mathbf{h}|^3}{a^3}\right) & \text{, pro } \ 0 \leq |\mathbf{h}| \leq a, \\ 0 & \text{, pro } \ |\mathbf{h}| > a. \end{cases}$$

1.5. Základní kriging

Základní kriging (ZK) je nejpoužívanější metodou typu kriging. Slouží pro odhad hodnoty na bodě nebo na celém definičním oboru, pro který je znám variogram, na základě dat v blízkém okolí odhadované lokality.

1.5.1. Bodový kriging

K odhadu hodnoty v místě $\mathbf{x}_0$, kde nebylo provedeno měření, jak je znázorněno na obrázku 5, nám pomůže lineární kombinace hodnot ze sousedních n bodů a vah w_α,

$$Z^*_{ZK}(\mathbf{x}_0) = \sum_{\alpha=1}^{n} w_\alpha Z(\mathbf{x}_\alpha).$$

Je zřejmé, že součet omezujících vah musí být roven jedné, protože v daném případě, kdy všechny hodnoty dat jsou rovny stejné konstantě, musí být i odha-

dovaná hodnota rovna této konstantě, proto:

$$\sum_{\alpha=1}^{n} w_\alpha = 1.$$

Díky jednotkovému součtu vah je též zaručena nestrannost,

$$E[Z^*(\mathbf{x}_0) - Z(\mathbf{x}_0)] = \left[\sum_{\alpha=1}^{n} w_\alpha Z(\mathbf{x}_\alpha) - Z(\mathbf{x}_0) \cdot \underbrace{\sum_{\alpha=1}^{n} w_\alpha}_{1} \right]$$

$$= \sum_{\alpha=1}^{n} w_\alpha E[Z(\mathbf{x}_\alpha) - Z(\mathbf{x}_0)] = 0,$$

protože střední hodnoty přírustků jsou nulové.

Rozptyl odhadu $\sigma_E^2 = \mathrm{var}(Z^*(\mathbf{x}_0) - Z(\mathbf{x}_0))$ je rozptyl lineární kombinace

$$Z^*(\mathbf{x}_0) - Z(\mathbf{x}_0) = \sum_{\alpha=1}^{n} w_\alpha Z(\mathbf{x}_\alpha) - 1 \cdot Z(\mathbf{x}_0) = \sum_{\alpha=0}^{n} w_\alpha Z(\mathbf{x}_\alpha),$$

s váhou $w_0 = -1$ a

$$\sum_{\alpha=0}^{n} w_\alpha = 0.$$

Proto podmínka, že váhy (číslované od 1 do n) dávající součet jedna rovněž říká, že použití variogramu je povoleno i při výpočtu rozptylu chyby odhadu. Tím, že minimalizujeme rozptyl odhadu s omezením pro váhy, získáme systém rovnic pro základní kriging systém (ZK)

$$\underbrace{\begin{pmatrix} \gamma(\mathbf{x}_1 - \mathbf{x}_1) & \dots & \gamma(\mathbf{x}_1 - \mathbf{x}_n) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma(\mathbf{x}_n - \mathbf{x}_1) & \dots & \gamma(\mathbf{x}_n - \mathbf{x}_n) & 1 \\ 1 & \dots & 1 & 0 \end{pmatrix}}_{\mathbf{C}_n} \underbrace{\begin{pmatrix} w_1^{ZK} \\ \vdots \\ w_n^{ZK} \\ \lambda_{ZK} \end{pmatrix}}_{\mathbf{w}} = \underbrace{\begin{pmatrix} \gamma(\mathbf{x}_1 - \mathbf{x}_0) \\ \vdots \\ \gamma(\mathbf{x}_n - \mathbf{x}_0) \\ 1 \end{pmatrix}}_{\mathbf{d}_n}$$

$\mathbf{C}_n$ ... charakterizují rozdíly mezi naměřenými daty,

$\mathbf{w}$ ... zahrnuje váhy přiřazené jednotlivým měřením,

λ ... Lagrangeův multiplikátor,

$\mathbf{d}_n$ ... charakterizuje rozdíly mezi měřenými hodnotami a odhadnutými hodnotami.

Neznámou v tomto případě je vektor vah, který jednoduše vypočítáme pomocí vztahu $\mathbf{w} = \mathbf{C}_n^{-1} d_n$ a následně dosadíme do rovnice $Z^*_{ZK}(\mathbf{x}_0)$.

Rozptyl odhadu základního krigingu je pak

$$\sigma^2_{ZK} = \lambda_{ZK} - \gamma(\mathbf{x}_0 - \mathbf{x}_0) + \sum_{\alpha=1}^{n} w_\alpha^{ZK} \gamma(\mathbf{x}_\alpha - \mathbf{x}_0).$$

Základní kriging je přesný interpolátor v tom smyslu, že pokud hodnota $\mathbf{x}_0$ je totožná s místem, kde bylo provedeno měření, pak bude odhadovaná hodnota shodná s hodnotou naměřenou, tedy

$$Z^*(\mathbf{x}_0) = Z(\mathbf{x}_\alpha), \quad \text{jestliže} \quad \mathbf{x}_0 = \mathbf{x}_\alpha.$$

Metodu kriging využíváme jako interpolační metodu k odhadu hodnot na pravidelné mřížce pomocí nepravidelně rozmístěných dat.

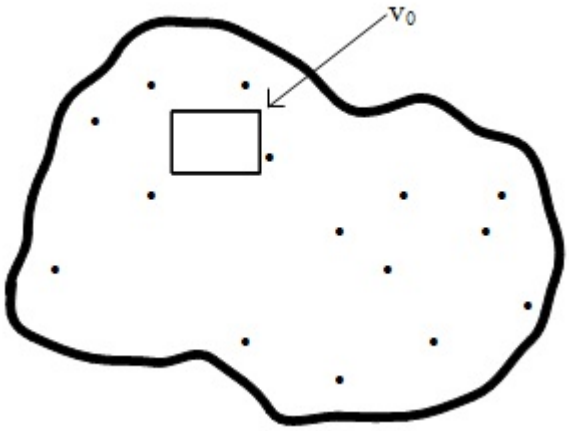
1.5.2. Blokový kriging

Základní kriging může být použit pro odhad bloku hodnoty namísto bodu hodnoty, jak je to znázorněno na obrázku 6. Při odhadu blokové hodnoty z bodových hodnot

$$Z^*_{v_0}(\mathbf{x}_0) = \sum_{\alpha=1}^{n} w_\alpha Z(\mathbf{x}_\alpha)$$

je základní kriging upraven následujícím způsobem pro blokový kriging (BK)

$$\begin{pmatrix} \gamma(\mathbf{x}_1 - \mathbf{x}_1) & \dots & \gamma(\mathbf{x}_1 - \mathbf{x}_n) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma(\mathbf{x}_n - \mathbf{x}_1) & \dots & \gamma(\mathbf{x}_n - \mathbf{x}_n) & 1 \\ 1 & \dots & 1 & 0 \end{pmatrix} \begin{pmatrix} w_1^{BK} \\ \vdots \\ w_n^{BK} \\ \lambda_{BK} \end{pmatrix} = \begin{pmatrix} \overline{\gamma}(\mathbf{x}_1 - \mathbf{v}_0) \\ \vdots \\ \overline{\gamma}(\mathbf{x}_n - \mathbf{v}_0) \\ 1 \end{pmatrix},$$



Obrázek 6: Definiční obor s nepravidelně rozmístěnými výběrovými body (černé tečky) a blok $\mathbf{v}_0$, pro který by měla být odhadnuta hodnota.

kde pravá strana nyní obsahuje průměrné variogramy $\overline{\gamma}(\mathbf{x}_\alpha, \mathbf{v}_0)$ každého výběrového bodu s pozovaným blokem.

Příslušný rozptyl blokového krigingu je

$$\sigma^2_{BK} = \lambda_{BK} - \overline{\gamma}(v_0, v_0) + \sum_{\alpha=1}^{n} w_\alpha^{BK} \overline{\gamma}(\mathbf{x}_\alpha, v_0).$$

2. Geostatistika pro kladná data

Nyní si představíme hlavní část bakalářské práce. Uvedeme si konkrétní případ geometrie kladných dat, který je speciálním případem geometrie kompozičních dat. V praxi se totiž často setkáváme s využitím geostatistiky s kladnými hodnotami, k čemuž je správné pochopení geometrických vlatností výběrového prostoru nezbytné. Příkladů kladných dat máme mnoho, od propustnosti k hodnotě atmosférického tlaku, od denních srážek k množství škodlivin atd. V této kapitole si tedy uvedeme geostatistiku pro kladná data, geostatiku pro kompoziční data najdeme v [10].

2.1. Kriging regionalizovaných kladných proměnných

Často jsou regionalizované kladné proměnné zpracovány předběžně použitím logaritmů [8]. Krigingové odhady jsou poté zpětně převedeny pomocí exponenciální funkce, často ovšem bez hlubšího pochopení geometrických souvislostí, např. i při srovnání s dalšími možnými trasformacemi dat (např. jejich odmocněním). Proto vznikl nový přístup, který bere v úvahu povahu výběrového prostoru a škály dat. Vzhledem k tomu, že mohou být tyto vlastnosti vhodně zachyceny pomocí algebraické struktury eukleidovského prostoru, můžeme definovat bodový i blokový kriging pro kladné proměnné (náhodné funkce). V následující kapitole se tak těmto uvahám budeme věnovat podrobněji.

2.1.1. Označení

Pro následující text si zavedeme nové značení. Nechť $\mathbf{x} \in D \subset \mathbb{R}^q$ je bod (nebo střed bloku $\mathbf{v}$) v definičním oboru D časoprostoru reálného prostoru, kde $q \in \{1, 2, 3, 4\}$. Nechť $Z(\mathbf{x}) \in \mathbb{R}_+$ je kladná náhodná funkce (RAF). Nechť $\{z(\mathbf{x}_1), z(\mathbf{x}_2), \ldots, z(\mathbf{x}_N)\}$ je pozorovaný výběr hodnot této náhodné funkce. Logaritmicky transformovaný soubor je $\zeta(\mathbf{x}_n) = \ln z(\mathbf{x}_n)$, který tvoří výběr $Z(\mathbf{x})$ v reálném prostoru. V této situaci je tedy geostatistika aplikována na logaritmy. Proto, pokud $\mathbf{h}$ představuje vektor posunutí mezi body definičního oboru, kova-

riance $C(\mathbf{h})$, variogramy $\gamma(\mathbf{h})$ a drift $m(\mathbf{h})$ jsou definovány jako hodnoty $\zeta(\mathbf{x}_n)$. Nechť pak ζ^*_{ZK} a $\sigma^2_{\zeta,ZK}$ označují základní kriging a rozptyl.

Navrhovaný odhad hodnoty $Z(\mathbf{x}_0)$ náhodné funkce v bodě je

$$z^+_K = \exp(\zeta^*_K),$$

nezávisle na tom, jaká krigingová metoda je pro tento odhad použita.

2.1.2. Výběrový prostor a měřítko pro kladná data

Argumenty ve prospěch odhadu $z^+_K = \exp(\zeta^*_K)$, oproti klasickému odhadu vyplývají z pojetí výběrového prostoru a měřítka. Tyto dva pojmy jsou v praxi používány ojediněle, ačkoli jejich role je ve statistice všeobecně známá. Výběrový prostor náhodné funkce je tvořen jejím oborem hodnot. Výběrový prostor provedeného měření je tedy rozsah hodnot, které mohou být pozorovány. Kladná data mají jako výběrový prostor kladnou reálnou osu, označenou $\mathbb{R}_+$. Soubor N lokalit kladných RAF má potom jako výběrový prostor $\mathbb{R}^N_+$. Měřítko spojené s náhodnou proměnnou je založeno na kritériu rozdílu mezi možnými hodnotami a je vybráno na základě předchozích zkušeností. Kladná měřítka pro kladná data zahrnují aditivní a multiplikativní měřítko. V praxi se ovšem daleko častěji vyskytuje měřítko multiplikativní.

Uvažujme např. měření počtu srážek ve dvou lokalitách. Řekněme, že jsme při dvou měřeních srážkových úhrnů obdrželi dvě dvojice hodnot, 5 a 10 mm pro první měření, respektive 100 a 105 mm při druhém měření. Pokud bude úkolem určit rozdíl mezi oběma srážkovými úhrny v daných lokalitách, potom v prvním případě bychom zřejmě došli k závěru, že ve druhé lokalitě napadlo dvojnásobně srážek ve srovnání s první lokalitou, zatímco ve druhém případě lze spíše hovořit o tom, že pršelo hodně a přibližně stejně v obou lokalitách. To znamená, že pro intuitivní určení rozdílů počtu srážek v daných lokalitách nebyly až tak podstatné rozdíly absolutních hodnot srážkových úhrnů jako spíše podíly mezi těmito hodnotami.

2.1.3. Pojmy z lineární algebry na $\mathbb{R}_+^N$

Formální definice eukleidovského prostoru lze nalézt v každé základní učebnici algebry, a proto budou následující pojmy uvedeny pouze pro úplnost. Z hlediska této práce je účelem nalézt strukturu pro pozitivní data.
Nechť E je výběrový prostor, v tomto případě $E = \mathbb{R}_+^N$ (v případě jedné lokality bychom pak obdrželi $N = 1$). Objekty tohoto prostoru, vektory, jsou označeny tučně malými písmeny, tj. $\mathbf{x} = (x_1, x_2, \ldots, x_p)$.
Nechť $\oplus : \mathbb{R}_+^N \times \mathbb{R}_+^N \to \mathbb{R}_+^N$ je vnitřní operace a $\odot : \mathbb{R} \times \mathbb{R}_+^N \to \mathbb{R}_+^N$ vnější operace. Tyto operace tvoří na $\mathbb{R}_+^N$ strukturu vektorového prostoru a jsou analogií sčítání vektorů,

$$\mathbf{z} = \mathbf{x} \oplus \mathbf{y} = (x_1 \cdot y_1, x_2 \cdot y_2, \ldots, x_N \cdot y_N),$$

a násobení vektoru skalárem

$$\mathbf{z} = \lambda \oplus \mathbf{x} = (x_1^\lambda, x_2^\lambda, \ldots, x_N^\lambda).$$

Neutrálním prvkem této struktury je vektor $\mathbf{n} = (1, 1, \ldots, 1)$, protože vždy splňuje $\mathbf{x} \oplus \mathbf{n} = \mathbf{x}$. Vezmeme-li jakýkoli vektor $\mathbf{x}$, vektor k němu opačný je $\ominus\mathbf{x} = (-1) \odot \oplus\mathbf{x} = (\frac{1}{x_1}, \frac{1}{x_2}, \ldots, \frac{1}{x_N})$.
V této souvislosti je lineární kombinace k vektorů $\mathbb{R}_+^N$ definována jako

$$\bigoplus_{i=1}^{k} \lambda_i \odot \mathbf{x}_i = \lambda_1 \odot \mathbf{x}_1 \oplus \lambda_2 \odot \mathbf{x}_2 \oplus \ldots, \oplus\lambda_k \odot \mathbf{x}_k.$$

Báze $\mathbb{R}_+^N$ je množina N vektorů takových, že:

1. žádný z nich není lineární kombinací zbytku množiny
2. všechny vektory prostoru jsou lineární kombinací báze.

Kanonická báze tohoto prostoru je tvořena N vektory tvaru

$$\mathbf{e}_i = (1, 1, \ldots, e, \ldots, 1),$$

kde číslo e je umístěno v $i-$té pozici, $i = 1, 2, \ldots, N$. Reálné koeficienty $\xi_1, \ldots, \xi_N$ potřebné k jednoznačnému vyjádření vektoru $\mathbf{x}$ jako lineární kombinace bázových

vektorů se nazývají souřadnice $\mathbf{x}$ vzhledem k této bázi. Vektory souřadnic budeme označovat řeckými písmeny, jako $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N)$. Kladný reálný orthant, kde kromě vnitřní a vnější operace definujeme též skalární součin

$$\langle \mathbf{x}, \mathbf{y} \rangle_+ = \sum_{i=1}^{N} \ln x_i \cdot \ln y_i,$$

je reálný eukleidovský prostor a značíme jej $\{\mathbb{R}_+^N, \oplus, \odot, \langle \cdot, \cdot \rangle_+\}$. Skalární součin je nástrojem pro stanovení úhlů a délek. Dva vektory jsou kolmé, jestliže jejich skalární součin je nulový. Norma nebo-li délka vektoru $\mathbf{x}$ je potom $||\mathbf{x}||_+ = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_+}$. Báze $\{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$ je ortonormální tehdy a jen tehdy, když všechny její vektory mají jednotkovou normu a jsou navzájem kolmé, což je vyjádřeno jako $\langle \mathbf{v}_i, \mathbf{v}_j \rangle_+ = \delta_{ij}$. Souřadnice vektoru $\mathbf{x}$ v ortonormální bázi jsou pohodlně vypočítány jako $\xi_i = \langle \mathbf{v}_i, \mathbf{x} \rangle_+$. Skalární součin lze také použít k definici vzdálenosti na $\mathbb{R}_+^N$,

$$d_+(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} \ominus \mathbf{y}||_+ = \sqrt{\langle \mathbf{x} \ominus \mathbf{y}, \mathbf{x} \ominus \mathbf{y} \rangle_+} = \sqrt{\sum_{i=1}^{N} \ln^2 \frac{x_i}{y_i}},$$

která je invariantní vůči posunutí, $d_+(\mathbf{x} \oplus \mathbf{z}, \mathbf{y} \oplus \mathbf{z}) = d_+(\mathbf{x}, \mathbf{y})$ a platí též $d_+(\lambda \odot \mathbf{x}, \lambda \odot \mathbf{y}) = |\lambda| \cdot d_+(\mathbf{x}, \mathbf{y})$. Toto popisuje multiplikativní měřítko na $\mathbb{R}_+^N$, jinak řečeno, je vhodná pro kladná data charakterizovaná relativním měřítkem.

2.1.4. Míra na $\mathbb{R}_+$

Míra, označená $\lambda(\cdot)$ na množině E je funkce, která:

1. přiřazuje kladnou hodnotu každé možné neprázdné množině a nulovou hodnotu prázdné množině,

2. splňuje, že pro každou posloupnost disjunktních množin $\{A_i\} \subset E$ je míra sjednocení těchto množin $\lambda(\bigcup_i A_i) = \sum_i \lambda(A_i)$ (tzv. $\sum -$aditivní vlastnost).

Například nejjednodušší a nejčastější podmnožina $\mathbb{R}$ je interval a standardní míra na tomto prostoru je délka tohoto intervalu, která se počítá jako vzdálenost mezi dvěma krajními body intervalu. V $\mathbb{R}^2$ nebo $\mathbb{R}^3$ se pak jedná o plochu, respektive objem dané oblasti. Jedná se o speciální případy Lebesqueovy míry. Z tohoto důvodu pro interval v $\mathbb{R}_+$, definovaný jeho krajními body x a y, je přirozená míra $\lambda_+(\overline{xy}) = |\ln x - \ln y|$. Zobecněním přirozené míry na $\mathbb{R}_+^N$ je

$$\lambda_+(\overline{\mathbf{xy}}) = \prod_{i=1}^{N} |\ln x_i - \ln y_i|.$$

Analogicky se tato míra nazývá Lebesgueova míra na $\mathbb{R}_+^N$, narozdíl od standardní Lebesgueovy míry, která je definována na $\mathbb{R}^N$.

2.1.5. Pravděpodobnost a statistika na $\mathbb{R}_+$

Jedno z nejčastějších rozdělení na $\mathbb{R}_+$ je lognormální. Kladná proměnná $Z \sim L(\mu, \sigma^2)$ se řídí lognormálním rozdělením, pokud její logaritmus je normálně rozdělená náhodná veličina se střední hodnotou μ a rozptylem σ^2. Její hustota vzhledem k Lebesgueově míře je dána jako

$$f_Z(z) = \frac{dP(z)}{d\lambda(z)} = \frac{1}{z\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{1}{2}\left[\frac{\ln z - \mu}{\sigma}\right]^2\right).$$

Je dobře známo, že její střední hodnota je

$$E[Z] = \int_{\mathbb{R}} z \cdot f_Z(z) d\lambda(z) = \exp(\mu + \frac{1}{2}\sigma^2).$$

Toto běžně označujeme jako $E-$typ odhadu Z, totiž, že argument z' minimalizuje očekávanou čtvercovou vzdálenost $E[d^2(Z, z')]$. Ovšem odtud vidíme, že $E-$typ odhadu závisí na volbě vzdálenosti. Totéž lze říct o rozptylu

$$\mathrm{var}[Z] = \int_{\mathbb{R}} (Z - E[Z])^2 \cdot f_Z(z) d\lambda(z) = E^2[Z] \cdot (\exp(\sigma^2) - 1).$$

Všimněme si, že eukleidovská vzdálenost odpovídá aditivnímu měřítku. Pokud uvažujeme multiplikativní měřítko, jako v tomto případě, ukazuje se jako vhodnější vyjádřit tuto hustotu vzhledem k Lebesgueově míře na $\mathbb{R}_+$,

$$f_Z^+(z) = \frac{dP(z)}{d\lambda_+(z)} = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{1}{2}\left[\frac{\ln z - \mu}{\sigma}\right]^2\right).$$

Vzhledem k izomorfismu mezi všemi reálnými eukleidovskými prostory stejné dimenze vyplývá, že pravděpodobnostní úvahy a statistická analýza by měly být aplikovány na souřadnice Z vzhledem k libovolné ortonormální bázi. V případě kladné proměnné pak tuto bázi tvoří triviální vektor $\mathbf{e}_1 = (e)$. Popsaný způsob uvažování nazýváme principem práce v souřadnicích. Všimněme si, že normální hustota na $\mathbb{R}_+$ se tímto principem řídí, jak ukazují následující výrazy pro střední hodnotu

$$E_+[Z] =: E[\ln Z] \odot \mathbf{e}_1 = \exp\left(\int_{\mathbb{R}} \ln Z \cdot d\,P(\ln Z)\right) = \exp(\mu)$$

a rozptyl

$$\mathrm{var}_+[Z] =: \mathrm{var}[\ln Z] = \sigma^2.$$

Z těchto charakteristik může být jen střední hodnota interpretována v původním prostoru a to jen proto, že je vyjádřena ve výše uvedené bázi. Zobecnění těchto výsledků do mnohorozměrného případu je jednoduché – pomocí mnohorozměrného normálního rozdělení a $N-$rozměrných souřadnic. Z toho důvodu je kovariance mezi dvěmi kladnými proměnnými definována jako

$$\mathrm{cov}_+[\mathbf{Z}, \mathbf{Y}] =: \mathrm{cov}[\ln Z, \ln Y] = \sigma_{ZY} = \sigma_{YZ}.$$

Pokud jsou tyto dvě pozitivní proměnné společně normálně rozděleny na $\mathbb{R}_+^2 (N = 2)$,

$$\mathbf{X} = \begin{bmatrix} Y \\ Z \end{bmatrix} \sim N_+^N \left(\boldsymbol{\mu} = \begin{bmatrix} \mu_Y \\ \mu_Z \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{YY}^2 & \sigma_{YZ}^2 \\ \sigma_{ZY}^2 & \sigma_{ZZ}^2 \end{bmatrix}\right),$$

pak je jejich hustota

$$f_{\mathbf{X}}^+(\mathbf{x}) = \frac{1}{\sqrt{|\boldsymbol{\Sigma}|(2\pi)^p}} \exp\left(-\frac{1}{2}(\ln \mathbf{x} - \boldsymbol{\mu})^t \cdot \boldsymbol{\Sigma}^{-1} \cdot (\ln \mathbf{x} - \boldsymbol{\mu})\right),$$

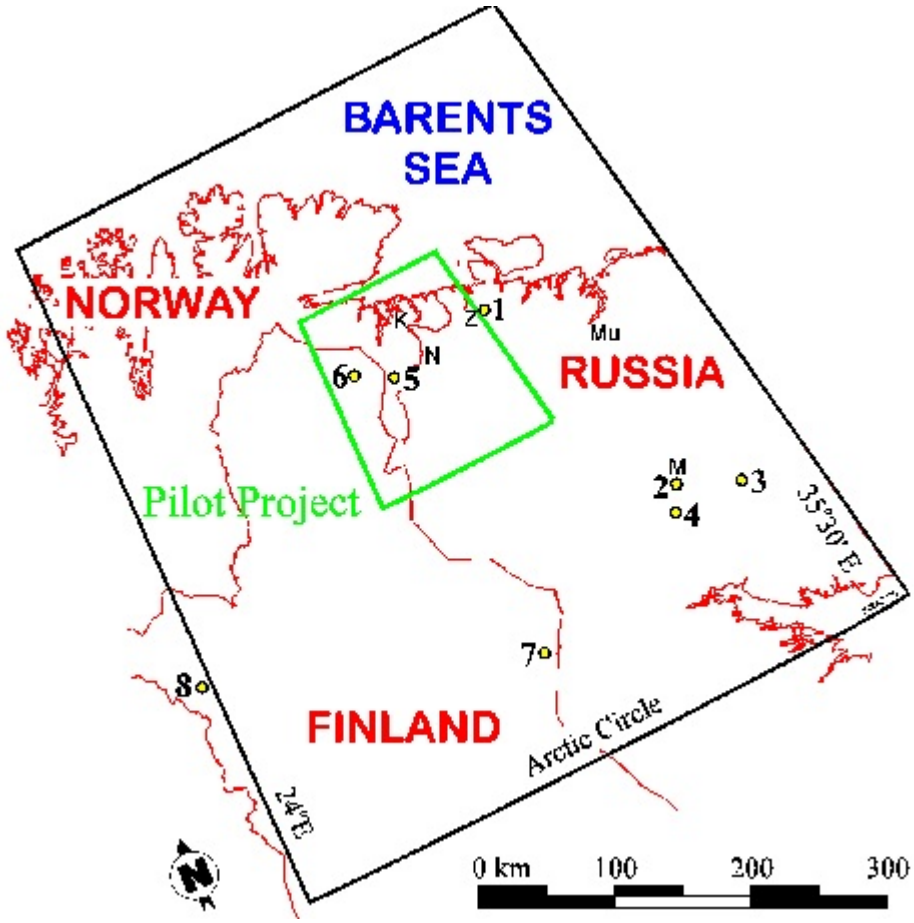
kde $\ln \mathbf{x} = (\ln z, \ln y)$ označuje vektor logaritmů složek $\mathbf{x}$. Všimněme si, že tvar hustoty $f^+_{\mathbf{X}}(\mathbf{x})$ odpovídá libovolné volbě N jako dimenze výběrového prostoru.

Poznamenejme, že analogicky bychom nyní mohli pro kladné proměnné definovat též všechny pojmy geostatistiky, představené v minulé kapitole. Např. pro kladnou náhodnou funkci Z, která je zobrazením kartézského součinu definičního oboru $D \subset \mathbb{R}^q$ a množiny výsledků náhodného pokusu Ω do množiny kladných reálných čísel $\mathbb{R}_+$, tj. $Z : D \times \Omega \longrightarrow \mathbb{R}_+$, bychom definovali variogram jako $\gamma_+(\mathbf{h}) = \frac{1}{2}\mathrm{Var}_+[Z(\mathbf{x}+\mathbf{h}) \ominus Z(\mathbf{x})]$. Jelikož ovšem v praxi pracujeme přímo v souřadnicích, odkazujeme pro detailní rozbor této situace na literaturu [8].

3. Aplikace s geologickými daty

Nyní přejdeme k praktické části bakalářské práce. Výše uvedené teoretické poznatky aplikujeme na reálná data získaná z velkého mezinárodního projektu, věnovaného geochemickému mapování na poloostrově Kola, probíhajícího v letech 1992-1998 [6]. Jednalo se tak o sběr a vyhodnocení geologických vzorků v oblasti západního regionu Murmanska a přilehlých oblastí Finska a Norska. Hlavní činností projetku bylo mapování těžkých kovů a radioaktivity, znečistění pozemních a vodních ekosystémů a vliv hlavních průmyslových činností na poloostrově na ekosystém.

Jedná se o jednu z nejkomplexnějších studií geochemických aspektů životního



Obrázek 7: Oblast projektu.

prostředí. Oblast tohoto projektu zahrnuje, jak už bylo zmíněno, nejsevernější část Norska, Finska a západní části poloostrova Kola. Tento region představuje jednu z nejvíce znečistěných a zároveň jednu z nejzachovalejších částí Evropy.

Poloostrov Kola, podle kterého byl pojmenován celý projekt, leží na severozápadě Ruska [5]. Na severu ho obklopuje Barentsovo moře, na východě a jihu Bílé moře. Poloostrov je tvořen převážně krystalickými horninami předkambria a starších prvohor. Jednou z částí poloostrova je Murmanský blok, který byl jedním z hlavních předmětů studie projektu. Zde, stejně jako na zbytku poloostrova, je velké bohatství rud a nerostů (např. železná ruda, slída, barevné kovy atd.). V naší práci se pak zaměříme na geochemickou analýzu nadmořské výšky v uvedené oblasti.

Výpočty budeme provádět pomocí volně šiřitelného statistického softwaru R [9]. Hlavní výhodou tohoto softwaru je, že je volně dostupný a také práce s ním je poměrně jednoduchá. Prvním krokem bude otevření knihoven určených pro načtení dat a následnou práci s nimi.

```
> library(geoR)
> library(StatDA)
```

Naměřená data byla odebraná v pěti různých vrstvách, my se zaměříme na svrchní vrstvu půdy, tzv. O-horizon. K tomu použijeme následující příkaz:

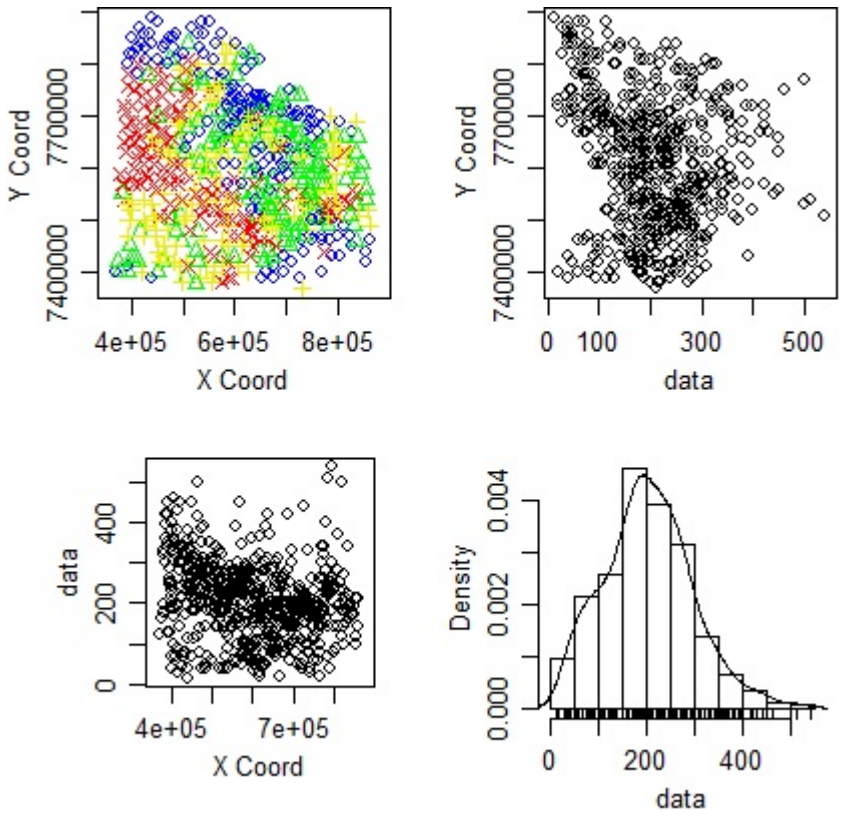
```
> data(ohorizon)
> head(ohorizon)
```

Nyní si načteme data, která chceme zkoumat, ale nejdříve si vysvětlíme zkratky, které v příkazu použijeme:

- XCOO - představuje zeměpisnou šířku
- YCOO - představuje zeměpisnou délku
- ELEV - na těchto datech je zaznamenaná nadmořská výška

```
> x=ohorizon[,c("XCOO","YCOO","ELEV")]
```

Dalším krokem bude, že si námi zvolená data převedeme na nové datové objekty tzv. geodata.



Obrázek 8: Vizualizace dat pomocí funkce `plot.geodata(x)`.

```
> x=as.geodata(x)
```

Takto jsme získali data, se kterými už můžeme pracovat. Předešlými kroky nám vznikla datová matice se třemi sloupci, z nichž první dva reprezentují souřadnice XCOO a YCOO, na kterých bylo měření provedeno a třetí sloupec ELEV hodnotu nadmořské výšky, která byla na tomto místě zaznamenána. Poznamenejme, že kromě nadmořské výšky je možné zkoumat také koncentrace chemických prvků, kterých bylo v průběhu projetku naměřeno více než 50 (např. zlato, stříbro, měd, vápník, atd.). Souřadnice jsou zadány v desetinných stupních, chceme-li je převést na klasické stupně, minuty a sekundy, můžeme použít [12].

Volbou funkce `summary(x)` zjistíme, že datový soubor obsahuje 617 bodů, tj. bylo provedeno 617 měření. Nejmenší vzdálenost mezi body je cca 0.5 km a největší vzdálenost činí cca 628 km.

```
> summary(x)
```

```
$n
[1] 617
$coords.summary
XCOO YCOO
min 372602 7370000
max 861309 7890000
$distances.summary
min max
529.0 627686.4
$data.summary
Min. 1st Qu. Median Mean 3rd Qu. Max.
15.0 140.0 200.0 204.9 260.0 540.0
attr(,"class")
[1] "summary.geodata"
```

Daná data zobrazíme pomocí další funkce `plot.geodata(x)`. Na obrázku 8 můžeme vidět čtyři grafy. Graf vlevo nahoře zobrazuje polohu bodů v rovině s barevným rozlišením hodnot nadmořských výšek rozdělených do několika tříd, graf vpravo dole zobrazuje histogram souboru dat a zbylé dva grafy ukazují hodnoty atributu (nadmořských výšek) v souřadnicích XCOO a YCOO.

Přejdeme k výpočtu variogramu, k tomu nám pomůže funkce `variog(x)`. Tato funkce nejdříve vytvoří tzv. mračno bodů, poté body v mračnu uspořádá do tříd podle vzdáleností v prostoru a vypočítá v nich příslušné průměry. Na závěr graficky vykreslí závislost hodnoty variogramu na vzdálenosti mezi body. Uvedeme si některé hodnoty, které nám variogram poskytuje:

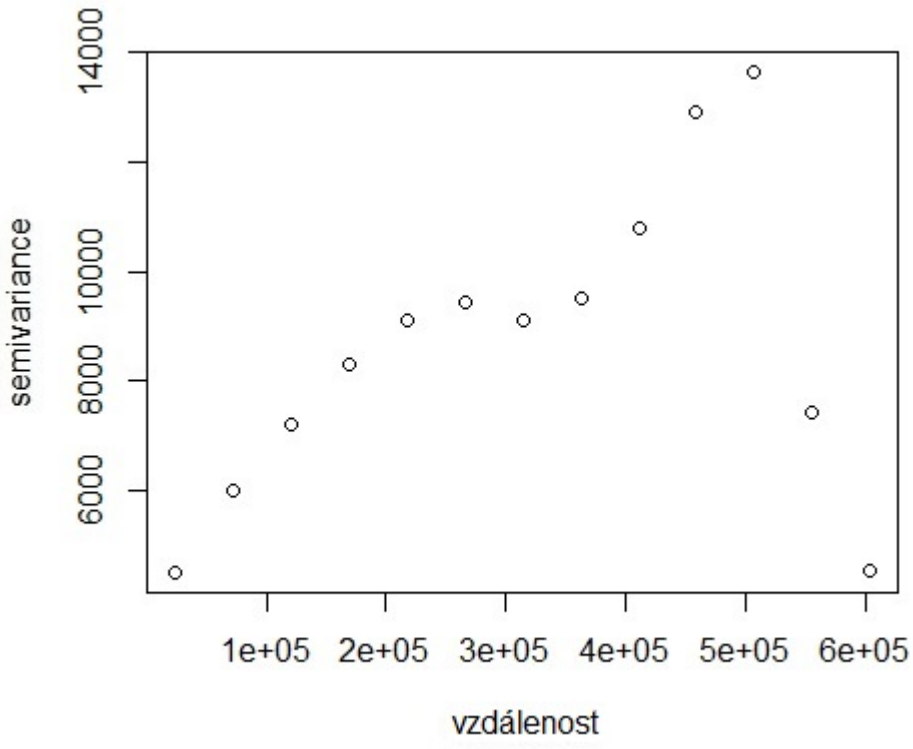
`u`... vektor vzdáleností

`v`... vektor odhadnutých hodnot variogramu vzdáleností uvedených v u

`n`... počet párů v každé třídě, do tříd jsou data uspořádána na základě jejich vzdálenosti a směru

`sd`... směrodatná odchylka hodnot v každé třídě

`var.mark`... rozptyl dat



Obrázek 9: Variogram.

`max.dist`... maximální vzdálenost mezi páry zvolenými ve výpočtu variogramu

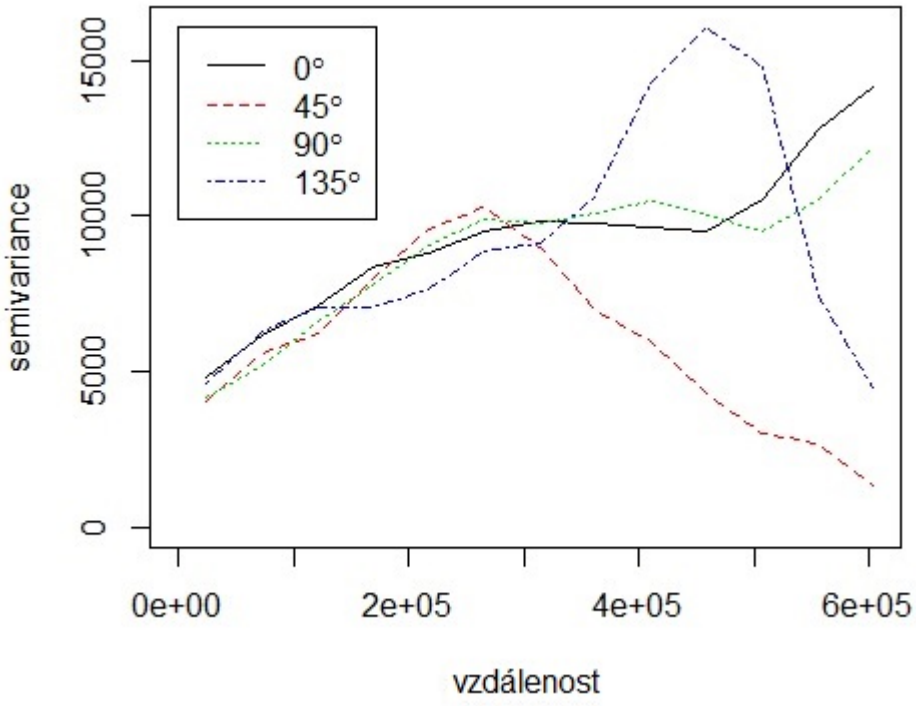
Díky provedenému výpočtu si můžeme variogram vykreslit použitím funkce:

```
> plot(variog(x)$u, variog(x)$v)
```

Obrázek 9 zobrazuje pro každou vzdálenost z vektoru vzdáleností `u` její odhadnutou hodnotu variogramu z vektoru `v`. Odtud vidíme, že se variogram zpočátku chová dle teoretických očekávání, po dosažení očekávaného prahu (pro dosah variogramu cca $3 \cdot 10^5$m = 300km), ovšem dojde k výraznému zašumění následujících hodnot.

Variogram můžeme vypočítat také zavoláním funkce `variog4`. Tato funkce vypočítá směrový variogram pro 4 směry. V jednotlivých směrech (0°, 45°, 90°, 135°) jsou obsaženy dvojice pozorování, jejichž spojnice se od daného směru odchyluje maximálně o 22.5°. Funkci zobrazíme pomocí příkazu `plot.variog4`.

Ukazuje se, že jednotlivé směry jsou velmi rozdílné, z čehož vyplývá, že směr

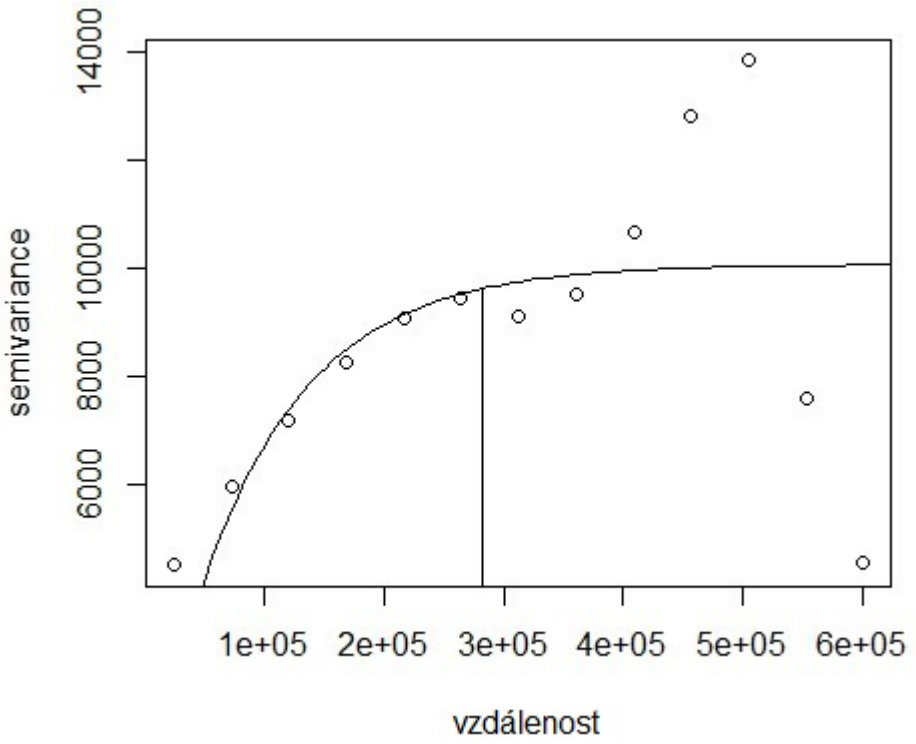


Obrázek 10: Směrový variogram.

má na nadmořskou výšku velký vliv.

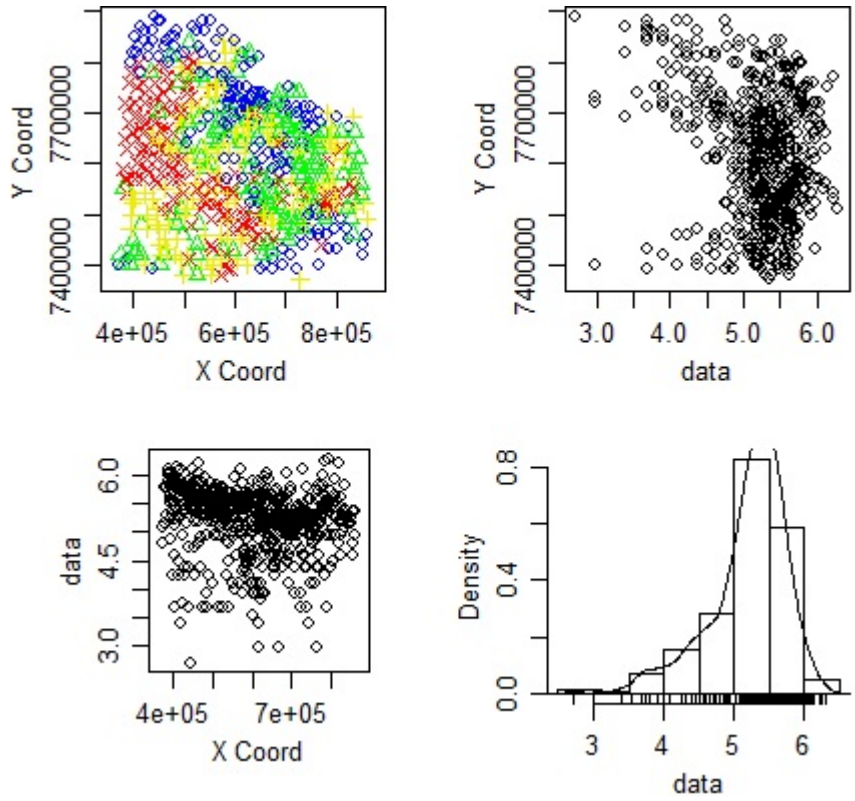
Hodnoty parametrů modelu odhadujeme pomocí metody nejmenších čtverců, v prostředí R nám odhad parametrů umožňuje funkce `variofit`. Experimentální variogram pro nadmořskou výšku proložíme sférickým variogramem. Dosah teoretického variogramu je 289 km, polovina částečného prahu je ve vzdálenosti 6.8 km, což ukazuje silnou autokorelaci až po tuto vzdálenost.

Podle předpokladu v druhé kapitole data zlogaritmujeme, výběrové hodnoty pak budou označeny symbolem $\zeta(\mathbf{x}_n)$. Výše uvedené výpočty a zobrazení provedeme také pro logaritmovaná data. Obrázek 12 znázorňuje rozložení logaritmovaných dat v rovině, hodnoty nadmořských výšek v souřadnicích XCOO a YCOO a také histogram souboru dat. Obrázky 13 a 14 zobrazují postupně variogram a směrový variogram. Můžeme vidět, že výstupy pro variogram se pro zlogaritmované hodnoty od původních hodnot pro menší vzdálenosti příliš neliší, rozdílnosti jsou zaznamenány, až pro větší vzdálenosti. Stejně jako v předchozím případě variogram proložíme sférickou funkcí, viz obrázek 15.

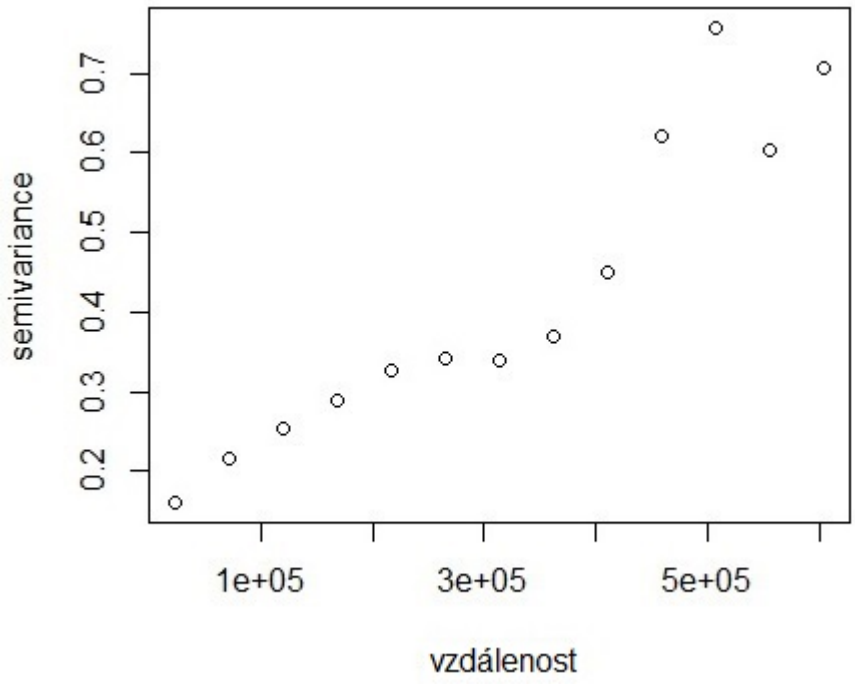


Obrázek 11: Experimentální variogram proložený sférickým variogramem.

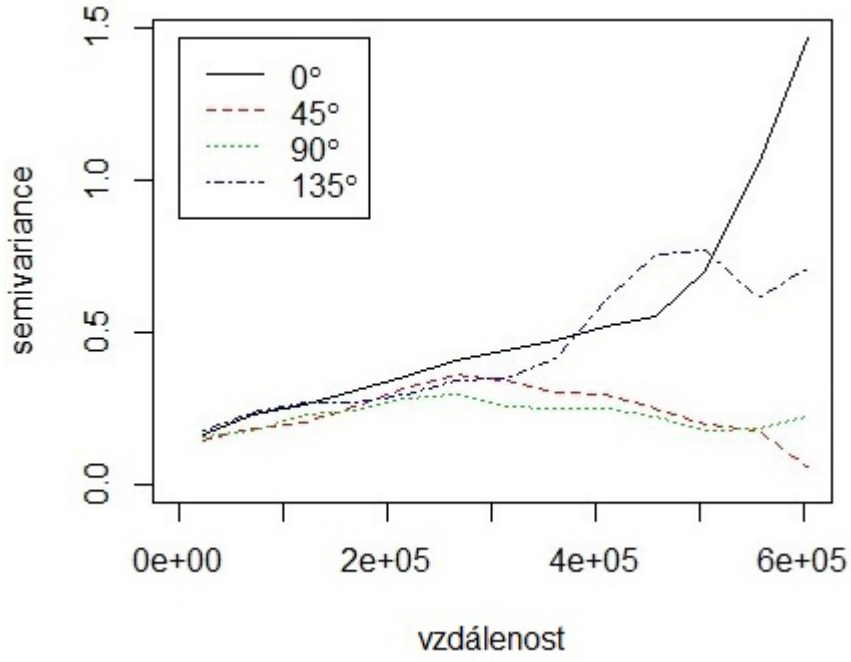
Nyní budeme odhadovat nadmořské výšky v místě, kde měření nebylo provedeno. K tomuto odhadu slouží lineární kombinace hodnot sousedních bodů nadmořských výšek prostřednictvím vah přidělených těmto bodům. Budeme tedy vytvářet kriging na pravidelné mřížce (gridu) vytvořené pomocí nerovnoměrně rozmístěných dat, která byla naměřena. Zvolením funkce `krige.conv` a `image(krig)` vypočítáme a následně zobrazíme konečný kriging i s izoliniemi. Čím bude barva světlejší, tím je nadmořská výška na zkoumané oblasti vyšší [7]. Na obrázku 16 jsme kriging aplikovali na původní data. Použitím krigingu na zlogaritmovaná data získáme výstupy na obrázku 17. Tyto výsledky následně odlogarimujeme a dostaneme obrázek 18. Porovnáním získaných výstupů zjistíme, že interpolace použitá na logaritmované hodnoty dává přesnější výsledky, protože lépe zachycuje strukturu sledované oblasti. Pro zajímavost můžeme získané výstupy porovnat s obrázkem 19, na kterém je zachycena struktura povrchu námi sledované oblasti.



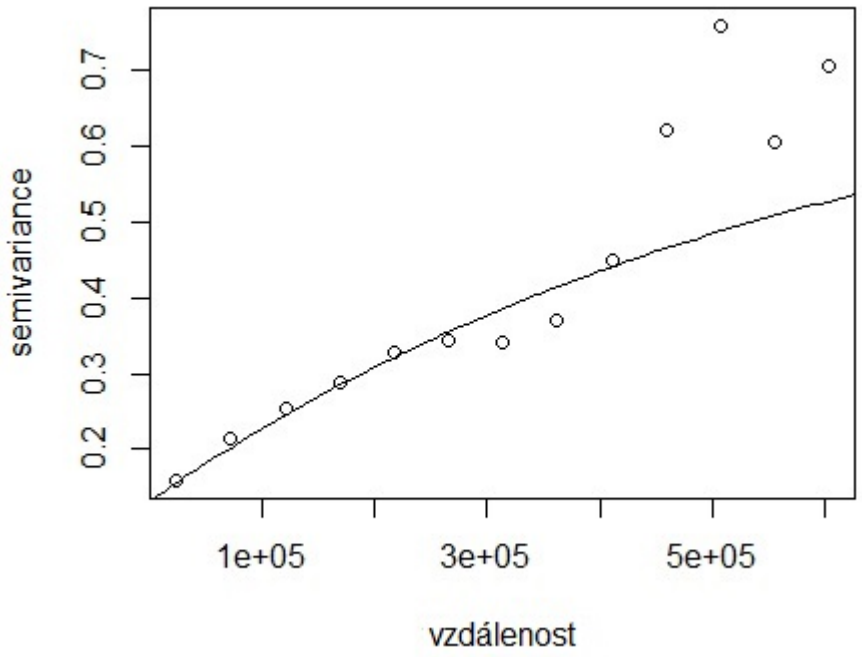
Obrázek 12: Vizualizace zlogaritmovaných dat pomocí funkce `plot.geodata(x)`.



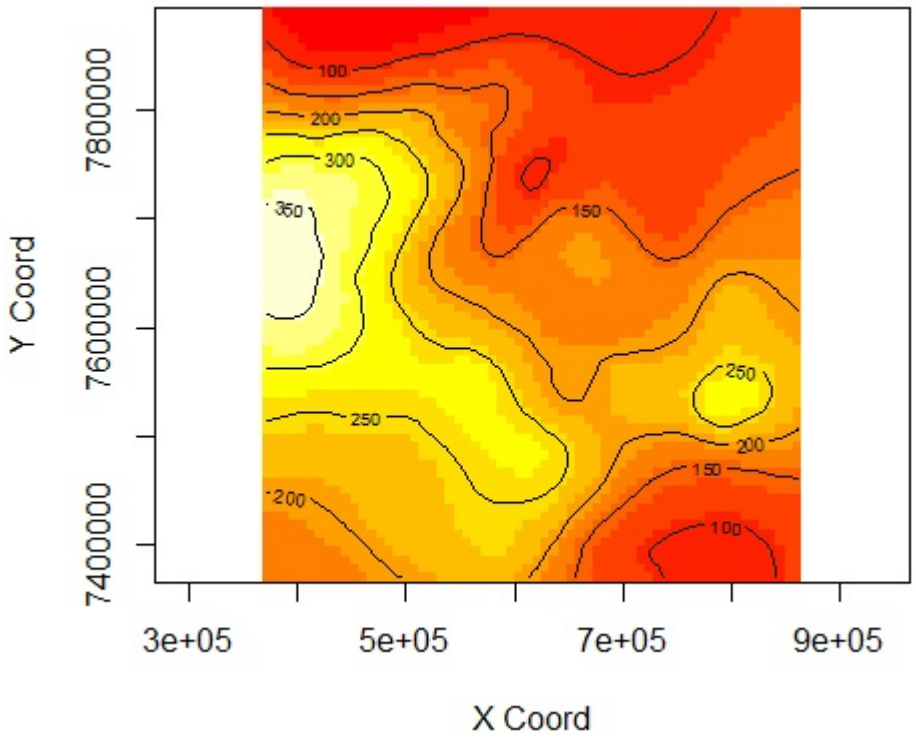
Obrázek 13: Variogram pro zlogaritmovaná data.



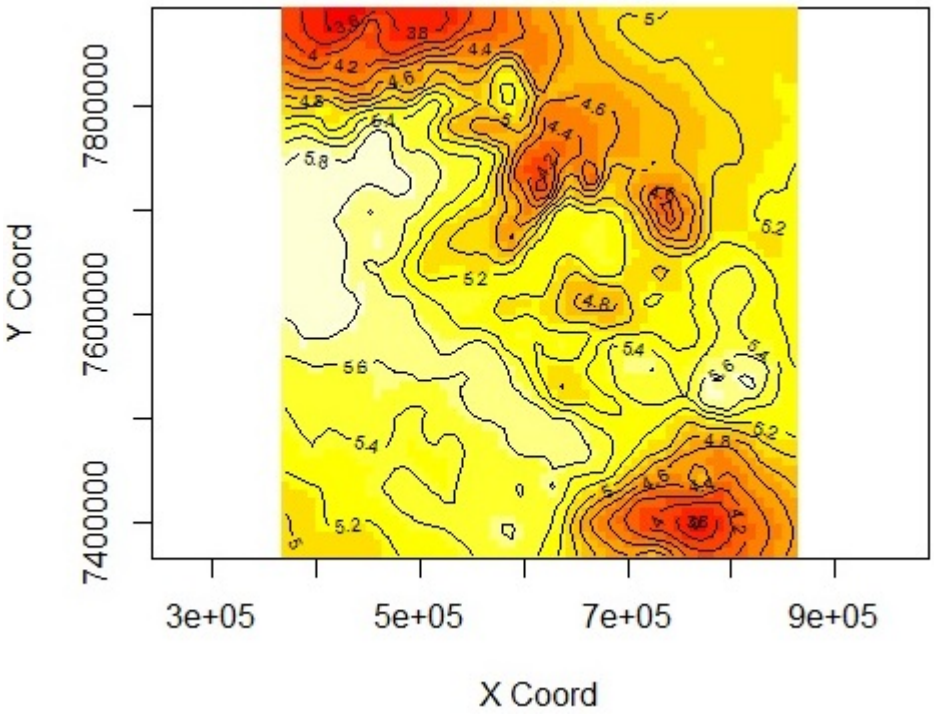
Obrázek 14: Směrový variogram pro zlogaritmovaná data.



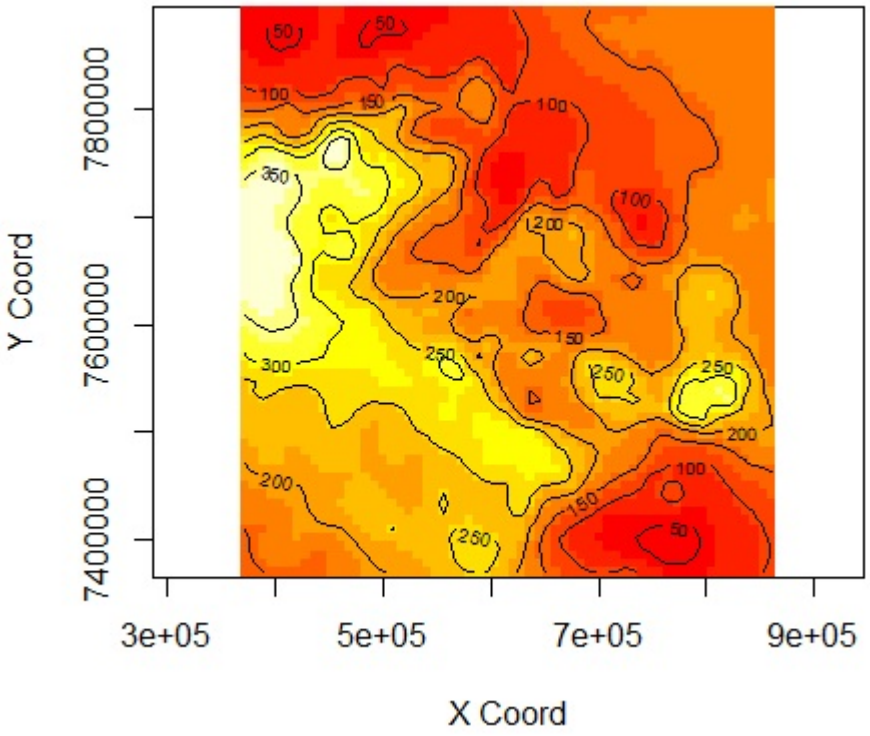
Obrázek 15: Experimentální variogram proložený sférickým variogramem.



Obrázek 16: Krigování původních dat s izoliniemi.



Obrázek 17: Krigování na zlogaritmovaných datech s izoliniemi.



Obrázek 18: Krigingování na odlogaritmovaných datech s izoliniemi.

Obrázek 19: Snímek terénu sledované oblasti.

Závěr

V této práci jsem se zabývala geostatistikou pro data s odlišnou geometrickou strukturou než jsou standardní geostatistická pozorování. Přestože byl původní záměr věnovat tuto práci přímo geostatistice pro kompoziční data, nakonec jsme se vzhledem k obtížnosti problematiky s mým školitelem rozhodli, že nejprve detailně prozkoumáme případ kladných dat, kde je ovšem princip práce s kompozičními daty též výrazně využit.

V teoretické části mé práce jsem za pomoci uvedené literatury popsala základní pojmy týkající se geostatistiky. Stěžejní se stala interpretace pojmu variogram a kriging, které jsem v praktické části využila ke zkoumání reálných dat. Druhá kapitola byla věnována popisu geostatistiky pro kladná data, kde se ukázalo, že nejvhodnější způsob předzpracování dat je jejich zlogaritmování.

V praktické části jsem vyšetřovala reálná data, naměřená v rámci mezinárodního projektu geochemického mapování. Zde jsem ověřila, že v souladu s provedenými teoretickými úvahami dávají přesnější výsledky (obzvláště při krigování) postupy založené na předchozím zlogaritmování dat.

Největším oříškem při tvorbě bakalářské práce pro mě byla její teoretická část, jelikož jsem se předtím s geostatistikou nikdy nesetkala. Ale po těžkém začátku mě tento obor velice zaujal a ráda bych se mu i v budoucnosti dále věnovala. Velkým přínosem pro mě byla i práce se sytémem TEX, ve kterém je práce zpracována.

Literatura

[1] Dzurov, M., Kitzbergerová, K., Šindelářová, L.: Geostatistika v R projektu, ČVUT, Praha, 2010.

[2] Geostatistický slovníček pro oblast modelování environmentálních dat [online], dostupné z: http://www.bilekarpaty.cz/csop/stahnout/geostat-slovnicek.pdf [citováno 1. 12. 2014].

[3] Geostatistika [online], dostupné z https://cs.wikipedia.org/wiki/Geostatistika [citováno 25. 2. 2014].

[4] Glosová, P.: Zjištění prostorové struktury typologických snímků lesní vegetace. Bakalářská práce. Masarykova univerzita, Brno, 2013.

[5] Kola [online], dostupné z: http://cs.wikipedia.cz/wiki/Kola_(poloostrov) [citováno 1. 2. 2014].

[6] Kola Ecogeochemistry, dostupné z: http://www.ngu.no/kola/ [citováno 8. 12. 2013].

[7] Nétek R.: Geostatistika - Seminární práce 2, Univerzita Palackého, Olomouc, 2006.

[8] Pawlowsky-Glahn, V., Tolosana-Delgado, R.: Kriging regionalized positive variables revisited: Sample space and scale considerations. Mathematical Geology, 39, 529-558.

[9] The R Project for Statistical Computing [online], dostupné z: http://www.rproject.org/ [citováno 1. 2. 2014].

[10] Tolosana-Delgado, R., van den Boogaart, K. G. and Pawlowsky-Glahn, V.: Geostatistics for compositions, In: Pawlowsky-Glahn, V., Buccianty, A. : Compositional data analysis: Theory and applications, John Wiley & Sons, Chichester, 2011, 73 - 86.

[11] Wackernagel, H.: Multivariate geostatistics: An introduction with applications, Heidelberg : Springer-Verlag, Berlin, 2003.

[12] Zeměpisné souřadnice: převod a formátování zeměpisných souřadnic [online], dostupné z: https://support.google.com/maps/answer/18539?hl=cs [citováno 30. 1. 2014].