

RIJKSINSTITUUT VOOR VOLKSGEZONDHEID EN MILIEUHYGIENE
BILTHOVEN

Rapport nr. 714802 001

Praktische inleiding tot Kriging,
een methode voor ruimtelijke statistische analyse

L.G. Barendregt en A.L.M. Dekkers

maart 1993

Dit onderzoek werd verricht in opdracht en ten laste van het
Directoraat-Generaal Milieubeheer
Directie Drinkwater, Water, Landbouw
afdeling Coördinatie Milieukwaliteit

Verzendlijst

1. Directoraat-Generaal voor Milieubeheer, DWL afdeling
Coördinatie Milieukwaliteit, t.a.v. ir. A. Roos
2. Directeur-Generaal Milieubeheer ir. M.E.E. Enthoven
3. Plv. Directeur-Generaal Milieubeheer dr. ir. B.C.J. Zoeteman
4. Plv. Directeur-Generaal Milieubeheer mr. G.J.R. Wolters
5. Depôt nederlandse publikaties en nederlandse bibliografie
6. Prof. P.A. Burrough (RUU)
7. Drs. P. Eilers (DCMR)
8. Dr. R. Helmers (CWI)
9. Dr. A. Stein (LUW)
10. Directie RIVM
11. Dr. R.M. van Aalst
12. Drs. J.M.M. Aben
13. Prof. dr. ir. C. van den Akker
14. Ir. B.P.M. Bloemberg
15. Ir. L.J.M. Boumans
16. Mw. ir. G. Doornbos
17. Ir. W. van Duijvenbooden
18. Ir. N.D. van Egmond
19. Drs. J.C.H. van Eijkeren
20. Drs. A. van der Giessen
21. Drs. S.H. Heisterkamp
22. Mw. drs. J.A. Hoekstra
23. Dr. A.S.M. Koeleman
24. Drs. L.H.M. Kohsiek
25. Dr. ir. J.W. de Kwaadsteniet
26. Ir. F. Langeweg
27. Dr. J.E.T. Moen
28. Dr. D. Onderdelinden
29. Drs. E.J. Pebesma
30. Ing. H. Snelting
31. Mw. drs. B.A.M. Staatsen
32. Bibliotheek RIVM
33. Hoofd Bureau Voorlichting en Public Relations
34. Bureau Projecten- en Rapportregistratie
- 35-60. Reserve exemplaren

Inhoudsopgave

Verzendlijst	ii
Inhoudsopgave	iii
Abstract	iv
Samenvatting	iv
1 Doel van het rapport	1
2 Historie	2
3 Wat is Kriging?	2
4 Verband tussen Kriging en de conventionele statistiek	5
5 Intrinsic Random Functions	8
6 Schatting van de variantie-covariantie-matrix	10
7 Ruimte-tijd-modellen	11
8 Kriging en splines	12
9 Het gebruik van Kriging voor de analyse van milieumeet- netten	13
Referenties	14

Abstract

This report deals with Kriging, a spatial interpolation-method that enables making predictions of the value of a measurable property on locations where this property has not actually been measured. The imprecision of these predictions can also be assessed. The relation is described between Kriging (also referred to as Optimum Interpolation) and some conventional statistical techniques. Non-stationary spatial patterns can be modelled using Intrinsic Random Functions. References are given to extensions of Kriging with time as an additional dimension.

The theoretical basis of specification of uncertainty in the predicted values following from uncertainty in the variogram, is as yet unsatisfactory.

Samenvatting

Dit rapport gaat over Kriging, een ruimtelijke interpolatiemethode waarmee ook op niet bemeeten locaties een schatting van de waarde van de meetgrootheid inclusief de betrouwbaarheid van die schatting verkregen kan worden. De relatie wordt gelegd tussen Kriging (ook wel aangeduid als Optimum Interpolatie) en conventionele statistische methoden. Er wordt aangegeven hoe ook niet-stationaire ruimtelijke patronen kunnen worden gemodelleerd met behulp van zogenaamde Intrinsic Random Functions. Referenties worden gegeven naar uitbreidingen van Kriging methoden met tijd als een extra dimensie.

De theoretische achtergrond voor het doorrekenen van de onzekerheid in het schatten van het variogram naar de onzekerheid in de voorspellers is nog een zwak punt in de theorie.

1 Doel van het rapport

De analyse van ruimtelijke patronen in (meet)gegevens wordt steeds belangrijker. Dankzij de spectaculaire ontwikkelingen in de prestaties van computer apparatuur kunnen gecompliceerde berekeningen die tot voor kort praktisch onmogelijk waren, momenteel met een zekere eenvoud worden uitgevoerd. Tevens kunnen allerlei gegevensbestanden aan elkaar gekoppeld worden via geografische informatie systemen. Dit maakt dat het toepassen van technieken voor ruimtelijke analyse op het RIVM een stormachtige ontwikkeling doormaakt. In dit kader heeft het CWM in nauwe samenwerking met (milieu)meetnet-deskundigen een pilotstudie uitgevoerd ter vergroting van de mathematische expertise ten behoeve van een adequate advisering aan diverse laboratoria.

De doelstelling van deze pilotstudie is dan ook tweeledig:

- Kennisverwerving betreffende ruimtelijke statistiek, speciaal gericht op Kriging met dit rapport als schriftelijk resultaat.
- Toepassing van deze techniek op diverse milieu compartimenten. Als onderwerp is daarbij gekozen voor de ruimtelijke analyse van het verband tussen zwavel-depositie en de concentratie van sulfaat in het grondwater ([Dekkers en Barendregt (1993)]).

Het doel van dit rapport is het verband te schetsen tussen Kriging en enkele andere bekende statistische methoden die voor interpolatie of verwante vraagstellingen geschikt zijn, resp. gebruikt worden.

Na een kort resumé over de historische achtergrond van Kriging wordt in hoofdstuk 3 in woorden aangeduid wat voor berekening Kriging is. Hoofdstukken 4 t/m 8 bevatten technische informatie over Kriging. In hoofdstuk 4 worden de Kriging-formules aannemelijk gemaakt met argumenten die geen al te sterk beroep doen op kennis van de matrix- en vectorrekening. Voor meer gedetailleerde informatie wordt verwezen naar [Christensen (1991)], [Corsten (1989)], [Cressie (1991)] en [Journel en Huijbregts (1978)]. In de hoofdstukken 5 en 6 van dit rapport worden enige eigenschappen van Kriging besproken die in de literatuur nogal eens onderbelicht of verkeerd beklemtoond worden.

Er wordt ook ingegaan op de mogelijkheden om Kriging uit te voeren op gegevens die zowel op verschillende tijdstippen als op verschillende plaatsen zijn verzameld.

Tenslotte wordt in hoofdstuk 9 ingegaan op het gebruik van Kriging bij de analyse van milieu-meetnetten.

2 Historie

Kriging is een ruimtelijke interpolatie-methode die berust op een stochastisch wiskundig model voor de meetgroottheden. Uitgaande van dit model wordt een schatting/voorspelling gemaakt van de meetwaarde in punten waar niet is gemeten. Ook is het mogelijk om gemiddelden van de meetgroottheid over een bepaald gebied te voorspellen. Bovendien is het mogelijk om voor elke voorspelling een standaardafwijking te berekenen; deze kan dan worden gebruikt als maat voor de betrouwbaarheid van de voorspelling.

Bij ruimtelijke interpolatie van meetgegevens die behept zijn met een meetfout kan een zeer grillig beeld ontstaan, als de voorspellingen op een rooster van niet-meetpunten grafisch worden uitgebeeld. Bij Kriging is het mogelijk om aan het stochastisch model voor de meetgroottheid een meetfout toe te voegen; als met dat model als uitgangspunt wordt gekrigit, wordt het voorspellingsoppervlak gladder. Hoe groter de variantie van de meetfout, hoe gladder het voorspellingsoppervlak.

Kriging is ontstaan buiten de kring van de wetenschappers die de gangbare statistische redeneringen hebben bedacht en gepropageerd. De naam Kriging bv. komt van Krige, een Zuidafrikaanse mijnbouwkundige die goudconcentraties wilde voorspellen. Krige raakte in gesprek met Matheron, een wiskundige verbonden aan de Ecole des Mines in Fontainebleau. Deze heeft, vermoedelijk na raadpleging van allerlei Sowjet-wiskundigen zoals Gel'fand, Vilenkin en Kolmogorov, een methode geïntroduceerd, die hij Kriging noemde. Via Geoffrey S. Watson (Princeton) is Kriging tenslotte bij de conventionele statistici bekend geworden.

In de meteorologie is een interpolatie-techniek onder de naam Optimum Interpolation Scheme geïntroduceerd door [Gandin (1965)]. Dit Schema blijkt gelijk te zijn aan Kriging en is geïnspireerd door het werk van Wiener en Kolmogorov. Zie bv. [Cressie (1990)], blz 247. In dit artikel kan verdere informatie over de ontstaansgeschiedenis van Kriging worden gevonden.

Een veel gebruikt boek bij de eerste kennismaking met Kriging is [Journel en Huijbregts (1978)].

3 Wat is Kriging?

Kriging is, zoals reeds eerder is opgemerkt, een interpolatie-methode. Uitgangspunt is een meetgroottheid die op elke plaats binnen een bepaald gebied gemeten zou kunnen worden. Het is noch zinnig, noch doenlijk om

die meting inderdaad op elk punt in dat gebied te verrichten; daarom is de meetgrootte slechts in een eindig aantal punten in het gebied gemeten. De meetpunten hoeven **niet** volgens een rechthoekig rooster gerangschikt te zijn.

Er moet nu voor een of meerdere punten waar niet gemeten is, een voorspelling van de waarde van de meetgrootte ter plaatse worden gegeven. Dat kan gebeuren door interpolatie vanuit de punten waar wel is gemeten, en wel door een gewogen gemiddelde te nemen van de waarnemingen op die plaatsen, waarbij de dichtbij gelegen punten een zwaarder gewicht hebben dan de veraf gelegen punten. Een voorbeeld van zo'n interpolatie is de "Distance Density weighting function interpolation" (DDI) beschreven in [van Egmond en Onderdelinden (1981)].

Ook een Kriging-schatting is op te vatten als een gewogen gemiddelde van de meetwaarden in de meetpunten. Sommige van de gewichten kunnen daarbij negatief worden, hoewel de som van de gewichten $+1$ is, net als bij DDI. Kenmerkend voor Kriging is dat de gewichten, die aan de diverse meetpunten moeten worden toegekend, worden berekend volgens een bepaald optimaliteits-criterium. [van Egmond en Onderdelinden (1981)] noemen Kriging dan ook "Optimum Interpolation" (OI). Het optimaliteits-criterium is: minimale variantie van de voorspeller, met als uitgangspunt een stochastisch model voor de meetgrootte in dat gebied.

Uit deze beschrijving blijkt dat er een stochastisch model voor de meetgrootte moet worden verondersteld alvorens er kan worden gekriged. De stochastische modellen die hiervoor kunnen gebruikt staan beschreven in onder meer [Christensen (1991)]. Die modellen zouden "stochastische functies" kunnen worden genoemd, want bij elk punt in het platte vlak hoort een stochastische variabele. In het Engels heten ze in elk geval "Random Fields".

Bij de keuze van het stochastisch model wordt vaak als hulpmiddel het zogenaamde **variogram** gebruikt. Het variogram is een grafische weergave van het oplopen van de verschillen tussen de meetwaarden van twee meetpunten, naarmate de afstand tussen die meetpunten groter wordt. Van elk puntenpaar (n meetpunten, dan $n(n-1)/2$ paren) wordt uitgerekend

- afstand
- gekwadrateerd verschil tussen de meetwaarden.

Men zou nu alle $n(n-1)/2$ paren in een punten-diagram kunnen uitbeelden met op de horizontale as de afstand tussen de meetpunten en op de

verticale as het gekwadrateerde verschil. Dat levert echter een puntenwolk met veel te veel punten op. In plaats daarvan worden de $n(n-1)/2$ paren gegroepeerd volgens afstandsklassen (bijvoorbeeld 10 afstandsklassen). Van elke afstandsklasse wordt het gemiddelde van de kwadratische meetwaardeverschillen berekend. Deze worden vervolgens in een punten-diagram uitgezet, dat variogram genoemd wordt. In een variogram staan dus niet zo veel punten (misschien maar 10), maar ze liggen doorgaans veel ordelijker dan de oorspronkelijke $n(n-1)/2$ punten. Door deze punten wordt een kromme getrokken, veelal berekend via een niet lineaire regressie.

Een stochastische functie (Random Field) heet **stationair**, als zijn eigenschappen invariant zijn onder ruimtelijke translaties. Als ze bovendien invariant zijn onder rotaties, dan is de stochastische functie **isotroop**.

Bij een stationaire stochastische functie moet derhalve zowel de verwachtingswaarde als de variantie plaats-onafhankelijk zijn. Als de functie bovendien isotroop is, dan is de correlatie tussen de functie-waarden op twee verschillende plaatsen uitsluitend afhankelijk van de onderlinge afstand en niet meer van de richting.

De uitdrukking "stochastische" functie suggereert misschien, dat zo'n functie continu is. Er zijn echter heel wat meetgrootheden die, zelfs als ze op twee zeer dicht bij elkaar gelegen punten zijn gemeten, twee sterk verschillende uitkomsten geven. Met name geldt dit voor concentraties in de bodem van een of andere stof. De reeds eerder genoemde Zuidafrikaanse mijnbouwkundige Krige nam dit waar voor goud-concentraties, en voerde de term nugget-effect in (nugget=goudklompje). Een meetgrootheid met een **nugget-effect** is een meetgrootheid waarbij het verschil tussen meetwaarden op twee verschillende plaatsen niet naar 0 nadert, als de afstand tussen die twee plaatsen willekeurig klein wordt. Een voorbeeld van een meetgrootheid die in het algemeen geen nugget-effect behoort te vertonen is de stijghoogte van grondwater. De concentratie van een of ander ion in dat grondwater zou echter weer wél een nugget-effect kunnen hebben.

Bij vele gegevensverzamelingen blijkt de meetgrootheid niet stationair te zijn (of beter uitgedrukt: blijkt dat er geen stationaire stochastische functie is, die geloofwaardig is als model voor de meetgrootheid). Exploratieve data-analyse moet dan de toepasbaarheid van een ander model suggereren.

Een van de mogelijkheden die zich kunnen voordoen is dat de data in een bepaalde richting geleidelijk oplopen. Men zegt dan dat de gegevens een **oplopende trend** hebben in die richting. Als model kan worden genomen: meetgrootheid = (lineaire functie van de coördinaten) + (stationaire stochastische functie).

4 Verband tussen Kriging en de conventionele statistiek

Aan het volgende voorbeeld kan worden gedemonstreerd hoe Kriging is bedacht, hoe het werkt en wat de overeenkomst is met de conventionele statistiek. Laat y_1 de meting zijn van een bepaalde grootte op de plaats u_1 en y_0 de fictieve meting van die grootte op de plaats u_0 ; y_1 is daadwerkelijk gemeten, y_0 echter niet; y_0 moet uit y_1 worden voorspeld.

We zouden een kansverdeling voor y_0 kunnen postuleren met verwachtingswaarde μ_0 ; μ_0 bekend. De voorspelling voor y_0 wordt dan eenvoudig μ_0 . We kunnen y_1 inschakelen bij de voorspelling door een gezamenlijke kansverdeling voor y_0 en y_1 te postuleren, met verwachtingswaarden resp. μ_0 en μ_1 , en met variantie-covariantie-matrix

$$\begin{pmatrix} \text{var}(Y_0) & \text{cov}(Y_0, Y_1) \\ \text{cov}(Y_1, Y_0) & \text{var}(Y_1) \end{pmatrix} = \begin{pmatrix} \sigma_{00} & \sigma_{01} \\ \sigma_{10} & \sigma_{11} \end{pmatrix} \quad (1)$$

waarbij natuurlijk $\sigma_{01} = \sigma_{10}$.

Als we ook nog veronderstellen dat (Y_0, Y_1) biviaat normaal verdeeld is, dan kunnen we afleiden dat de voorwaardelijke verdeling van Y_0 , gegeven $Y_1 = y_1$ normaal is met verwachtingswaarde $\mu_0 + \sigma_{01}\sigma_{11}^{-1}(y_1 - \mu_1)$ en variantie $\sigma_{00} - \sigma_{01}^2\sigma_{11}^{-1}$. Een voor de hand liggende voorspelling voor y_0 is nu

$$E(y_0|Y_1 = y_1) = \mu_0 + \sigma_{01}\sigma_{11}^{-1}(y_1 - \mu_1) \quad (2)$$

Opmerkingen:

We maken hier gebruik van het gecorreleerd zijn van Y_0 met Y_1 . Als $\sigma_{01} = 0$, dan is de voorspeller voor y_0 weer μ_0 .

De voorspeller voor y_0 wordt vaak genoteerd als $\hat{y}_0$. Alleen al om die reden zijn velen geneigd om tijdens informele gedachtenwisselingen $\hat{y}_0$ de schatter van y_0 te noemen. Maar in de mathematische statistiek is de term 'schatting' gereserveerd voor uitspraken over een onbekende, niet-stochastische *parameter* van een verdeling, vandaar de term voorspelling.

De veronderstelling van bi-variate normaliteit is lang niet altijd geloofwaardig. Toch is de voorspelformule (2) ook in dat geval nog verdedigbaar. Hij voldoet namelijk aan de volgende eigenschap:

Van alle lineaire voorspellers $\hat{y}_0 = \alpha + \beta y_1$ is $\hat{y}_0 = \mu_0 + \sigma_{01}\sigma_{11}^{-1}(y_1 - \mu_1)$ degene waarvoor $E(\hat{Y}_0 - Y_0)^2$ minimaal is.

Bewijs:

$$E(\hat{Y}_0 - Y_0)^2 = E(\alpha + \beta Y_1 - Y_0)^2 = E(\alpha + \beta \mu_1 - \mu_0 + \beta(Y_1 - \mu_1) - (Y_0 - \mu_0))^2 =$$

$$(\alpha + \beta\mu_1 - \mu_0)^2 + \text{var}(\beta(Y_1 - \mu_1) - (Y_0 - \mu_0)) =$$

$$(\alpha + \beta\mu_1 - \mu_0)^2 + \beta^2\sigma_{11} - 2\beta\sigma_{01} + \sigma_{00}$$

Door $\alpha = \mu_0 - \beta\mu_1$ te kiezen wordt de beginterm minimaal, nl. 0. Er blijft een kwadratische uitdrukking in β over. Deze is minimaal voor $\beta = \sigma_{01}/\sigma_{11}$. Dus de $\hat{Y}_0$ waarvoor $E(\hat{Y}_0 - Y_0)^2$ minimaal is, is

$$\hat{Y}_0 = \mu_0 - \beta\mu_1 + \beta Y_1 = \mu_0 + \sigma_{01}\sigma_{11}^{-1}(Y_1 - \mu_1)$$

Deze $\hat{Y}_0$ heet de Best Linear Unbiased Predictor (BLUP).

De definitie van het begrip BLUP kan worden uitgebreid naar situaties met meer dan twee meetplaatsen, en naar situaties waarin niet alle verwachtingswaarden bekend zijn. **Kriging is hetzelfde als het geven van een Best Linear Unbiased Prediction volgens deze uitgebreide definitie.**

In de algemene situatie is de grootte y gemeten op n verschillende plaatsen $u_1, \dots, u_n$ met als meetresultaat resp. $y_1, \dots, y_n$. Er moet een voorspelling worden gemaakt voor y_0 , de waarde van y op de plaats u_0 . (Ik ga hierbij uit van de notatie van [Christensen (1991)]). Op elke plaats zijn naast y ook nog enige covariabelen $x_1, \dots, x_p$ gemeten. Tot die covariabelen behoren vaak de twee plaatscoördinaten.

De n -dimensionale vector $Y = (y_1, \dots, y_n)'$ heeft verwachtingswaarde

$$E(Y) = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}. \quad (3)$$

In verkorte vorm : $E(Y) = X\beta$ met $\beta_1, \dots, \beta_p$ onbekende parameters. De variantie-covariantie-matrix van Y is de symmetrische matrix Σ , waarvan de coëfficiënten wél bekend moeten zijn. Voorts:

$$E(Y_0) = (x_{01}, \dots, x_{0p}) \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \quad (4)$$

$$\text{cov}(Y_i, Y_0) = \sigma_{i0}$$

$$\Sigma_{Y0} = (\sigma_{10}, \dots, \sigma_{n0})'$$

Deze veronderstellingen zijn dezelfde als die welke gedaan worden in het algemene lineaire model dat ten grondslag ligt aan de multiële regressie. De

gewichten-matrix Σ kan daar in beginsel algemeen zijn, maar als het tot een daadwerkelijke regressie-berekening komt, wordt Σ toch meestal diagonaal genomen. Bij Kriging daarentegen is het niet-diagonaal zijn van Σ juist essentieel.

Met de formule $EY = X\beta$ kunnen o.a. verschillen tussen deelgebieden worden gemodelleerd, zoals verschillen in grondsoort. Stel er kunnen drie grondsoorten worden onderscheiden en stel dat moet worden aangenomen dat de verwachtingswaarde van y per grondsoort verschilt. Dan wordt voor X een $n \times 3$ -matrix genomen van de vorm

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \quad (5)$$

In dit voorbeeld zijn de waarnemingen op volgorde van grondsoort gezet: eerst 3 van grondsoort 1, dan 2 van grondsoort 2 en tenslotte 3 van grondsoort 3.

De voorspelformule volgens Kriging voor $\hat{y}_0$ wordt gegeven op blz. 268 in [Christensen (1991)] als :

$$\hat{y}_0 = x'_0 \hat{\beta} + \Sigma'_{Y_0} \Sigma^{-1} (Y - X \hat{\beta}) \quad (6)$$

Als β bekend zou zijn dan zou deze formule een generalisatie van de tweepunts voorspelling (2) zijn. Nu β onbekend is, wordt hij vervangen door een schatter $\hat{\beta}$, waarbij $\hat{\beta}$ wordt geschat volgens

$$\hat{\beta} = (X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} Y \quad (7)$$

hetgeen de gewogen kleinste kwadratenschatter van β uit $y_1, \dots, y_n$ en X is, met gewichten-matrix Σ . In [Christensen (1987)], maar ook op andere plaatsen in de literatuur, kan men bewezen vinden dat $\hat{y}_0$ volgens (6) met $\hat{\beta}$ gekozen volgens (7) inderdaad de BLUPredictor is.

Bij de berekening van $\hat{y}_0$ heeft men de inversie van de $n \times n$ matrix Σ nodig. Dat is heden ten dage goed uitvoerbaar, maar vele gebruikers wagen zich daar toch niet aan. Zij gebruiken bij daadwerkelijke berekeningen van

(6) alleen de naburige punten van de plaats u_0 . Het computer-programma BLUEPACK [Delfiner e.a. (1978)] neemt bv. alleen de 8 meest naburige waarnemingsplaatsen in de berekening mee.

Een bijkomend voordeel van deze handelswijze is dat de procedure robuust is tegen afwijkingen van de stationariteit.

5 Intrinsic Random Functions

Toen Delfiner omstreeks 1977 Kriging in Nederland introduceerde, wees hij met nadruk op de mogelijkheid om het toe te passen op niet-stationaire, in een plat vlak gemeten variabelen. Hij heeft in BLUEPACK de mogelijkheid verwerkt om als model voor de te meten variabele een Intrinsic Random Function te nemen. Dit begrip is in 1973 door Matheron ingevoerd.

Intrinsic Random Functions kunnen worden opgevat als onderdeel van een nog grotere familie, de zgn. Generalized Random Functions van Gel'fand en Vilenkin. **Intrinsic Random Functions zijn intrinsiek niet-stationair, d.w.z. zij zijn niet stationair te maken door aftrek van een ruimtelijke trend.**

Van de door Matheron en Delfiner aangegeven mogelijkheid om bij Kriging met Intrinsic Random Functions te werken is in het grondwater-onderzoek vanaf het begin veelvuldig gebruik gemaakt. Het doet in dat licht bezien vreemd aan dat in een aantal exposé's over Kriging deze mogelijkheid onderbelicht of zelfs onbelicht is gebleven. Wat voor vele Nederlandse onderzoekers een dagelijks gebruikt analyse-instrument is (zie bv. [Stein (1988)], fig 2a en 2b), werd door auteurs als [Ripley (1981)] en [Cressie (1990)] niet vermeldenswaard gevonden.

Het eenvoudigste voorbeeld van een stationaire Intrinsic Random Function is het model, dat een lineair variogram heeft. Als $Y(u)$ de stochastische variabele is, dan geldt per definitie

$$E(Y(u_1) - Y(u_2)) = 0$$

$$\text{var}(Y(u_1) - Y(u_2)) = 2\theta|u_1 - u_2| \quad (8)$$

$u_1, u_2 \in R^2, \theta > 0$ bekend. Het variogram hangt dus lineair van de afstand $|u_1 - u_2|$ af.

Meer algemeen:

$$\text{cov}(Y(u_1) - Y(u_2), Y(u_3) - Y(u_4)) = \theta[|u_1 - u_4| + |u_2 - u_3| - |u_1 - u_3| - |u_2 - u_4|]$$

met $u_1, \dots, u_4 \in R^2$.

Hiermee zijn alleen de paarsgewijze verschillen gemodelleerd, maar nog niet de waarden van $Y(u)$ zelf. Het blijkt niet mogelijk te zijn om $\text{var}(Y(u))$ onafhankelijk van u te definiëren zonder het bovenstaande geweld aan te doen. Met andere woorden: er bestaat geen stochastische functie $Y(u)$ met stationaire variantie en auto-covarianties, die voldoet aan (8).

Zie overigens ook [Christensen (1991)], blz. 288 en 289, waar het model met lineair variogram en nugget effect wordt geformuleerd.

De volgende vragen rijzen:

1. Waarom zou men zo'n model poneren?
2. Is het wel consistent; leiden de voorwaarden (8) niet tot tegenstrijdigheden, zoals lineaire combinaties met negatieve varianties?
3. Hoe kan zo'n model op natuurlijke wijze ontstaan?

Ad 1. Vele empirische variogrammen blijken lineair te zijn, zoals vermeld o.a. door [Delhomme (1978)]. Hierdoor ontstond bij geostatistici de behoefte aan een stochastisch model met eigenschap (8).

Ad 2. Het model kan worden opgevat als de limiet van de volgende familie $\{Y_\alpha(u); \alpha > 0\}$:

$$\begin{aligned} E(Y_\alpha(u)) &= \mu \quad \forall \alpha \\ \text{cov}(Y_\alpha(u_1), Y_\alpha(u_2)) &= \theta \cdot \alpha \cdot \exp\left(-\frac{|u_1 - u_2|}{\alpha}\right) \end{aligned} \quad (9)$$

Laat nu $\alpha \rightarrow \infty$. Dan is $\lim_{\alpha \rightarrow \infty} \text{var}(Y_\alpha(u)) = +\infty$, maar

$$\begin{aligned} \lim_{\alpha \rightarrow \infty} \text{var}(Y_\alpha(u_1) - Y_\alpha(u_2)) &= \\ \lim_{\alpha \rightarrow \infty} \theta \cdot 2\alpha \left(1 - \exp\left[-\frac{|u_1 - u_2|}{\alpha}\right]\right) &= 2\theta \cdot |u_1 - u_2| \end{aligned}$$

Van het model met α eindig is bekend dat het consistent is (het wordt ook gebruikt in de geostatistiek), dus het is plausibel dat de limiet ervan voor $\alpha \rightarrow \infty$ ook nog consistent is.

Ad 3. Uit Ad 2 blijkt dat het model (8) met het lineaire variogram als een limiet kan worden beschouwd van het model met exponentieel covariogram en toenemende 'correlatie-afstand'.

Nog enkele eigenschappen:

Als $Y(u)$ slechts beschouwd wordt op een rechte lijn in het platte vlak, dan gedraagt hij zich als een Brownse beweging. $Y(u)$ is dus wèl continu, maar nergens differentieerbaar. Dat geldt trouwens ook voor het proces (9).

Als het model met lineair variogram wordt gebruikt om te krijgen moet de $n \times p$ -matrix X minimaal de niveau-vector $(1, \dots, 1)'$ bevatten.

6 Schatting van de variantie-covariantie-matrix

Tot hiertoe is steeds aangenomen dat Σ , de variantie-covariantie-matrix, bekend was. Als dat niet zo is, dan moet Σ uit de waarnemingen worden geschat. Aangezien de symmetrische $n \times n$ -matrix $\Sigma \frac{1}{2}n(n+1)$ onafhankelijke parameters heeft, zal dat niet uit de n waarnemingen kunnen gebeuren, tenzij men voor Σ een model met een beperkt aantal vrije parameters wil veronderstellen.

Het is goed om hier nog even in herinnering te brengen hoe bij lineaire regressie, waar $\Sigma = \sigma^2 I$, wordt gehandeld. Hier bevat Σ slechts één onbekende parameter, nl. σ^2 , waarvoor een schatter bestaat. Als van Y mag worden verondersteld, dat hij multivariaat normaal verdeeld is, dan is $\hat{\sigma}^2$ verdeeld als $\sigma^2 \cdot \chi^2(n-p)/(n-p)$, dus $\hat{\sigma}^2$ is asymptotisch raak (de verdeling van $\hat{\sigma}^2$ verdicht zich tot een identieke verdeling rond σ als n naar $+\infty$ gaat). Voor $\hat{\beta}$ en daarvan afgeleide grootheden, zoals voorspellers, kunnen betrouwbaarheidsgebieden worden berekend, die zijn gebaseerd op een F-verdeling met $n-p$ vrijheidsgraden in de noemer.

Bij Kriging kan men in eerste instantie analoog te werk gaan, nl. door $\Sigma = \theta \cdot V$ te veronderstellen, met θ onbekend, V bekend (zie ook [Christensen (1991)], blz. 277). Dan kunnen voor de voorspellers gestudentizeerde betrouwbaarheidsintervallen worden opgesteld. Bij de meeste toepassingen zal V echter niet bekend zijn, maar nog één of meer parameters θ_2 bevatten. Die parameters kunnen vaak nog uit de waarnemingen worden geschat, maar er is geen algemeen geaccepteerde theorie voorhanden, waarmee men betrouwbaarheidsintervallen voor de voorspellers zou kunnen berekenen.

Bij gebrek aan theorie kan men een reeks van 'geloofwaardige' waarden voor de parameters van Σ invullen, en dan zien hoe de schattingen $\hat{Z}(x_0)$ en de varianties van $\hat{Z}(x_0) - Z(x_0)$ variëren. Geloofwaardig betekent hier: ruimschoots vallend binnen het 95%-betrouwbaarheidsgebied voor θ_2 . De ervaring van vele auteurs (zie o.a. [Barendregt (1987)], table 2) is dat de voorspeller $\hat{Z}(x_0)$ niet veel varieert, maar dat de variantie van $\hat{Z}(x_0) - Z(x_0)$ in aanzienlijke mate varieert. Christensen vermeldt op blz.

279 [Christensen (1991)] dat de varianties van voorspellers die een geschatte Σ gebruiken (zgn. plug-in-estimators), misleidend klein zijn.

Deze constatering doet afbreuk aan de betekenis van de betrouwbaarheidsintervallen voor Kriging-voorspellers.

7 Ruimte-tijd-modellen

Met behulp van meetnetten worden over de tijd herhaalde ruimtelijke gegevensverzamelingen verkregen. Het ligt voor de hand om Kriging ook op dergelijke gegevens toe te passen door de tijd als een extra-dimensie toe te voegen aan de ruimte. Formulering van een (co)variogram maakt het wiskundig gereedschap compleet.

Een voorbeeld van toepassing van Kriging in ruimte en tijd wordt gevonden in [Bilonick (1985)]. De gegevens-verzameling bestaat uit sulfaat-depositie-metingen in neerslag in het Noordoosten van de Verenigde Staten, welke zijn gedaan tussen 1965 en 1981. De meetwaarden $Z(x, t)$ hangen af van de plaatscoördinaten $x \in R_2$ en de tijd $t \in R_1$. Anders geformuleerd: Z hangt af van $(x, t) \in R_2 \times R_1 = R_3$. Bilonick kiest als variogram voor $Z(x, t)$:

$$\text{var}[Z(x, t) - Z(x + \Delta x, t + \Delta t)] = \gamma_s(\Delta x) + \gamma_t(\Delta t) \quad (10)$$

dus een optelsom van een puur ruimtelijk variogram en een puur tijds-variogram. Wiskundig gezien leidt dit niet tot inconsistenties, maar de vraag is of de auteur zich heeft gerealiseerd dat

$$\text{var}[Z(x_1, t_1) - Z(x_1, t_2) - Z(x_2, t_1) + Z(x_2, t_2)] = 0$$

Het verschil tussen de meetwaarden op twee tijdstippen wordt dus over het gehele veld constant verondersteld. [Egbert en Lettenmaier (1986)] bestuderen eveneens sulfaat-concentraties in neerslag in het Noordoosten van de Verenigde Staten. Zij schatten de variogrammen van week-gemiddelden, seizoen-gemiddelden en jaar-gemiddelden op een parameter-vrije manier. [van Egmond en Onderdelinden (1981)] gebruiken gegevens die op verschillende plaatsen en tijdstippen zijn gemeten. Zij beschrijven in hun inleidende hoofdstuk het Optimum Interpolation Scheme (Kriging dus) en enige verwante interpolatie-technieken. In een later hoofdstuk beschrijven zij hoe

zij bepaalde covarianties schatten. De daarin gegeven afleidingen zijn echter zeer moeilijk na te trekken. De methodes worden toegepast op concentratiemetingen van enkele verontreinigende gassen in de lucht.

Rodriguez-Iturbe en Mejia nemen als model voor het covariogram:

$$\text{cov}[Z(x, t), Z(x + \Delta x, t + \Delta t)] = \sigma(\Delta x) \cdot \rho(\Delta t) \quad (11)$$

waarbij $\sigma(\Delta x)$ een zuiver ruimtelijk covariogram is en $\rho(\Delta t)$ een correlogram in de tijd [Rodriguez-Iturbe en Mejia (1974)]. Dit model heeft niet het bezwaar dat het model (10) van Bilonick heeft.

Het model (11) zou nog kunnen worden uitgebreid tot

$$\text{cov}[Z(x, t), Z(x + \Delta x, t + \Delta t)] = \sigma(\Delta x - v \cdot \Delta t) \rho(\Delta t)$$

waarbij v de snelheidsvector van de heersende windrichting is.

8 Kriging en splines

Een bekende ruimtelijke interpolatie-methode is de "thin-plate smoothing spline". Zie [Cressie (1991)]. Dit is een twee-dimensionale variant van de "cubic smoothing spline", die als doel heeft het trekken van een kromme door een aantal punten. Uit [Cressie (1991)], blz. 181 kan worden opgemaakt dat thin-plate smoothing spline equivalent is met Kriging met als gegeneraliseerde covariantie:

$$K(u_1, u_2) = |u_1 - u_2|^2 \cdot \log(|u_1 - u_2|) + C \cdot \delta(|u_1 - u_2|)$$

waarbij $\delta(\cdot)$ de Kronecker-delta is en C een geschikt gekozen positieve constante. Het bijbehorende Intrinsic Random Field is een IRF(1). Dit houdt in dat de niveau-vector en de beide plaatscoördinaten verplicht in de X -matrix moeten worden opgenomen. Een nog duidelijker referentie voor deze constatering is [Dubrule (1983)].

Cubic smoothing spline kan eveneens als een bijzonder geval van Kriging worden opgevat. De gegeneraliseerde covariantie is hierbij:

$$K(u_1, u_2) = |u_1 - u_2|^3 + C \cdot \delta(|u_1 - u_2|)$$

9 Het gebruik van Kriging voor de analyse van milieumeetnetten

Zoals in hoofdstukken 3 en 4 beschreven, kan Kriging worden opgevat als regressie-analyse aangepast voor ruimtelijke data. Het essentiële verschil met gewone regressie is dat de metingen niet onafhankelijk verondersteld worden. Kriging wordt op het RIVM toegepast op data uit milieu meetnetten, zie bijvoorbeeld [Pebesma (1992)].

Regressie-analyse is de meest frequent gebruikte statistische techniek. Ze wordt aangewend voor het aantonen en kwantificeren van verbanden tussen variabelen, met daarbij aangegeven de mate van onzekerheid. Voorbeelden zijn relaties tussen gemeten gehalten van stoffen, tussen gehalten en andere (ruimtelijk variërende) kenmerken als bodemtype, verstedelijking, gebruik van de bodem, etc.

Modellering van de ruimtelijke samenhang met Kriging maakt het ook mogelijk fijnmaziger schattingen op niet gemeten plaatsen te geven dan met klassieke regressie mogelijk is. De prijs hiervoor is een grotere afhankelijkheid van de modelveronderstellingen: de optimaliteit van de interpolatie en de juistheid van de weergegeven statistische onbetrouwbaarheid staat of valt met de juistheid van het toegepaste model voor de ruimtelijke correlatie.

Bij het implementeren van de interpolator moet daarom grote zorg worden besteed aan het onderzoek naar de ruimtelijke structuur. Daarbij speelt naast gegevensanalyse proceskennis een belangrijke rol. Zo kan men bij resultaten van metingen in het diepe grondwater verwachten dat bodemtype en bodemgebruik alsook de verplaatsingssnelheid van het grondwater van belang zijn (cf. [Pebesma (1992)]). Bij een analyse van de gegevens van geheel Nederland dienen deze variabelen dus in het model te worden opgenomen.

Anderzijds kan bestudering van de ruimtelijke correlatie in de metingen een bijdrage aan de kennis van het ruimtelijke verspreidingsproces leveren. Wanneer de meetpunten zover uiteen liggen of anderszins zo zijn gesitueerd dat er geen ruimtelijke correlatie is, is het in feite niet mogelijk voorspellingen te doen op een fijner grid dan dat van de meetpunten, tenzij door gebruik te maken van externe proceskennis. Uitbreidingen van Kriging/Optimale Interpolatie in de richting van Kalmanfiltering kunnen de benodigde proceskennis integreren met de data. Dit vraagt om een gezamenlijke inzet van milieudeskundigen en experts van diverse wiskundige disciplines.

Bovengenoemde mogelijkheden van Kriging zijn direkt inzetbaar voor de

optimalisatie en integratie van de RIVM milieumeetnetten. Optimale schattingen worden bereikt door gebruik te maken van de ruimtelijke correlatie en van de onderlinge verbanden tussen metingen. De weergegeven onzekerheden in de schattingen kunnen worden gebruikt bij de verdere optimalisatie van het milieumeetnet. Kriging verschaft het instrument om op doorzichtige wijze metingen uit verschillende milieucompartimenten met elkaar in verband te brengen.

Een nog vrij nieuwe en zeer belangrijke ontwikkeling binnen de ruimtelijke statistiek zijn de tijd-ruimte modellen. Daarin wordt naast de ruimtelijke afhankelijkheid ook de tijdsafhankelijkheid zoals dat in tijdreeksanalyse al wordt gedaan, gemodelleerd. Deze modellen zijn veelbelovend voor bijvoorbeeld de gegevens die in het Landelijk Luchtmeetnet beschikbaar zijn. Een verdere studie en toepassing van deze modellen is voor het RIVM zeer aan te bevelen.

Milieumeetnetten kunnen alleen dan optimaal worden benut wanneer de eruit resulterende data systematisch worden beheerd en grondig worden geanalyseerd. Het beschikbaar hebben van voldoende menskracht en gebruikersvriendelijk computerinstrumentarium is daarvoor vereist. Het is aan te bevelen voor het RIVM goed gevalideerde en gedocumenteerde, gebruikersvriendelijke en krachtige Kriging programmatuur aan te schaffen. Hierbij dient dan de koppeling van deze programmatuur met de bestaande GIS-programmatuur zoals ArcInfo centraal te staan. Daarnaast is het wenselijk dat er programmatuur beschikbaar is waarmee voordat er gekrigit wordt, een grondige exploratieve data-analyse uitgevoerd kan worden.

Referenties

- [Barendregt (1987)] Barendregt, L.G.
The Estimation of the Generalized Covariance when it is a linear combination of two known generalized Covariances
Water Resources Research, Vol. 23, no. 4, pg 583-590, 1987
- [Bilonick (1985)] Bilonick, R.A.
The space-time-distribution of sulfate deposition
in the northeastern United States
Athmospheric Environment, Vol. 19, ppg 1829-1845, 1985

- [Christensen (1987)] Christensen, R.
Plane Answers to Complex Questions: The Theory of Linear Models
Springer-Verlag, New York, 1987
- [Christensen (1991)] Christensen, R.
Linear Models for Multivariate, Time Series, and Spatial Data
Springer-Verlag, New York, 1991
- [Corsten (1989)] Corsten, L.C.A.
Interpolation and optimal linear prediction
Statistica Neerlandica, Vol. 43, ppg 69-84, 1989
- [Cressie (1990)] Cressie, N.
The Origins of Kriging
Mathematical Geology, Vol.22, ppg 239-252, 1990
- [Cressie (1991)] Cressie, N.
Spatial Data Analysis
Wiley, New York, 1991
- [Dekkers en Barendregt (1993)] Dekkers, A.L.M. en Barendregt, L.G.
Een ruimtelijke analyse van het verband tussen zwavel-depositie en de
concentratie van sulfaat in het grondwater
RIVM-rapport 714802 002
- [Delhomme (1978)] Delhomme, J.P.
Kriging in the Hydrosiences
Advances in Water Resources, Vol. 1, No. 5, ppg 251-266, 1978
- [Delfiner e.a. (1978)] Delfiner, P. e.a.
BLUEPACK-3D-manual. Centre de Geostatistique, Ecole des Mines,
Fontainebleau, 1978
- [Delfiner (1976)] Delfiner, P.
Linear estimation of nonstationary spatial phenomena.
In: Advanced Geostatistics in the Mining Industry,
M. Guarascio, M. David and C. Huijbregts, eds. Reidel, Dordrecht, ppg
49-68, 1976
- [Dubrule (1983)] Dubrule, O.
Two methods with Different Objectives: Spline and Kriging
Mathematical Geology, Vol. 15, ppg 245-257, 1983

- [Egbert en Lettenmaier (1986)] Egbert, G.O. en Lettenmaier, D.P.
Stochastic modeling of the space-time-structure of atmospheric chemical decomposition
Water Resources Research, Vol. 22, pgg 165-179, 1986
- [van Egmond en Onderdelinden (1981)] van Egmond, N.D., en Onderdelinden, D.
Objective analysis of air pollution monitoring network data; spatial interpolation and network density
Atmospheric Environment, Vol. 15, pgg 1035-1046, 1981
- [Gandin (1965)] Gandin, L.S.
Objective analysis of meteorological fields. Translated by Israel Program for Scientific Translations, Jerusalem, 1965
- [Journel en Huijbregts (1978)] Journel, A.G. en Huijbregts, Chr.
Mining geostatistics, Academic Press, New York, 1978
- [Matheron (1973)] Matheron, G.
The intrinsic Random functions and their applications.
Advances in Applied Probability, Vol. 5, pgg 439-468, 1973
- [Pebesma (1992)] Pebesma, E.J.
Een landelijk beeld van grondwaterkwaliteitsvariabelen
deel1: een onderzoek naar interpolatiemethoden
RIVM-rapport 721801 002
- [Ripley (1981)] Ripley, B.D.,
Spatial Statistics
John Wiley and Sons, New York, 1981
- [Rodriguez-Iturbe en Mejia (1974)] Rodriguez-Iturbe, I. en Mejia, J.M.
The design of rainfall networks in time and space
Water Resources Research, Vol. 10, pgg 713-728, 1974
- [Stein (1988)] Stein, A., Van Dooremolen, W., Bouma, J. en Bregt, A.K.
Cokriging point data on moisture deficit
Soil Society of America Journal, Vol. 52, pgg 1418-1423, 1988
opgenomen in het proefschrift A. Stein