

Geostatistika v R-projektu

Martin Dzurov*, Kristýna Kitzbergerová*, Lucie Šindelářová*

Abstrakt

Geostatistika se zabývá odhady a předpovědí spojitých jevů v prostoru za použití dat jen z omezeného počtu míst v tomto prostoru. Geostatistika aplikuje obecné statistické principy na modelování a vyvozování závěrů o geostatistických problémech. R-projekt je volně dostupný open source software pro statistické výpočty a grafiku, který pracuje pod operačními systémy UNIX, Windows a MacOS. Obsahuje celou řadu balíčků včetně balíčků pro geostatistiku. Cílem projektu je vytvoření interaktivního výukového textu, který bude obsahovat úvod do jazyka R a popis jednotlivých funkcí v geostatistických balíčcích, zejména v balíčku geoR. Celý text bude doplněn o příklady zpracování reálných dat.

Úvod

Mnoho informací, se kterými se denně setkáváme a které používáme, má prostorový charakter. Polohově lokalizovaná data, neboli geodata, nesou tematické informace vázané k určitému místu nebo času, jsou určena geometrickým tvarem a polohou na zemském povrchu, popř. časovým údajem.

Vzhledem k tomu, že se prostorová data bezprostředně týkají našeho každodenního života, byly vyvinuty metody, pomocí nichž je možné taková data zkoumat, analyzovat, odhadovat jejich vývoj, řešit problémy, atd.

Analýzou prostorových dat je možné popsat prostorovou variabilitu atributových hodnot v rámci řešeného území, je možné zjistit zajímavé vlastnosti dat, detekovat shluky, vysoké nebo nízké hodnoty a v neposlední řadě také vypočítat odhady atributových hodnot v místech, které neobsahují naměřené hodnoty. Toto vše nám poskytuje obor geostatistika.

Cílem této práce je poskytnout interaktivní učební text zabývající se geostatistikou v prostředí R a to zejména interpolační metodou kriging. k tomuto účelu byl použit balíček geoR, který poskytuje nástroje pro geostatistickou analýzu dat. Data, která jsou v práci použita představují nadmořské výšky na území České Republiky v měřítku 1 : 500 000.

Tato práce byla podpořena grantem Studentské grantové soutěže ČVUT č. OHK1-065/10.

1 Geostatistika

Geostatistika je vědní obor, který se v poslední době velmi rychle rozvíjí a poprvé byla použita v polovině padesátých let 20. století v těžebním průmyslu pro zlepšení výpočtu

*Katedra matematiky, FSv, ČVUT v Praze

zásoby železné rudy. První výpočty byly provedeny v Jižní Africe těžebním inženýrem D.G. Kriegem a statistikem H.S. Sichelem.

Původně vyvinutý obor pro výpočet odhadu zásoby železné rudy se však rozšířil do mnohých odvětví zabývajících se prostorovými daty spolu s rozvojem výpočetní techniky. Dnes je tento obor velmi populární především v průmyslu, kde je zapotřebí vyhodnocovat prostorově nebo časově určená data. Při rozboru problematiky geostatistiky je zapotřebí brát v úvahu tři oblasti:

Popis dat – je nutné vědět, zda se jedná o data prostorově nebo časově určená, o data s více proměnnými, zda se v datech vyskytují odlehlé hodnoty, které mohou zamaslovat skutečnou strukturu dat. Dnešní výpočetní technika nám poskytuje grafická prostředí, která umožňují vytvářet dynamické náhledy na data (histogramy, korelační diagramy, variogramy, ...).

Interpretace dat – grafická zobrazení získaná z numerických informací jsou dále interpretována již definovanými modely, které se nejvíce přimykají daným datům. Tento model lze brát jako interpretaci prostorových dat a vztahů mezi nimi.

Odhad – našli-li jsme nejlepší model dat, dalším krokem je odhad hodnot pod různými měřítky, v různých lokacích, která nejsou součástí měřených dat. Metody vytváření těchto odhadů jsou založeny na metodě nejmenších čtverců.

1.1 Jazyk R

R je software určený pro manipulaci s daty, provádění výpočtů a grafické zobrazení těchto dat. Mimo jiné také:

- je efektivním nástrojem pro zpracování dat
- obsahuje sadu operátorů pro pole, zejména pak pro výpočty s maticemi
- obsahuje velkou kolekci nástrojů pro průběžnou analýzu dat
- je napsán v jednoduchém a efektivním programovacím jazyku, který zahrnuje podmínky, cykly, uživatelem definované rekurzivní funkce a vstupní a výstupní proudy.

R je nástroj pro vývoj nových metod určených k analýze dat. Jeho vývoj rapidně stoupá a je stále rozšiřován pomocí tzv. balíčků. Nicméně většina programů napsaná v R je stále napsána jen pro určitý typ dat a jejich analýzu.

Prostředí R je vhodné pro statistické rozborů dat a obsahuje již naimplementované klasické i moderní techniky. Běžně používané statistické funkce jsou obsažené v základní sadě R, přejeme-li si použít funkci, která není součástí základního balíčku, musíme použít speciální balíček, který danou funkci obsahuje. Existuje okolo 25 balíčků dodávaných k R a lze je rozdělit do dvou hlavních kategorií – *standardní* a *doporučené*. Řada těchto balíčků je dostupná skrze tzv. CRAN rodinu¹.

¹<http://CRAN.R-project.org>

1.2 Balíček GeoR

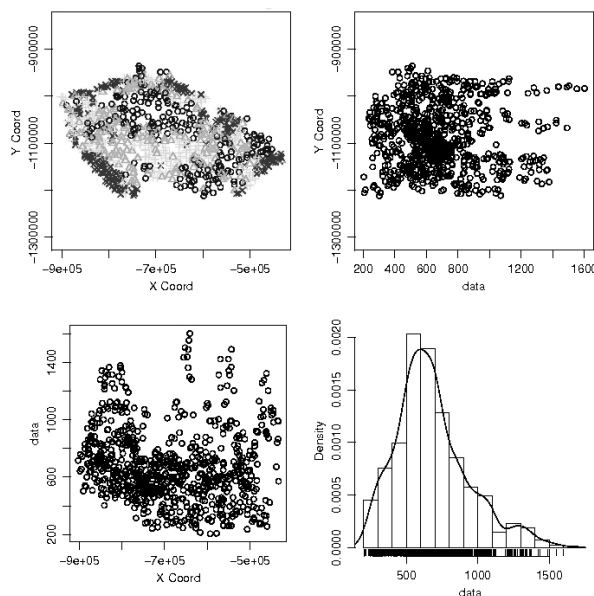
Balíček `geoR` poskytuje funkce pro geostatistickou analýzu dat v prostředí R. Po spuštění softwaru R je nutné balíček nainstalovat.

Data, se kterými chceme pracovat, musí být uložena v objektu typu `geodata`. Objekt tohoto typu obsahuje alespoň jednu souřadnici místa pozorování a hodnotu atributu v tomto místě. Nejčastěji se setkáme s daty uloženými v ASCII formátu, která musíme načíst do R.

Po načtení dat nadmořských výšek a zavoláním funkce `summary()` zjistíme, že datový soubor obsahuje 797 bodů, nejmenší x -ová souřadnice je -902215.0 m a největší -1212091.1 m, nejmenší y -ová souřadnice je -433083.5 m. a největší -935732.4 m. Nejmenší vzdálenost mezi body činí cca 1.2 km a největší vzdálenost mezi dvěma body je 488 km.

Pomocí další vestavěné funkce `plot.geodata()` zobrazíme daná data.

Graf na obrázku č. 1 vykreslený funkcí `plot.geodata(vysky)` zobrazuje 2D polohu bodů s barevným rozlišením hodnot atributu (nadmořských výšek) rozdělených do několika tříd, dále graf hodnot atributu v souřadnicích X a Y a poslední z grafů zobrazuje histogram souboru dat.



Obrázek 1: Grafický výstup funkce `plot.geodata()`

1.3 Variogramy

Variogram je základním geostatistickým nástrojem pro vizualizaci, modelování a prostorovou autokorelaci naměřených hodnot. Variance poté udává míru rozptylu, jak se mění hodnoty proměnné při změně vzdálenosti h z bodu x do bodu $x + h$.

Vzorec pro výpočet variogramu je:

$$\gamma(h) = \frac{1}{2n} \sum_{i=1}^{\infty} [z(x_i) - z(x_i + h)]^2 \quad (1)$$

Vypočteme jej jako součet čtverců eukleidovských vzdáleností mezi každou dvojicí bodů, dělený dvojnásobkem počtu bodů.

V balíčku geoR jsou implementovány dva typy odhadů empirických variogramů:

- **Klasický odhad**

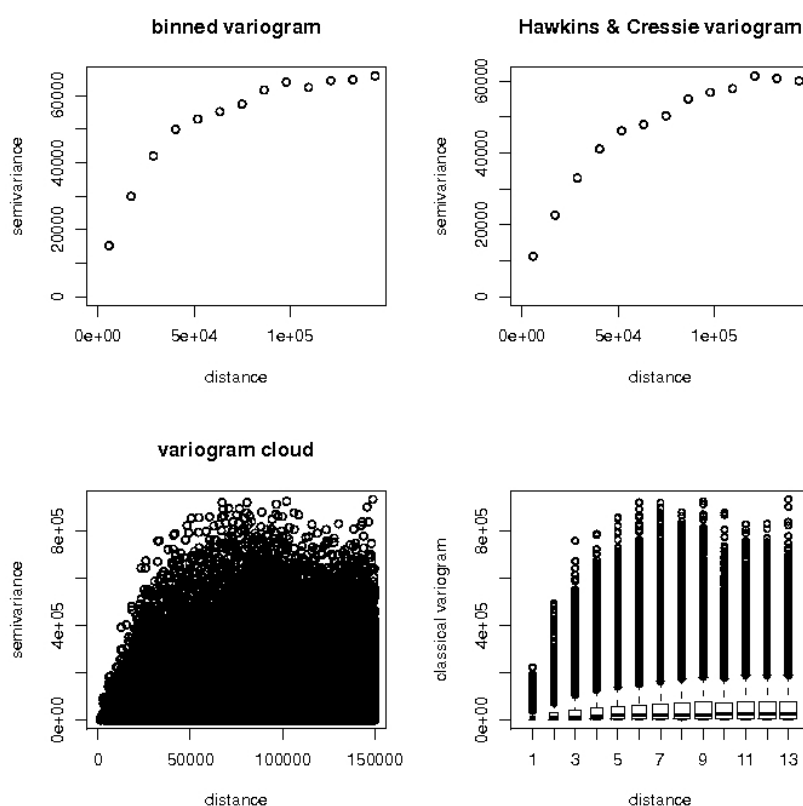
$$\gamma(h) = \frac{1}{2n} \sum_{i=1}^{\infty} [z(x_i) - z(x_i + h)]^2 \quad (2)$$

- **Hawkins-Cressieův robustní odhad**

$$\gamma(h) = \frac{\left[\frac{1}{n} \sum_{i=1}^{\infty} [z(x_i) - z(x_i + h)]^2 \right]^{\frac{1}{4}}}{0.914 + \frac{0.998}{n}} \quad (3)$$

Algoritmus výpočtu variogramu funguje zjednodušeně následovně:

1. nejdříve vytvoří tzv. mračna bodů (variogram cloud)
2. body v mračnu uspořádá do tříd (bins) podle vzdáleností
3. v každé třídě vypočítá průměr
4. v posledním kroku graficky vykreslí závislost hodnoty variogramu γ na vzdálenosti mezi body, viz obrázek č. 2



Obrázek 2: Variogramy

1.3.1 Odhady parametrů variogramu a kovarianční funkce

Experimentální variogramy jsou nahrazeny variogramy teoretickými a to z důvodu, aby měl model variogramu fyzikální význam. Použití teoretického variogramu zajišťuje, že variance jakékoliv lineární kombinace náhodně vybraných hodnot bude vždy kladná. "Napasování" teoretického variogramu na empirický se většinou děje od oka, protože ani empirický variogram přesně neodpovídá sekvenci naměřených hodnot.

Kovarianční funkce je definována:

$$\begin{cases} Ez(x) = m & \text{pro } x \in \mathcal{D} \\ Ez(x) \cdot z(x+h) - m^2 = C(h) & \text{pro } x, x+h \in \mathcal{D} \end{cases} \quad (4)$$

Vlastnosti kovarianční funkce:

1. Kovarianční funkce je omezená a její absolutní hodnota není větší než variance

$$|C(h)| \leq C(0) = \text{var}(z(x)) \quad (5)$$

2. sudá funkce

$$C(-h) = C(h) \quad (6)$$

3. kovarianční funkce podělená variancí se nazývá *korelační funkce*

$$\rho(h) = \frac{C(h)}{C(0)}, \quad -1 \leq \rho(h) \leq 1 \quad (7)$$

4. kovarianční funkce je *pozitivně definitní funkce*, což znamená, že použijeme-li kovarianční funkci $C(h)$ pro výpočet variance lineární kombinace $n + 1$ náhodných proměnných $z(x_i)$, je tato variance vždy kladná.

$$\text{var} \left(\sum_{i=0}^n w_i z(x_i) \right) = \sum_{i=0}^n \sum_{j=0}^n w_i w_j C(x_i - x_j) = \mathbf{w}^T \mathbf{C} \mathbf{w} \quad (8)$$

Vztah mezi kovarianční funkcí a variogramem je poté následující:

$$\gamma(h) = C(0) - C(h) \quad (9)$$

Modely kovariančních funkcí

Pro teoretické variogramy existuje několik modelů, zde uvedeme základní čtyři typy:

1. **Nugget-effekt model** – kovarianční funkce $C(h)$ modeluje nespojitost na počátku – tzv. *nugget* – viz obrázek č. 3.

$$C_{nug}(h) = \begin{cases} b, & h = 0 \\ 0, & h > 0 \end{cases} \quad (10)$$

a platí, že

$$\lim_{h \rightarrow \infty} \gamma(h) = b, \quad b > 0 \quad (11)$$

2. **Exponenciální model** – exponenciální kovarianční funkce klesá exponenciálně s rostoucí vzdáleností bodů a je definována následovně:

$$C_{exp}(h) = \sigma^2 \exp\left(-\frac{3h}{R}\right), \quad \sigma, R > 0 \quad (12)$$

σ je tzv. *sill* (neboli výška variogramu) – viz obrázek č. 3.

R je tzv. *range* a vyjadřuje, jak rychle klesá kovariance. Pro hodnoty vzdálenosti $h = 3R$ klesne kovarianční funkce o 95% své původní hodnoty. Jinými slovy lze říci, že vyjadřuje, v jak velkém okolí se naměřené hodnoty ještě ovlivňují – viz obrázek č. 3.

Exponenciální model je spojitý a na počátku se asymptoticky blíží nule:

$$\lim_{h \rightarrow \infty} C_{exp} = 0 \quad (13)$$

3. **Sférický model** – je definován následovně:

$$C_{sph}(h) = \begin{cases} \sigma^2 \left(1 - \frac{3h}{2R} + \frac{h^3}{2R^3}\right), & 0 \leq h \leq R \\ 0, & h > R \end{cases} \quad (14)$$

4. **Gaussův model** – je definován následovně:

$$C_{gauss}(h) = \sigma^2 \exp\left(-\frac{3h^2}{R^2}\right) \quad (15)$$

1.4 Cross-validace

Cross-validace je jednoduchý způsob, jak porovnat různé předpoklady, např. o modelu (typ variogramu a jeho parametry, velikost nejbližšího okolí použitého pro následnou interpolaci), o datech (detekce odlehlých hodnot nebo anomálií).

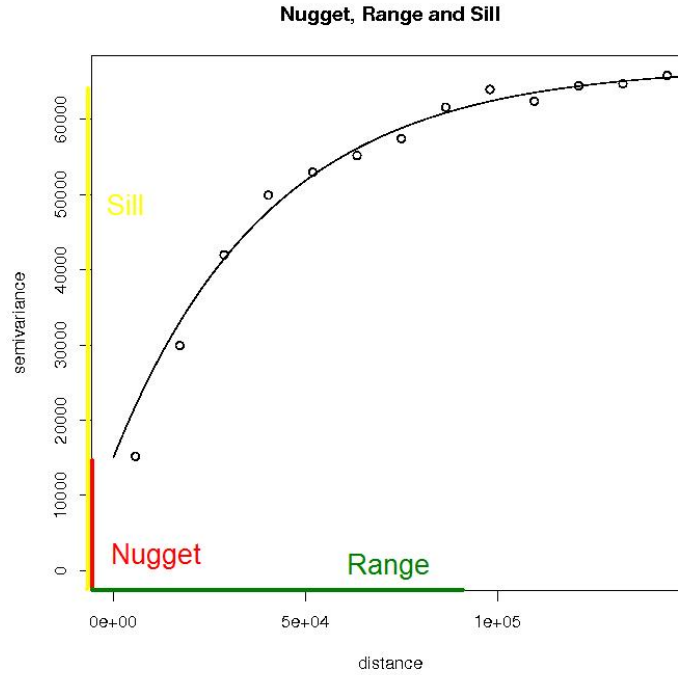
V procesu cross-validace je každá hodnota $z(x_\alpha)$, na které se cross-validace provádí, odstraněna z množiny dat a hodnota v tomto bodě $z^*(x_\alpha)$ je poté odhadnuta pomocí zbývajících $n - 1$ hodnot.

Rozdíl mezi hodnotou naměřenou a hodnotou určenou odhadem poté ukazuje, jak dobře zvolený model odpovídá skutečným naměřeným hodnotám.

$$z(x_\alpha) - z^*(x_\alpha) \quad (16)$$

Pokud je aritmetický průměr chyby cross-validace blízký nule, pak lze tvrdit, že se nejedná o zjevné zkreslení, avšak pokud je tento průměr značně kladný nebo záporný, lze předpokládat, že došlo k systematickému nadhodnocení nebo podhodnocení.

$$\frac{1}{n} \sum_{\alpha=1}^n (z(x_\alpha) - z^*(x_\alpha)) \cong 0 \quad (17)$$



Obrázek 3: Znázornění parametrů variogramu – nugget, sill a range

Pokud se následující rovná jedné, pak lze tvrdit, že se aktuální chyba odhadu rovná v průměru chybě zvoleného modelu. Pokud je tato podmínka splněna, lze zvolený model a jeho parametry považovat za adekvátní.

$$\frac{1}{n} \sum_{\alpha=1}^n \frac{(z(x_{\alpha}) - z^*(x_{\alpha}))^2}{\sigma_{\alpha}^2} \cong 1 \quad (18)$$

σ_{α} je směrodatná odchylka reprezentující chybu odhadu modelu

V prostředí R je funkce pro cross-validaci již naimplementovaná a lze ji spustit příkazem `xvalid()`. Jako argumenty vstupují měřená data, zvolený model, který má být ověřen a případně logická hodnota, zda se parametry modelu mají znovu odhadnout.

2 Interpolační metoda Ordinary Kriging

Metoda Ordinary Kriging (OK) je nejrozšířenější metodou typu kriging. Slouží k odhadu hodnoty na bodě nebo na celém regionu, pro který známe variogram, za použití dat v blízkém okolí tohoto bodu.

Přejeme-li si odhadnout hodnotu v místě x_0 , kde nemáme provedeno měření, získáme tuto hodnotu lineární kombinací hodnot ze sousedních bodů a vah, které jsou těmto bodům přidělené:

$$\hat{z}(x_0) = \sum_{\alpha=1}^n w_{\alpha} z(x_{\alpha}), \quad \sum_{\alpha=1}^n w_{\alpha} = 1 \quad (19)$$

Je však nutné dodržet několik předpokladů:

- *pozitivní semidefinitnost*

- Matice A je pozitivně semidefinitní, pokud její kvadratická forma je pozitivně semidefinitní:

$$x^T A x \geq 0, \quad x \neq 0 \quad (20)$$

- *Best Linear Unbiased Predictor (BLUP)* – nejlepší lineární nezávislý prediktor

- $\hat{h}_0 = \sum_{\alpha=1}^n w_{\alpha} h_{\alpha} + w_0$ je lineární prediktor podle definice
- $\hat{h}_0$ je nezávislý prediktor, pokud platí $E \{ \hat{h}_0 - h_0 \} = 0$
- $\hat{h}_0$ je nejlepší, pokud platí $E \{ \| \hat{h}_0 - h_0 \|^2 \}$ je minimální

Minimalizací tedy získáme systém rovnic pro metodu kriging:

$$\underbrace{\begin{pmatrix} \gamma(x_1 - x_1) & \cdots & \gamma(x_1 - x_n) & 1 \\ \vdots & \cdots & \vdots & \vdots \\ \gamma(x_n - x_1) & \cdots & \gamma(x_n - x_n) & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix}}_{\mathbf{C}_n} \underbrace{\begin{pmatrix} w_1 \\ \vdots \\ w_n \\ \lambda \end{pmatrix}}_{\mathbf{w}_n} = \underbrace{\begin{pmatrix} \gamma(x_1 - x_0) \\ \vdots \\ \gamma(x_n - x_0) \\ 1 \end{pmatrix}}_{\mathbf{d}_n} \quad (21)$$

$\mathbf{C}_n$ vyjadřuje rozdíly mezi naměřenými daty

$\mathbf{w}_n$ obsahuje váhy přiřazené měřením

λ je Langrangův multiplikátor

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

Neznámou v tomto případě je matice vah, kterou snadno vypočítáme a dosadíme do rovnice č. 19.

$$w_n = C_n^{-1} d_n \quad (22)$$

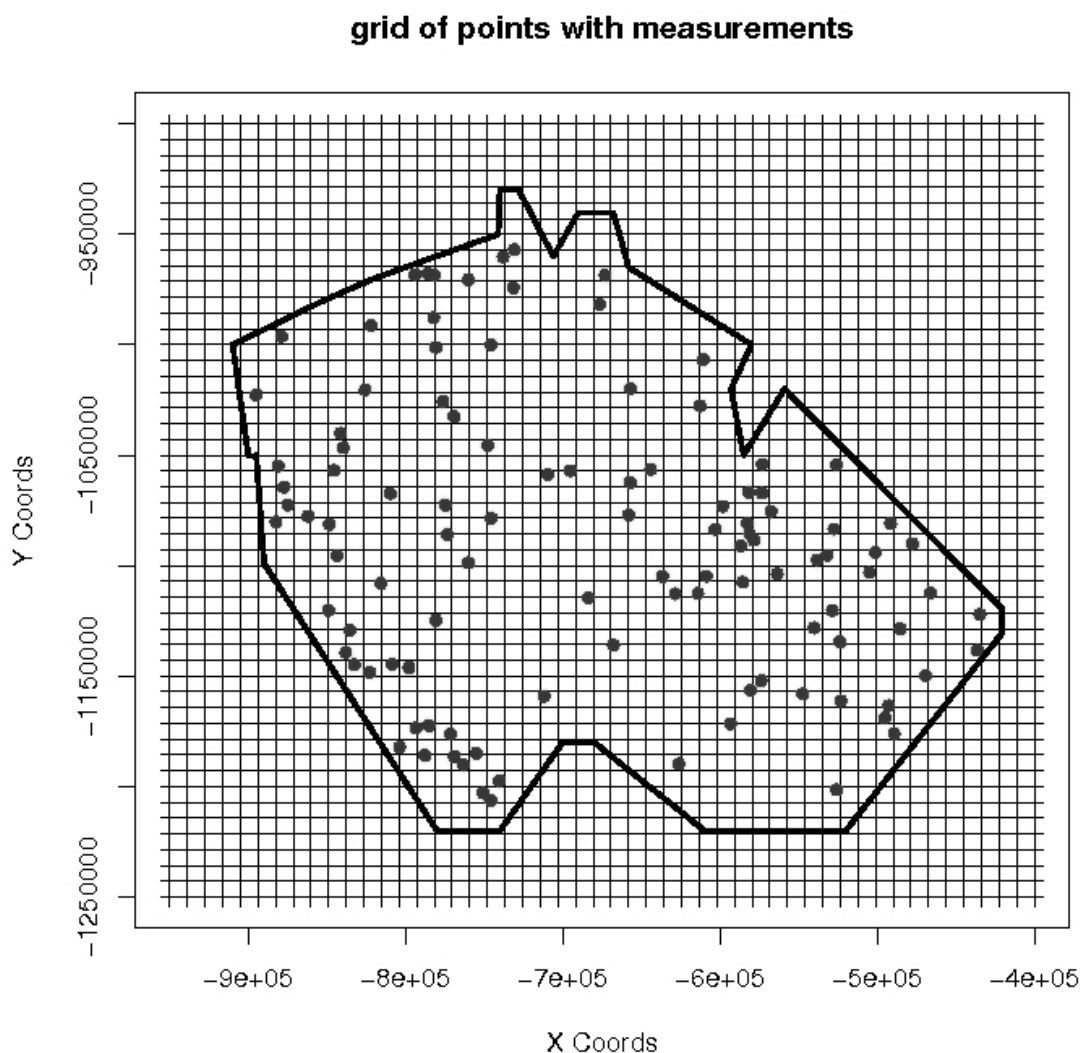
Variance odhadnutých hodnot je pak dána následovně:

$$\sigma^2 = \lambda - \gamma(x_0 - x_0) + \sum_{\alpha=1}^n w_{\alpha} \gamma(x_{\alpha} - x_0) \quad (23)$$

Ordinary Kriging je tzv. přesný interpolátor, to znamená, že pokud hodnota x_0 je identická s místem, kde bylo provedeno měření, bude vyinterpolovaná hodnota identická s hodnotou naměřenou:

$$\hat{z}(x_0) = z(x_{\alpha}), \quad x_0 = x_{\alpha} \quad (24)$$

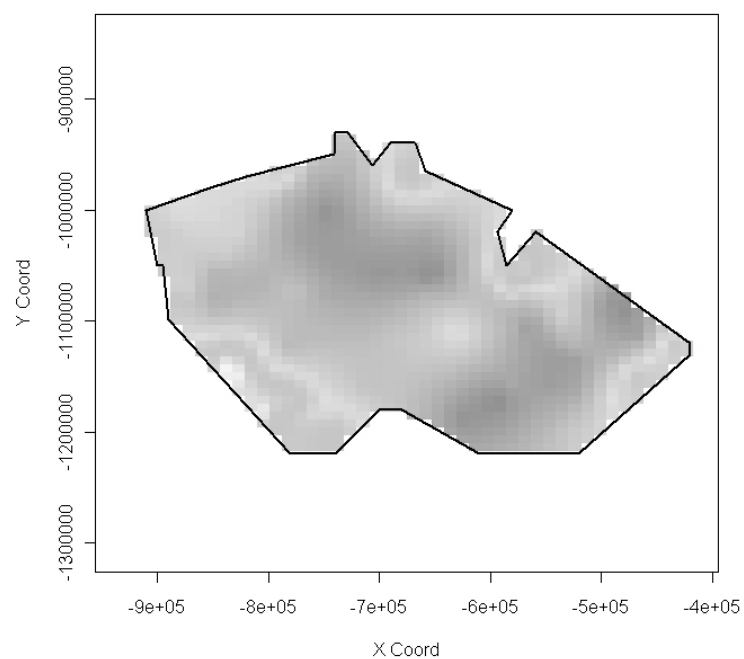
Metodu kriging lze použít jako interpolační metodu pro odhad hodnot na pravidelné mřížce pomocí nepravidelně rozmístěných dat. Tuto metodu jsme využili i pro interpolaci výšek ČR. Obrázek č. 4 zobrazuje pravidelnou mřížku, na které budou odhadnuty hodnot nadmořských výšek, červené body zobrazují 100 náhodně vybraných bodů, na kterých známe hodnotu hledaného atributu. Celý výpočet bude proveden uvnitř polygonu, který lemuje hranice území ČR.



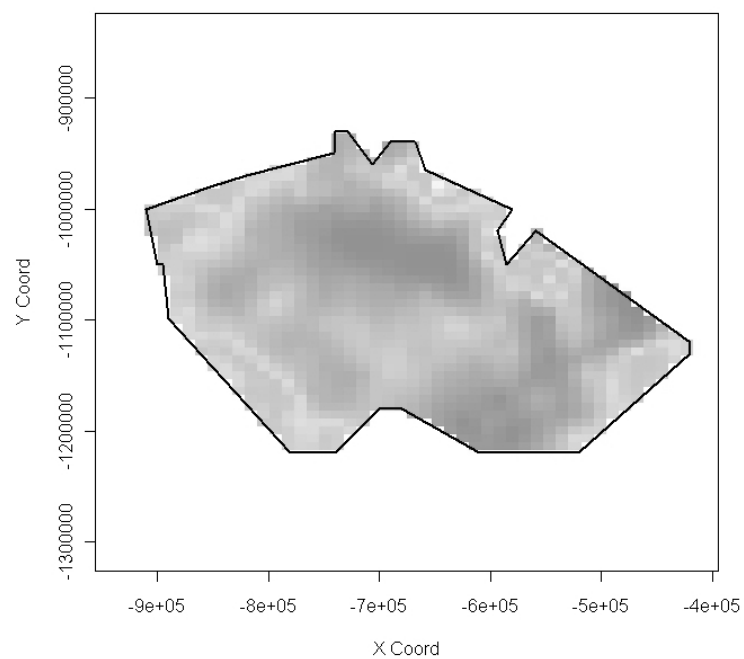
Obrázek 4: Pravidelná mřížka, na které bude proveden výpočet metodou OK

Výsledný graf vyinterpolovaných nadmořských výšek je zobrazen na obrázku č.5, střední chyby interpolace zobrazuje obrázek č. 7. Jako barevná stupnice byla použita stupnice `terrain.colors()`, kde tmavá barva vyjadřuje vysoké hodnoty a světlá naopak hodnoty velmi nízké. Pro vykreslení byla v prostředí R použita funkce `image()`.

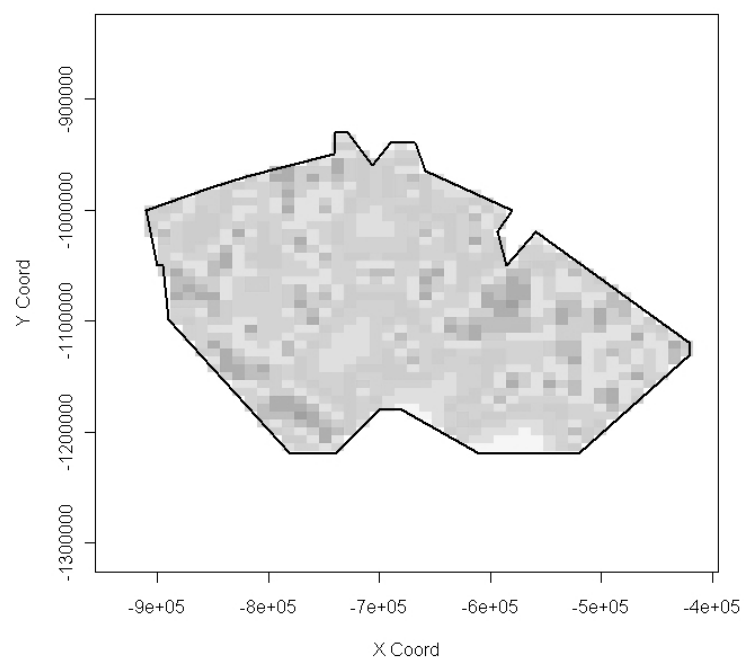
Na obrázku č. 5 a č. 6 lze porovnat výsledek interpolace za použití pouze 100 náhodně vybraných bodů pro interpolaci a za použití celého souboru měření. Stejně porovávání lze provést na grafech č. 7 a č. 8 se středními chybami určení nadmořské výšky.



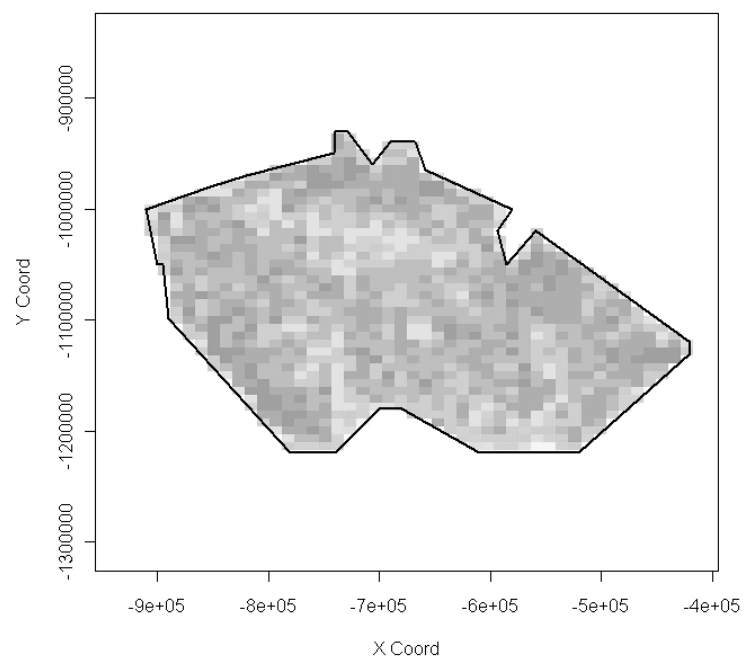
Obrázek 5: Predikce nadmořských výšek pomocí OK ze 100 náhodně vybraných bodů



Obrázek 6: Predikce nadmořské výšky pomocí OK ze všech hodnot měření



Obrázek 7: Směrodatná odchylka odhadnutých nadmořských výšek ze 100 náhodně vybraných bodů



Obrázek 8: Variance odhadnutých nadmořských výšek ze všech hodnot měření