Geostatistik og analyse af spatielle data

Allan Aasbjerg Nielsen

Resumé— **Denne note omhandler geostatistiske mål for spatiel korrelation, nemlig autokovariansfunktionen og semivariogrammet, samt deterministiske og geostatistiske metoder til spatiel interpolation, hhv. afstandsvægtning og *kriging*. En række semivariogrammodeller nævnes, specielt beskrives den sfæriske, den eksponentielle og den gaussiske model. Ligningssystemer til udførelse af simpel og ordinær *kriging* udledes. Andre former for *kriging* nævnes, og der gives referencer til international litteratur, Internet adresser og *state-of-the-art software* på området. Der gives et meget simpelt eksempel til illustration af beregningerne samt et mere realistisk eksempel med højdedata fra et område ved Slagelse. Endelig opremses en række attraktive egenskaber ved *kriging*, og der nævnes en simpel prøvetagningsstrategisk overvejelse baseret på *kriging*variansens afhængighed af afstand og retning til de nærmeste observationer.**

I. Introduktion

OFTE har man brug for at kunne integrere punktinformationer med de vektor- og rasterdata, der i øvrigt er lagret i et geografiske informationssystem (GIS). Dette kan gøres ved at lænke punktinformationen til en geografisk koordinat i databasen. Hvis man har mange punktdata, vil det være et fristende alternativ at generere et interpoleret kort, så man fra sine punktdata beregner rasterdata, som kan indgå i en senere analyse på lige fod med andre rasterdata.

Dette kapitel omhandler geostatistiske metoder til beskrivelse af spatiel eller rumlig korrelation mellem punktmålinger samt deterministiske og geostatistiske metoder til udførelse af den ønskede interpolation.

Den grundlæggende idé i geostatistikken består i at betragte observerede værdier af geokemiske, geofysiske eller andre naturlige variable som realisationer af en stokastisk proces i planen eller rummet. For hver postion $\boldsymbol{r}$ i et domæne $\mathcal{D}$, som er en del af det Euklidiske rum, findes en målbar størrelse $z(\boldsymbol{r})$, som kaldes en *regionaliseret variabel*. $z(\boldsymbol{r})$ er en realisation af en *stokastisk variabel* $Z(\boldsymbol{r})$. Mængden af stokastiske variable $\{Z(\boldsymbol{r}) \mid \boldsymbol{r} \in \mathcal{D}\}$ udgør en *stokastisk funktion*. $Z(\boldsymbol{r})$ har middelværdi eller forventningsværdi (engelsk: *expectation value*) $\mathrm{E}\{Z(\boldsymbol{r})\} = \mu(\boldsymbol{r})$ og autokovariansfunktion $\mathrm{Cov}\{Z(\boldsymbol{r}), Z(\boldsymbol{r}+\boldsymbol{h})\} = C(\boldsymbol{r},\boldsymbol{h})$, hvor $\boldsymbol{h}$ kaldes forskydningsvektoren. Hvis $\mu(\boldsymbol{r})$ er konstant over $\mathcal{D}$, d.v.s. $\mu(\boldsymbol{r}) = \mu$, siges Z at være første ordens stationær. Hvis $C(\boldsymbol{r},\boldsymbol{h})$ også er konstant over $\mathcal{D}$, d.v.s. $C(\boldsymbol{r},\boldsymbol{h}) = C(\boldsymbol{h})$, siges Z at være anden ordens stationær.

Denne statistiske synsmåde er inspireret af arbejde udført af Georges Matheron i 1962-1963. Den er beskrevet i bl.a. [1], [2]. [3] giver en god praktisk og dataanalytisk orienteret introduktion til geostatistik. [4] er et indlæg i en artikelsamling, som beskriver mange forskellige teknikker og deres anvendelser indenfor geo-videnskaberne. [5] omhandler geostatistik og andre relevante emner i forbindelse med analyse af spatielle data. Geostatistiske fremstillinger i GIS-sammenhæng findes i [6], [7]. [8] omhandler multivariat geostatistik, altså studier af flere variables spatielle samvarians. *The International Association for Mathematical Geology (IAMG)* udgiver bl.a. tidsskriftet *Mathematical Geology*. Her publiceres mange resultater vedrørende geostatistisk forskning. *State-of-the-art software* findes i *GSLIB*, [9], og *Variowin*, [10]. Andet letopnåeligt programmel er *GeoEAS* og *Geostatistical Toolbox*. Alle tre nævnte pakker kan findes på http://www-sst.unil.ch/research/variowin/ (eller kan findes via en søgemaskine). Desuden findes en del kommercielt geostatistisk programmel.

Fremstillingen her er præget af [11].

Allan Aasbjerg Nielsen, Informatik og Matematisk Modellering, Danmarks Tekniske Universitet, Bygning 321, 2800 Lyngby, Telefon 45 25 34 25, Telefax 45 88 13 97, E-post: aa@imm.dtu.dk, Internet http://www.imm.dtu.dk/∼aa/.

II. Spatiel korrelation

I dette afsnit omtales metoder til beskrivelse af lighed mellem målinger af naturlige variable i planen eller rummet. Specielt introduceres autokovariansfunktionen og semivariogrammet. Der gives ligeledes en relation mellem disse størrelser.

A. Semivariogrammet

Betragt to skalare størrelser $z(\boldsymbol{r})$ og $z(\boldsymbol{r}+\boldsymbol{h})$ målt i to punkter i planen eller rummet $\boldsymbol{r}$ og $\boldsymbol{r}+\boldsymbol{h}$ adskilt af forskydningsvektoren $\boldsymbol{h}$. Vi betragter z som en realisation af en stokastisk variabel Z. Variabiliteten kan beskrives v.h.a. *autokovariansfunktionen* (idet vi antager eller påtvinger første ordens stationaritet, d.v.s., at middelværdien er steduafhængig)

$$C(\boldsymbol{r},\boldsymbol{h}) \quad = \quad \mathrm{E}\{[Z(\boldsymbol{r}) - \mu][Z(\boldsymbol{r}+\boldsymbol{h}) - \mu]\}.$$

Variogrammet, 2γ, defineres som

$$2\gamma(\boldsymbol{r},\boldsymbol{h}) \quad = \quad \mathrm{E}\{[Z(\boldsymbol{r}) - Z(\boldsymbol{r}+\boldsymbol{h})]^2\},$$

som er et mål for gennemsnitlige, kvadrerede forskelle på måleværdier som funktion af afstand og retning mellem observationer. I det generelle tilfælde vil variogrammet afhænge af stedvektoren $\boldsymbol{r}$ og af forskydningsvektoren $\boldsymbol{h}$. Geostatistikkens *intrinsiske hypotese* siger at *semivariogrammet*, γ, er uafhængigt af stedvektoren og at det udelukkende afhænger af forskydningsvektoren, d.v.s.

$$\gamma(\boldsymbol{r},\boldsymbol{h}) \quad = \quad \gamma(\boldsymbol{h}).$$

Hvis $Z(\boldsymbol{r})$ er anden ordens stationær (d.v.s., at dens autokovariansfunktion er steduafhængig), gælder den intrinsiske hypotese, hvorimod det omvendte ikke nødvendigvis er tilfældet.

Antager eller påtvinger vi anden ordens stationaritet gælder følgende relation mellem autokovariansfunktionen og semivariogrammet

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

Bemærk, at $C(\boldsymbol{0}) = \sigma^2$, den stokastiske variabels varians.

Givet et sæt punktmålinger kan semivariogrammet beregnes v.h.a. følgende estimator, som beregner (det halve af) middelværdien af de kvadrerede differenser mellem alle par af målinger $z(\boldsymbol{r}_k)$ og $z(\boldsymbol{r}_k+\boldsymbol{h})$ adskilt af forskydningsvektoren $\boldsymbol{h}$

$$\hat{\gamma}(\boldsymbol{h}) \quad = \quad \frac{1}{2N(\boldsymbol{h})} \sum_{k=1}^{N(\boldsymbol{h})} [z(\boldsymbol{r}_k) - z(\boldsymbol{r}_k+\boldsymbol{h})]^2.$$

$N(\boldsymbol{h})$ er antallet af punktpar adskilt af $\boldsymbol{h}$. $\hat{\gamma}$ kaldes det *eksperimentelle semivariogram*. Der beregnes ofte gennemsnit af $\hat{\gamma}$ over intervaller $\boldsymbol{h} \pm \Delta\boldsymbol{h}$ for både længde (magnitude) og vinkel (argument) af $\boldsymbol{h}$. Gennemsnit for længden af $\boldsymbol{h}$ ($h \pm \Delta h$) beregnes for at få tilstrækkeligt høje $N(\boldsymbol{h})$ til opnåelse af lav estimationsvarians for semivariogramværdien. Gennemsnit over intervaller af argumentet af $\boldsymbol{h}$ beregnes for at kontrollere for eventuel

anisotropi. Anisotropi betegner det forhold, at autokovariansfunktionen (og semivariogrammet) ikke opfører sig ens for alle retninger af forskydningsvektoren mellem observationer. Denne eventuelle anisotropi kan også konstanteres ved at beregne 2-D semivariogrammer, [3], [12], [11], [10], [9].

B. Semivariogrammodeller

For at kunne definere karakteristiske egenskaber ved semivariogrammet parameteriseres dette v.h.a. forskellige semivariogrammodeller. En meget udbredt model, γ^*, er den *sfæriske model* (vi antager isotropi, altså det forhold, at semivariogrammet kun afhænger af afstanden og ikke af retningen mellem observationerne, og betegner med h længden af $\boldsymbol{h}$)

$$\gamma^*(h) = \begin{cases} 0 & h = 0 \\ C_0 + C_1 \left[\frac{3}{2}\frac{h}{R} - \frac{1}{2}\frac{h^3}{R^3}\right] & 0 < h < R \\ C_0 + C_1 & h \geq R\,, \end{cases}$$

hvor C_0 er den såkaldte *nugget* effekt og R kaldes *range of influence* eller blot *range*; $C_0/(C_0 + C_1)$ er den relative *nugget* effekt og $C_0 + C_1$ kaldes *the sill* ($= \sigma^2$) (i mangel af udbredte danske termer). Parametrene C_0 og C_1 må ikke forveksles med autokovariansfunktionen $C(\boldsymbol{h})$. *Nugget* effekt er en diskontinuitet i semivariogrammet for $h = 0$, som skyldes både måleunøjagtigheder og mikrovariabilitet, der ikke kan studeres på den anvendte skala. *Range of influence* er den afstand ved hvilken kovariation mellem målinger ophører; målinger taget længere fra hinanden er ukorrelerede.

To andre hyppigt anvendte modeller er den eksponentielle (se figur 4)

$$\gamma^*(h) = \begin{cases} 0 & h = 0 \\ C_0 + C_1 \left[1 - \exp\left(-\frac{3h}{R}\right)\right] & 0 < h \end{cases}$$

og den gaussiske (se figur 5)

$$\gamma^*(h) = \begin{cases} 0 & h = 0 \\ C_0 + C_1 \left[1 - \exp\left(-\frac{3h^2}{R^2}\right)\right] & 0 < h\,. \end{cases}$$

For de to sidstnævnte modeller gælder, at de aldrig når men nærmer sig *sill* asymptotisk. Den gaussiske model er på grund af sin vandrette tangent for $h \to 0$ god til beskrivelse af meget kontinuerte fænomener.

Andre semivariogrammodeller som lineære og potensfunktioner anvendes også. I øvrigt kan kombinationer af modeller (til modellering af *nested structures*, det forhold at semivariogrammet har forskellig struktur afhængig af længden og eventuelt retningen af forskydningsvektoren mellem observationer) være nyttige.

Modelparametrene kan estimeres v.h.a. iterative, ikke-lineære mindste kvadraters metoder. Disse minimerer den kvadrerede afvigelse mellem det eksperimentelle semivariogram og modellen opfattet som funktion af parameteren $\boldsymbol{\theta}$, her $\boldsymbol{\theta} = [C_0\ C_1\ R]^T$

$$\min_{\boldsymbol{\theta}} ||\hat{\gamma}(\boldsymbol{h}) - \gamma^*(\boldsymbol{\theta}, \boldsymbol{h})||^2.$$

For eksempler på et eksperimentelt semivariogram og forskellige modeller, se figurerne 4 og 5.

Bemærk, at $C(\mathbf{0})$ er er autokovariansfunktionen for forskydningsvektor $\mathbf{0}$, og at C_0 er en parameter i semivariogrammodellen.

C. Eksempler

Figur 1 viser et meget simpelt eksempel med tre observationer til illustration af beregningerne, $z_1 = 1, z_2 = 3$ og $z_3 = 2$

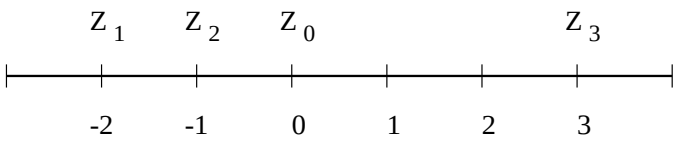

Fig. 1. Simpelt eksempel med tre observationer

med (1-D) koordinaterne –2, –1 og 3. Semivariogrammet $\hat{\gamma}$ med $\Delta h = 1,5$ beregnes således (*lags* er afstandsgrupper defineret ved $h \pm \Delta h$)

lag	h	N	$\hat{\gamma}$
0	$0 < h \leq 3$	1	$1/2(1-3)^2 = 2$
1	$3 < h \leq 6$	2	$1/4((1-2)^2 + (3-2)^2)$ $= 1/2$

Som et andet mere realistisk eksempel ses i figur 2 et kort over prøvetagningssteder. Hver cirkel er anbragt med centrum i et målepunkt, og radius er proportional med den målte størrelse, som er højden over grundvandet i et 10 km × 10 km område ved Slagelse. Figur 3 viser et histogram for disse data.

I figur 4 ses alle kvadrerede differenser som funktion af afstanden mellem observationerne for højdedata fra Slagelse (der er antaget isotropi). Desuden er indtegnet en eksponentiel variogrammodel estimeret direkte på denne punktsky. *Nugget* effekt er 0 m^2, den effektive *range* er 3.840 m og *sill* er 840 m^2 (svarende til 420 m^2 for semivariogrammodellen).

I figur 5 ses det tilsvarende eksperimentelle semivariogram. Δh er her 100 m, og der igen er antaget isotropi. Det ses på det eksperimentelle semivariogram, at en gaussisk model i dette tilfælde nok er bedre end den eksponentielle. Der er derfor indtegnet en gaussisk model estimeret på det eksperimentelle semivariogram. *Nugget* effekt er 18 m^2, *range* er 1.890 m og *sill* er 364 m^2.

III. SPATIEL INTERPOLATION

Denne sektion omhandler deterministiske interpolationsformer som afstandsvægtning og statistiske former kendt under

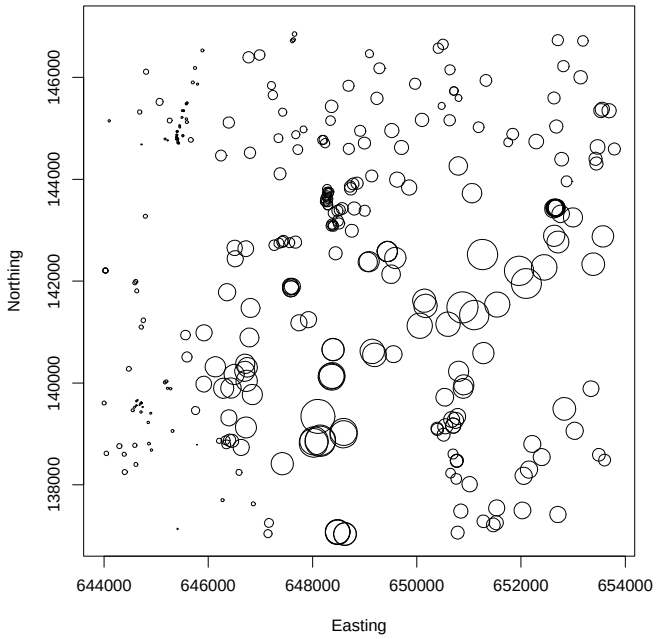

Fig. 2. Prøvetagningssteder, hvor hver cirkel er anbragt med centrum i et målepunkt, radius er proportional med den målte størrelse, som er højden over grundvandet i et 10 km × 10 km område ved Slagelse

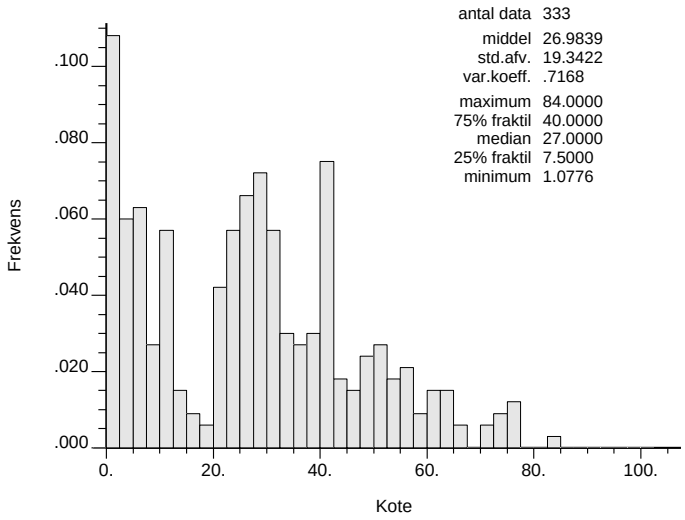

Fig. 3. Simpel statistik og histogram for højdedata ved Slagelse

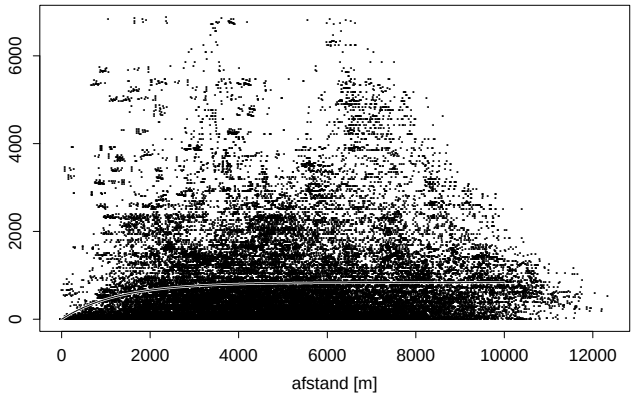

Fig. 4. Alle mulige kvadrerede differenser som funktion af forskydningsvektorens længde; eksponentiel variogrammodel indtegnet

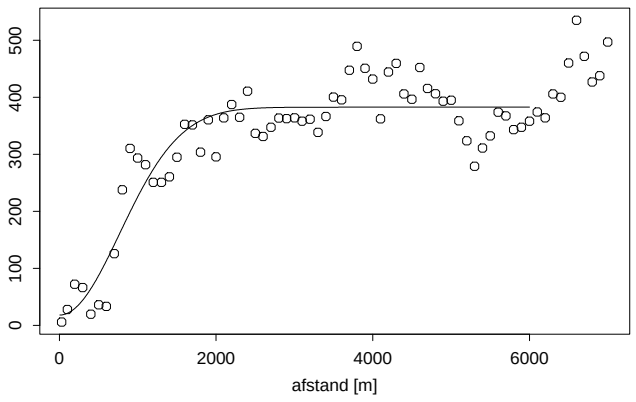

Fig. 5. Eksperimentelt semivariogram som funktion af forskydningsvektorens længde; gaussisk semivariogrammodel indtegnet

fællesnavnet *kriging*. Specielt udledes ligningssystemerne for simpel og ordinær *kriging*.

A. Afstandsvægtning

Noget af det simplest tænkelige til udførelse af interpolation består i at tildele et punkt, hvori værdien ikke kendes, samme værdi som den nærmeste nabo. En forbedring heraf består i at tildele højere vægte til observationer, der ligger tættere på punktet, hvortil det skal interpoleres. En oplagt måde at gøre dette på er at tildele alle N punkter, der indgår i interpolationen, vægte, som er proportionale med den inverse afstand til det ønskede punkt. For det $i$te punkt fås vægten

$$w_i = \frac{1/d_i}{\sum_{j=1}^{N} 1/d_j},$$

hvor d_i er afstanden til det punkt, hvortil der interpoleres. Dette udvides let til at vægte med forskellige potenser, $p > 0$, af den inverse afstand

$$w_i = \frac{1/d_i^p}{\sum_{j=1}^{N} 1/d_j^p}.$$

Andre deterministiske interpolationsmetoder anvender (Delaunay) triangulering, regressionsanalyse til bestemmelse af *trend surfaces*, minimal krumning etc., [13], [3].

A.1 Eksempler

Vi ønsker nu at interpolere til Z_0 med positionen $r = 0$ i figur 1 v.h.a. vægtning med den inverse afstand. d_i er afstanden til Z_0. Vi beregner let følgende vægte

r	d_i	$1/d_i$	$(1/d_i)/\sum(1/d_i)$
-2	2	$1/2$	$3/11$ $(=0,2727)$
-1	1	1	$6/11$ $(=0,5455)$
3	3	$1/3$	$2/11$ $(=0,1818)$

For forskellige potenser af d_i fås vægtene

r	d_i	$p=0,1$	$p=2,0$	$p=10,0$
-2	2	0,3298	0,1837	0,0010
-1	1	0,3535	0,7347	0,9990
3	3	0,3167	0,0816	0,0000

Vi ser, at for lave værdier af p nærmer vi os en ens vægtning af de indgående punkter. For høje værdier af p nærmer vi os en vægtning med 1 af den nærmeste nabo.

B. Kriging

Kriging (efter den sydafrikanske mineingeniør og professor Danie Krige) er en betegnelse for en familie af metoder for minimum fejlvarians estimation. Betragt et lineært (eller rettere affint) estimat $\hat{z}_0 = \hat{z}(\boldsymbol{r}_0)$ på stedet $\boldsymbol{r}_0$ baseret på N målinger $\boldsymbol{z} = [z(\boldsymbol{r}_1), \ldots, z(\boldsymbol{r}_N)]^T = [z_1, \ldots, z_N]^T$

$$\hat{z}_0 \quad = \quad w_0 + \sum_{i=1}^{N} w_i z_i = w_0 + \boldsymbol{w}^T \boldsymbol{z},$$

hvor w_i er de på z_i anvendte vægte og w_0 er en konstant.

Vi betragter z_i som realisationer af stokastiske variable Z_i, $\boldsymbol{Z} = [Z(\boldsymbol{r}_1), \ldots, Z(\boldsymbol{r}_N)]^T = [Z_1, \ldots, Z_N]^T$. Vi tænker på $Z(\boldsymbol{r})$ som bestående af en middelværdi og et residual $Z(\boldsymbol{r}) = \mu(\boldsymbol{r}) + \epsilon(\boldsymbol{r})$ med konstant varians σ^2, $\mathrm{E}\{\epsilon\} = 0$. For den lineære estimator får vi

$$\hat{Z}_0 \quad = \quad w_0 + \boldsymbol{w}^T \boldsymbol{Z}. \tag{1}$$

Estimationsfejlen $z_0 - \hat{z}_0$ kendes ikke, vi kender ikke z_0, som vi jo netop vil estimere. Men for estimationsfejlens forventningsværdi får vi

$$\begin{aligned} \mathrm{E}\{Z_0 - \hat{Z}_0\} &= \mathrm{E}\{Z_0 - w_0 - \boldsymbol{w}^T\boldsymbol{Z}\} \\ &= \mu_0 - w_0 - \boldsymbol{w}^T\boldsymbol{\mu}, \end{aligned} \quad (2)$$

hvor $\mu_0 = \mu(\boldsymbol{r}_0)$ er forventningsværdien af Z_0 og $\boldsymbol{\mu}$ er en vektor af forventningsværdier for $\boldsymbol{Z}$

$$\boldsymbol{\mu} = \begin{bmatrix} \mu(\boldsymbol{r}_1) \\ \vdots \\ \mu(\boldsymbol{r}_N) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_N \end{bmatrix}.$$

Vi ønsker, at vores estimator skal være central, d.v.s. vi kræver at $\mathrm{E}\{Z_0 - \hat{Z}_0\} = 0$ eller

$$\mu_0 - w_0 - \boldsymbol{w}^T\boldsymbol{\mu} = 0. \quad (3)$$

Estimationsfejlens varians er

$$\begin{aligned} \sigma_E^2 &= \mathrm{Var}\{Z_0 - \hat{Z}_0\} \\ &= \mathrm{Var}\{Z_0\} + \mathrm{Var}\{w_0 + \boldsymbol{w}^T\boldsymbol{Z}\} \\ &\quad -2\mathrm{Cov}\{Z_0, w_0 + \boldsymbol{w}^T\boldsymbol{Z}\} \\ &= \sigma^2 + \boldsymbol{w}^T(\boldsymbol{C}\boldsymbol{w} - 2\mathrm{Cov}\{Z_0, \boldsymbol{Z}\}), \end{aligned}$$

hvor $\boldsymbol{C}$ er dispersions- eller kovariansmatricen for de stokastiske variable, $\boldsymbol{Z}$, der indgår i estimationen.

Det i afsnit III-B hidtil beskrevne gælder alle lineære estimatorer. Idéen i *kriging* er nu at finde den lineære estimator, der minimerer estimationsvariansen.

B.1 Simpel *kriging*

I simpel *kriging* (SK) antager vi, at $\mu(\boldsymbol{r})$ er kendt. Fra ligningerne 1 og 3 får vi

$$\hat{Z}_0 - \mu_0 = \boldsymbol{w}^T(\boldsymbol{Z} - \boldsymbol{\mu}).$$

Vægtene w_i findes ved at minimere estimationsvariansen σ_E^2. Dette gøres ved at sætte de partielt afledede til nul

$$\frac{\partial \sigma_E^2}{\partial \boldsymbol{w}} = 2\boldsymbol{C}\boldsymbol{w} - 2\mathrm{Cov}\{Z_0, \boldsymbol{Z}\} = \boldsymbol{0},$$

hvilket giver SK systemet

$$\boldsymbol{C}\boldsymbol{w} = \mathrm{Cov}\{Z_0, \boldsymbol{Z}\}$$

eller

$$\begin{bmatrix} C_{11} & \cdots & C_{1N} \\ \vdots & \ddots & \vdots \\ C_{N1} & \cdots & C_{NN} \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_N \end{bmatrix} = \begin{bmatrix} C_{01} \\ \vdots \\ C_{0N} \end{bmatrix},$$

hvor $C_{ij},\ i,j = 1,\ldots,N$ er kovariansen mellem punkterne i og j blandt de N punkter, der indgår i estimationen af punkt 0. $C_{0j},\ j = 1,\ldots,N$ er kovariansen mellem punkt j og punkt 0, det punkt hvortil der interpoleres. Disse kovarianser findes fra semivariogrammodellen (idet man husker, at $\gamma(\boldsymbol{h}) = C(\boldsymbol{0}) - C(\boldsymbol{h})$) som *sill* minus semivariogramværdien for den aktuelle afstand (og eventuelt retning) mellem observationer. (*Kriging*systemet kan alternativt formuleres v.h.a. semivariogrammet; for at undgå nuller på $\boldsymbol{C}$s diagonal foretrækkes af numeriske årsager autokovariansformuleringen.) C_{ij} her må ikke forveksles med semivariogramparametrene C_0 og C_1.

Den minimale kvadrerede estimationsfejl, som kaldes den simple *kriging*varians, er

$$\begin{aligned} \sigma_{\mathrm{SK}}^2 &= \sigma^2 + \boldsymbol{w}^T(\boldsymbol{C}\boldsymbol{w} - 2\mathrm{Cov}\{Z_0, \boldsymbol{Z}\}) \\ &= \sigma^2 - \boldsymbol{w}^T\mathrm{Cov}\{Z_0, \boldsymbol{Z}\}. \end{aligned}$$

I SK er middelværdien kendt. I praksis antages den konstant for hele domænet (eller studieområdet), eller man må estimere $\mu(\boldsymbol{r})$ forud for estimationen eller konstruere en estimationsalgoritme, som ikke kræver kendskab til middelværdien, se næste afsnit.

B.2 Ordinær *kriging*

I ordinær *kriging* (OK) antager vi, at middelværdien er konstant lig μ_0 for Z_0 og de N punkter, der indgår i estimationen af Z_0. Fra ligningerne 2 og 3 får vi

$$\mathrm{E}\{Z_0 - \hat{Z}_0\} = \mu_0(1 - \boldsymbol{w}^T\boldsymbol{1}) - w_0 = 0$$

for alle μ_0. $\boldsymbol{1}$ er en vektor bestående af lutter et-taller. Dette er kun muligt hvis $w_0 = 0$ og $\boldsymbol{w}^T\boldsymbol{1} = 1$.

Vægtene w_i findes ved at minimere σ_E^2 under bibetingelsen $\boldsymbol{w}^T\boldsymbol{1} = 1$. En standardteknik til minimering under en sådan bibetingelse går ud på at indføre en funktion F med en såkaldt Lagrange multiplikator (her -2λ) ganget med bibetingelsen sat til 0 og minimere

$$F = \sigma_E^2 + 2\lambda(\boldsymbol{w}^T\boldsymbol{1} - 1)$$

uden bibetingelser. Igen sættes de partielt afledede til nul

$$\begin{aligned} \frac{\partial F}{\partial \boldsymbol{w}} &= 2\boldsymbol{C}\boldsymbol{w} - 2\mathrm{Cov}\{Z_0, \boldsymbol{Z}\} + 2\lambda\boldsymbol{1} = \boldsymbol{0} \\ \frac{\partial F}{\partial \lambda} &= 2(\boldsymbol{w}^T\boldsymbol{1} - 1) = 0, \end{aligned}$$

hvilket giver OK systemet

$$\begin{aligned} \boldsymbol{C}\boldsymbol{w} + \lambda\boldsymbol{1} &= \mathrm{Cov}\{Z_0, \boldsymbol{Z}\} \\ \boldsymbol{1}^T\boldsymbol{w} &= 1 \end{aligned}$$

eller

$$\begin{bmatrix} C_{11} & \cdots & C_{1N} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ C_{N1} & \cdots & C_{NN} & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_N \\ \lambda \end{bmatrix} = \begin{bmatrix} C_{01} \\ \vdots \\ C_{0N} \\ 1 \end{bmatrix}.$$

De ønskede værdier for C_{ij} findes som angivet i sidste afsnit om SK.

Den minimale kvadrerede estimationsfejl, som kaldes den ordinære *kriging*varians, er

$$\begin{aligned} \sigma_{\mathrm{OK}}^2 &= \sigma^2 + \boldsymbol{w}^T(\boldsymbol{C}\boldsymbol{w} - 2\mathrm{Cov}\{Z_0, \boldsymbol{Z}\}) \\ &= \sigma^2 - \boldsymbol{w}^T\mathrm{Cov}\{Z_0, \boldsymbol{Z}\} - \lambda. \end{aligned}$$

OK indebærer en implicit re-estimation af μ_0 for hver ny punktkonstellation, en attraktiv egenskab, der gør OK velegnet til interpolation i situationer, hvor middelværdien ikke er konstant (altså ved manglende første ordens stationaritet).

B.3 Eksempler

Vi betragter igen data fra figur 1. Vi ønsker nu at interpolere til positionen $r = 0$ v.h.a. ordinær *kriging*. Til beregningerne bruger vi en påstået semivariogrammodel, nemlig den sfæriske model med $C_0 = 0, C_1 = 1$ og $R = 6$. Dette giver, idet

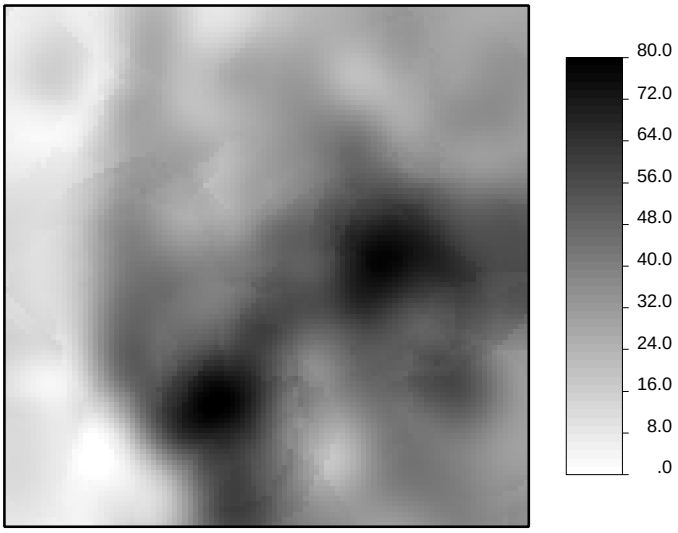

Fig. 6. *Kriged* kort over højder over grundvandet (enheden er m)

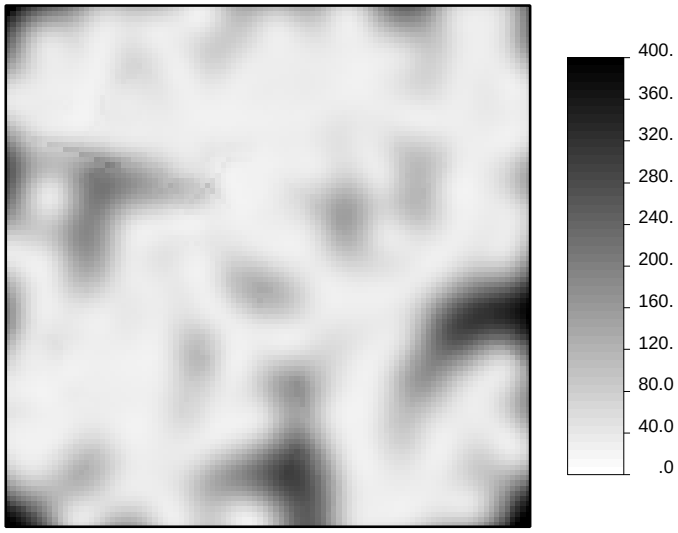

Fig. 7. *Kriging*varians svarende til figur 6 (enheden er m^2)

$C(h) = C(0) - \gamma(h)$, autokovariansfunktionen (som i dette tilfælde, fordi $C_0 + C_1 = 1$, er det samme som autokorrelationsfunktionen)

h	$\hat{\gamma}(h)$	$C(h)$
0	0,0000	1,0000
1	0,2477	0,7523
2	0,4815	0,5185
3	0,6875	0,3125
4	0,8519	0,1481
5	0,9606	0,0394
6	1,0000	0,0000

OK systemet får derfor følgende udseende

$$\begin{bmatrix} 1,0000 & 0,7523 & 0,0394 & 1 \\ 0,7523 & 1,0000 & 0,1481 & 1 \\ 0,0394 & 0,1481 & 1,0000 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \lambda \end{bmatrix} = \begin{bmatrix} 0,5185 \\ 0,7523 \\ 0,3125 \\ 1 \end{bmatrix},$$

hvor talværdierne fås ved opslag i $C(h)$ tabellen. Løsningen er $w_1 = -0,0407, w_2 = 0,7955, w_3 = 0,2452$ og $\lambda = -0,0489$, hvilket giver *kriging*variansen $0,3949$. Vi ser, at selvom Z_1 ligger tættere på Z_0 end Z_3, er vægten på Z_1 meget mindre end vægten på Z_3. Dette er en attraktiv egenskab ved *kriging*, idet der herved tages hensyn til eventuel *clustering* af prøvetagningspunkterne. Man siger, at Z_2 *screen*er for Z_1. Denne *screening* bliver svagere for højere *nugget* effekt for helt at forsvinde for ren *nugget* effekt (altså $C_1 = 0$ og $R = 0$ for de her viste modeller), hvor alle vægtene bliver ens.

Som et mere realistisk eksempel viser figur 6 et *kriged* (OK) kort over højder over grundvandet for det nævnte område ved Slagelse. Interpolationen er baseret på den isotrope gaussiske model for det eksperimentelle semivariogram (*nugget* effekt 18 m^2, *range* 1.890 m og *sill* 364 m^2; søgeradius er 2.000 m og der indgår minimalt 2 og maximalt 20 punkter i estimationen af hvert punkt). Vi ser, at det interpolerede kort stemmer godt overens med kortet over prøvetagningssteder i figur 2. Figur 7 viser de tilsvarende OK varianser. Vi ser, at *kriging*variansen er stor, hvor der er langt til nærmeste prøver.

B.4 Andre *kriging* former

Hvis man ønsker at estimere gennemsnitlige (også kaldet regulariserede) værdier over et volumen snarere end punktværdier, kan man benytte blok*kriging*, som kan kombineres med flere andre *kriging*former.

Hvis flere variable studeres simultant kan de beskrevne metoder til etablering af spatiel korrelation udvides til at håndtere den spatielle samvarians mellem alle par af variable i form af krydssemivariogrammer. Tilsvarende kan flere variable interpoleres simultant i *cokriging*. *Cokriging* har størst praktisk værdi, hvis en variabel er målt færre steder end andre korrelerede variable.

Universal kriging er en metode for det tilfælde, at middelværdien kan skrives som linearkombinationer af kendte funktioner, som ideelt er bestemt af fysikken i det problem, man arbejder med. Ligeledes findes metoder til ikke-lineær *kriging*, som *lognormal kriging*, *multiGaussian kriging*, rang *kriging*, indikator *kriging* og *disjunctive kriging*.

Referencer til ovenstående er [1], [3], [9].

IV. Konklusion

Ovenstående afsnit og eksempler viser følgende fordele ved *kriging*:

- *Kriging* er en interpolationsform, der giver både et estimat baseret på den spatielle struktur af den variabel, man studerer, som udtrykt ved krydskovariansfunktionen (eller semivariogrammet), og en estimationsvarians, som er minimeret.
- *Kriging* estimatoren er den bedste lineære estimator, *best linear unbiased estimator* (BLUE) i betydningen lavest estimationsvarians, og den er også eksakt, d.v.s. at hvis et punkt, hvortil der interpoleres, falder sammen med et prøvetagningspunkt, giver *kriging* samme værdi som den målte, og *kriging*variansen er 0.
- *Kriging*systemet og *kriging*variansen afhænger kun af kovariansfunktionen (eller semivariogrammet) og det spatielle *layout* af prøvetagningspunkterne og ikke af selve de målte værdier. Dette kan anvendes til planlægning af en fornuftig prøvetagningsstrategi.
- Løsningen af *kriging*systemet indebærer en statistisk afstandsvægtning af de datapunkter, der indgår i estimationen. Ligeledes bliver de estimerede vægte re-skaleret, så de summerer til 1. Ydeligere tages der hensyn til redundans i form af mulig *clustering* af prøvetagningspunkter; denne *de-clustering* er

grunden til føromtalte *screening*effekt.

• Den implicitte re-estimation af middelværdien for hver ny punktkonstellation gør, at OK egner sig til situationer, hvor middelværdien varierer over studieområdet, altså hvor der ikke er første ordens stationaritet.

Desuden gælder, at *kriging*systemet har en unik løsning hvis og kun hvis kovariansmatricen $\boldsymbol{C}$ (sektion III-B) er positiv definit; dette sikrer også en ikke-negativ *kriging*varians.

*Kriging*s styrke kan tilskrives en kombination af ovenstående egenskaber.

Kriging fordrer den nævnte intrinsiske hypotese, altså at semivariogrammet må antages konstant over hele studieområdet, hvis man vælger at formulere *kriging*systemet v.h.a. semivariogrammet. Hvis man vælger at formulere *kriging*systemet v.h.a. autokovariansfunktionen, hvilket foretrækkes p.g.a. numeriske forhold, må man antage anden ordens stationaritet, altså konstant autokovariansfunktion over hele studieområdet.

Dette kan virke som en ulempe ved *kriging*, men hvis deterministiske metoder anvendes forudsættes implicit noget lignende. Og det kan næppe betegnes som en ulempe ved geostatistiske metoder, at man tvinges til at overveje betimeligheden af de nævnte forhold.

V. Afslutning

*Kriging*variansens afhængighed af afstanden til nærmeste prøver kan anvendes til at lægge en prøvetagningsstrategi. Hvis semivariogrammet og prøvetagningsstederne kendes, kan man inden selve prøvetagningen udregne *kriging*varianser. Hvis disse varianser i visse dele af studieområdet bliver for store, kan man modificere placeringen af prøverne for at opnå lavere varians. For at opnå et godt estimat af *nugget* effekten, som er en meget vigtig parameter for resultatet af *kriging*, kan det desuden være en fordel at forsøge at lægge en del prøver tæt på hinanden.

I multivariate studier, hvor flere variable studeres simultant, kan man i stedet for at interpolere de oprindelige variable, interpolere kombinationer af disse variable. F.eks. kan man interpolere principale komponenter eller faktorer fra en faktoranalyse eller fra en spatiel faktoranalyse, [14], [15], [11], [16], [17]. Generelle referencer til multivariat statistik er f.eks. [18], [19]. [20] er skrevet specielt for geografer, [13] for geologer.

I den senere tid er temporale aspekter blevet genstand for interesse i forbindelse med anvendelse af data, der varierer i både rum og tid. Spatio-temporale semivariogrammer og spatio-temporal *kriging* behandles i bl.a. [21], [22]. Et GIS til håndtering af tidsvarierende data er beskrevet i [23].

Referencer

[1] A. G. Journel and Ch. J. Huijbregts, *Mining Geostatistics*, Academic Press, London, 1978, 600 pp.

[2] Isobel Clark, *Practical Geostatistics*, Elsevier Applied Science, London, 1979, http://www.stokos.demon.co.uk/.

[3] Edward H. Isaaks and R. Mohan Srivastava, *An Introduction to Applied Geostatistics*, Oxford University Press, New York, 1989, 561 pp.

[4] K. Conradsen, A. A. Nielsen, and K. Windfeld, "Analysis of geochemical data sampled on a regional scale," in *Statistics in the Environmental and Earth Sciences*, A. Walden and P. Guttorp, Eds., pp. 283–300. Griffin, 1992.

[5] Noel A. C. Cressie, *Statistics for Spatial Data*, Wiley, New York, revised edition, 1993.

[6] P. A. Burrough and R. A. MacDonnell, *Principles of Geographical Information Systems*, Oxford University Press, 1998.

[7] M. F. Goodchild, B. O. Parks, and L. T. Steyaert, *Environmental Modeling with GIS*, Oxford University Press, 1993.

[8] H. Wackernagel, *Multivariate Geostatistics*, Springer, 1995.

[9] C. V. Deutsch and A. G. Journel, *GSLIB: Geostatistical Software Library and User's Guide*, Oxford University Press, second edition, 1998, Internet http://www.gslib.com/.

[10] Y. Pannatier, *VARIOWIN: Software for Spatial Data Analysis in 2D*, Springer, 1996, Internet http://www-sst.unil.ch/research/variowin/.

[11] Allan Aasbjerg Nielsen, *Analysis of Regularly and Irregularly Sampled Spatial, Multivariate, and Multi-temporal Data*, Ph.D. thesis, Informatics and Mathematical Modelling, Technical University of Denmark, Lyngby, 1994, Internet http://www.imm.dtu.dk/∼aa/phd/.

[12] Allan Aasbjerg Nielsen, "2D semivariograms," in *Proceedings of the Fourth South African Workshop on Pattern Recognition*, Paul Cilliers, Ed., Simon's Town, South Africa, 25–26 November 1993, pp. 25–35.

[13] J. C. Davis, *Statistics and Data Analysis in Geology*, John Wiley & Sons, second edition, 1986.

[14] E. C. Grunsky and F. P. Agterberg, "Spatial and multivariate analysis of geochemical data from metavolcanic rocks in the Ben Nevis area, Ontario," *Mathematical Geology*, vol. 20, no. 7, pp. 825–861, 1988.

[15] E. C. Grunsky and F. P. Agterberg, "SPFAC: a Fortran program for spatial factor analysis of multivariate data," *Computers & Geosciences*, vol. 17, no. 1, pp. 133–160, 1991.

[16] Allan Aasbjerg Nielsen, Knut Conradsen, John L. Pedersen, and Agnete Steenfelt, "Spatial factor analysis of stream sediment geochemistry data from South Greenland," in *Proceedings of the Third Annual Conference of the International Association for Mathematical Geology*, Vera Pawlowsky-Glahn, Ed., Barcelona, Spain, September 1997, pp. 955–960.

[17] Allan Aasbjerg Nielsen, Knut Conradsen, John L. Pedersen, and Agnete Steenfelt, "Maximum autocorrelation factorial kriging," in *Proceedings of the 6th International Geostatistics Congress (Geostats 2000)*, W. J. Kleingeld and D. G. Krige, Eds., Cape Town, South Africa, April 2000.

[18] Knut Conradsen, *Introduktion til Statistik*, Informatics and Mathematical Modelling, Technical University of Denmark, 1984.

[19] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, John Wiley, New York, second edition, 1984.

[20] D. A. Griffith and C. G. Amrhein, *Multivariate Statistical Analysis for Geographers*, Prentice Hall, 1997.

[21] A. K. Ersbøll, "A comparison of two spatio-temporal semivariograms with use in agriculture," in *Lecture Notes in Statistics*, T. G. Gregoire et al., Ed., vol. 122, pp. 299–308. Springer-Verlag, 1997.

[22] Annette Kjær Ersbøll and Bjarne Kjær Ersbøll, "On spatio-temporal kriging," in *Proceedings of the Third Annual Conference of the International Association for Mathematical Geology (IAMG'97)*, Vera Pawlowsky-Glahn, Ed., Barcelona, Spain, September 1997, pp. 617–622.

[23] Thomas Knudsen, *Busstop - A Spatio-Temporal Information System*, Ph.D. thesis, Niels Bohr Institute, Department of Geophysics, University of Copenhagen, Denmark, 1997.

Stikord

afstandsvægtning
autokorrelationsfunktion
autokovariansfunktion
BLUE
dataanalyse
estimationsvarians
geostatistik
GIS
GSLIB
intrinsisk hypotese
kovariansfunktion
kriging
*kriging*varians
Lagrange multiplikator
mindste kvadraters metode
multivariat statistik
nested structures
nugget effekt
ordinær *kriging*
prøvetagningsstrategi
range of influence
regionaliseret variabel
*screening*effekt
semivariogram
semivariogrammodel
sill
simpel *kriging*
spatiel interpolation
spatiel korrelation
stationaritet
statistik
stokastisk funktion
stokastisk proces
stokastisk variabel
variogram
Variowin

Allan Aasbjerg Nielsen er lektor ved Danmarks Tekniske Universitet (DTU), Informatik og Matematisk Modellering (IMM), Sektion for Geoinformatik; professorvikar ved IMM, DTU fra 1995-2001. Civilingeniør i 1978 fra Afdelingen for Elektrofysik, DTU. Ph.d. i 1994 fra IMM, DTU. Arbejde ved Forsvarets Forskningstjeneste 1977-1978. Arbejde på Lavenergihusprojektet ved Laboratoriet for Varmeisolering, DTU, fra 1978 til 1985. Siden 1985 på IMM, DTU. Har arbejdet på flere nationale og internationale projekter om anvendelse af statistiske metoder og *remote sensing* indenfor kortlægning, mineralefterforskning, geologi, geodæsi, oceanografi, landbrug, miljøovervågning og sikkerhed med finansiering fra industri, den Europæiske Union, Danida og forskningsrådene. Meget interesseret i og stor erfaring med udvikling og implementering af matematiske og statistiske metoder vedrørende spatielle, multi-/hyper-variate og multi-temporale data samt anvendelser heraf. *National point of contact* (NPOC) for *concerted actions* om anvendelse af avancerede metoder indenfor *remote sensing* finansieret af den Europæiske Union. Medlem af værts-, program- og videnskabelige kommiteer for internationale konferencer. *Referee* for internationale tidsskrifter og konferencer.