

**RUIMTELIJKE STATISTIEK
VAN BODEM EN WATER**



COMMISSIE VOOR HYDROLOGISCH ONDERZOEK TNO
RAPPORTEN EN NOTA'S No. 24

**RUIMTELIJKE STATISTIEK
VAN BODEM EN WATER**

CIP DATA

Statistiek

Ruimtelijke statistiek van bodem en water. (Auteurs: P.A. Burrough et al.; samensteller: J.C. Hooghart), Delft; Commissie voor Hydrologisch Onderzoek TNO. - ill. - (Rapporten en Nota's/Commissie voor Hydrologisch Onderzoek TNO, no. 24).

Rapport met de inleidingen van de lezingendag in De Reehorst te Ede, op 24 januari 1991.

Met literatuuropgave

ISBN 90-6743-185-0

SISO 631.2 UDC 556.3

Trefwoord: statistiek, hydrologie

Copyright © NEDERLANDSE ORGANISATIE VOOR TOEGEPAST
NATUURWETENSCHAPPELIJK ONDERZOEK, 1991

**RUIMTELIJKE STATISTIEK
VAN BODEM EN WATER**



Lezingendag, 24 januari 1991,
De Reehorst te Ede

COMMISSIE VOOR HYDROLOGISCH ONDERZOEK TNO

Delft, 1991

RAPPORTEN EN NOTA'S No. 24



VOORWOORD

De lezingendag, met als thema "Ruimtelijke statistiek van bodem en water" op 24 januari 1991 is een vervolg op de 6e CHO-studiebijeenkomst op 22 oktober 1986 en de lezingendag "Tijdreeksen in bodem en water" die op 26 januari 1989 is gehouden. De lezingdagen werden georganiseerd door de NRLO-werkgroep Ruimtelijke variabiliteit van bodem en water en de Studiegroep Statistiek in de Hydrologie van de CHO-TNO.

Deze twee samenwerkende groepen hebben zich als taak gesteld het onderzoek naar, en de toepassing van statistische methoden in de bodemkunde en de hydrologie te stimuleren. Daartoe worden regelmatig lezingen/discussiemiddagen georganiseerd, waarop specialisten kennis en ervaringen kunnen uitwisselen. Daarnaast organiseren ze lezingendagen met een thema, zoals "Ruimtelijke statistiek in bodem en water" met het doel kennis en inzichten naar een groter publiek te presenteren.

De aanleiding voor het organiseren van deze bijeenkomst en tot het uitgeven van een publikatie over "ruimtelijke statistiek" is dat er ten aanzien van dit thema bij diverse instanties vele vragen liggen. Bovendien is het vakgebied sterk in beweging, waardoor er sinds 1986 nieuwe ontwikkelingen te melden zijn.

Middels deze publikatie kunnen naast de deelnemers aan de lezingendag ook andere geïnteresseerden kennis nemen van de gepresenteerde lezingen.

Delft, januari 1991



INHOUDSOPGAVE

	PAG.
AUTEURS	1
1 DE RUIMTELIJKE VARIABILITEIT VAN BODEM EN WATER	3
A. Stein	
Samenvatting	3
1 Inleiding	3
2 Twee hoofdthema's	4
3 Omgevingsstudies	7
4 Statistiek	9
5 (Co)kriging als een regressie procedure	10
6 Toekomst	14
Literatuur	15
2 KLASSIEKE STEEKPROEFTHEORIE IN DE RUIMTELIJKE STATISTIEK	17
J.J. de Gruijter	
Samenvatting	17
1 Inleiding	17
2 Design-based en model-based benadering in een voorbeeld	20
3 Conclusies	27
Literatuur	28
3 EEN SEQUENTIËLE BEMONSTERINGSPROCEDURE VOOR HET SCHATTEN VAN HET GEMIDDELDE VAN EEN RUIMTELIJK VARIËREND KENMERK	31
P.A. Finke	
Samenvatting	31
1 Inleiding	31
2 Statistische achtergrond	33
3 Bepaling van de N-mineraal inhoud van een landbouwperceel	37

	voor het verkrijgen van een betrouwbaar bemestingsadvies	
4	Conclusies	42
	Literatuur	43
4	ONTWERP VAN HET MEETNET VOOR LUCHTVERONTREINIGING	45
	D. Onderdelinden	
	Samenvatting	45
	1 Inleiding	45
	2 Uitgangspunten	46
	3 Statistische eigenschappen	47
	4 Ontwerp meetnet	53
	5 Discussie	53
	Literatuur	54
5	DE FUNDAMENTEN VAN KRIGING EN CO-KRIGING	55
	P.R. Defize	
	Samenvatting	55
	1 Inleiding	56
	2 Ruimtelijke correlatie	56
	3 Formules	58
	4 Vergelijking Kriging versus Co-Kriging	61
	5 Conclusie	71
	Literatuur	72
6	HET KARTEREN VAN BODEMVERONTREINIGING IN HET GEULDAL DOOR CO-KRIGING MET AANVULLENDE HOOGTE- INFORMATIE	73
	H. Leenaers	
	Samenvatting	73
	1 Inleiding	73
	2 Het stroomgebied van de Geul	74
	3 Werkwijze	77
	4 Het ruimtelijk verband tussen Zn-gehalte en hoogteligging	80
	5 Discussie	84

6 Conclusies	86
Literatuur	87
7 KALMAN FILTERING BIJ RUIMTELIJKE INTERPOLATIE VAN GRONDWATER STIJGHOOGTEN	89
F.C. van Geer	
Samenvatting	89
1 Introductie	90
2 Uitgangspunten	91
3 Rekenschema	93
4 Toepassing	98
5 Discussie	101
Literatuur	103
8 COVARIANTIESTRUCTUUR VAN KAARTEN	105
J.V. Witter, A.C. van Eijnsbergen en M.A.J. van Montfort	
Samenvatting	105
1 Inleiding	106
2 Dispersie variantie	107
3 Interpolatie en dispersie variantie	110
4 Conditionele simulatie	113
5 Temporele covariantiestructuur	114
6 Voorbeeld	115
7 Conclusies	121
Literatuur	122
Appendix I : Turning bands methode	123
Appendix II : Afleiding van vergelijking 12	125
Appendix III : Verband tussen DP^2 en s^2	127
9. DE ROL VAN DE GEOSTATISTIEK IN EEN INTELLIGENT GIS	129
P.A. Burrough	
Samenvatting	129
1 Inleiding	130

2	Geografische Informatiesystemen	131
3	Modelleren met GIS	133
4	De kwaliteit van modelresultaten	136
5	Wat ontbreekt bij de huidige GIS?	137
6	De ruimtelijke statistiek als component van een intelligent GIS	138
7	Kennissystemen	143
	Dankwoord	145
	Literatuur	145

OVERZICHT RAPPORTEN EN NOTA'S

AUTEURS

- P.A. Burrough Rijksuniversiteit van Utrecht
 Geografisch Instituut, Postbus 80115
 3508 TC UTRECHT
- P.R. Defize Instituut voor Toegepaste Informatica TNO
 Postbus 214
 2600 AE DELFT
- A.C. van Eijnsbergen Landbouwwuniversiteit Wageningen
 Vakgroep Wiskunde, Postbus 8003
 6700 EB WAGENINGEN
- P.A. Finke Landbouwwuniversiteit Wageningen
 Vakgroep Bodemkunde en Geologie, Postbus 37
 6700 AA WAGENINGEN
- F.C. van Geer Dienst Grondwaterverkenning TNO
 Postbus 285
 2600 AG DELFT
- J.J. de Gruijter Staring Centrum
 Postbus 125
 6700 AC WAGENINGEN
- H. Leenaers CSO, Adviesbureau voor Milieuonderzoek
 Postbus 30
 3734 ZG DEN DOLDER

M.A.J. Montfort	Landbouwniversiteit Wageningen, Vakgroep Wiskunde Postbus 8003 6700 EB WAGENINGEN
D. Onderdelinden	Rijksinstituut voor Volksgezondheid en Milieuhygiëne Postbus 1 3720 BA BILTHOVEN
A. Stein	Landbouwniversiteit Wageningen, Vakgroep Bodemkunde en Geologie, Postbus 37 6700 AA WAGENINGEN
J.V. Witter	Hoogheemraadschap West-Brabant Postbus 2212 4800 CE BREDA

DE RUIMTELIJKE VARIABILITEIT VAN BODEM EN WATER

A. Stein

ABSTRACT

Waarnemingen in de ruimte kunnen met behulp van statistische methoden onderzocht worden. Een belangrijk gegeven betreft de samenhang tussen de waarnemingen, die beschreven wordt met behulp van covariantie functies, of met semivariogrammen. In dit artikel wordt ingegaan op de verschillende toepassingsvelden waarin aandacht besteed wordt aan ruimtelijke variabiliteit. Vervolgens wordt aangetoond hoe Cokriging, d.w.z. het uitvoeren van voorspellingen naar niet bezochte lokaties, beschreven kan worden met behulp van een regressie model. De activiteit van het voorspellen is een bindend element in verschillende studies naar het opzetten van optimale bemonsteringsschema's en studies naar interpolatie. Deze studie wordt afgesloten met een korte visie op de toekomst.

1 INLEIDING

Sinds het begin van de zeventiger jaren wordt in vele omgevingsstudies meer en meer aandacht besteed aan een kwantitatieve analyse van kenmerken ruimtelijke waargenomen zijn. Verwerking van dergelijke gegevens op een statistisch verantwoorde manier heeft in de afgelopen decennia zijn beslag gekregen. We zouden dit kunnen samenvatten als 'Ruimtelijke Statistiek'. Hoewel de ruimtelijke statistiek als zodanig veel ouder is, en we veel van het werk van Fisher en anderen hier onder kunnen rangschikken is het vandaag de bedoeling dat met name aandacht wordt besteed aan datgene dat populair wel de

'Geostatistiek' genoemd wordt.

Ruimtelijke statistiek heeft raakvlakken met vele vakgebieden van de statistiek, terwijl toepassingen te vinden zijn in talloze omgevingsstudies. Aan verschillende facetten van de ruimtelijke statistiek zullen we vandaag aandacht besteden.

Het is verheugend dat er opnieuw sprake is van een studiedag naar de ruimtelijke variabiliteit van bodem en water (zie ook CHO-TNO, 1987). Het toont aan dat de belangstelling voor het bestuderen van de ruimtelijke variabiliteit onverminderd groot is.

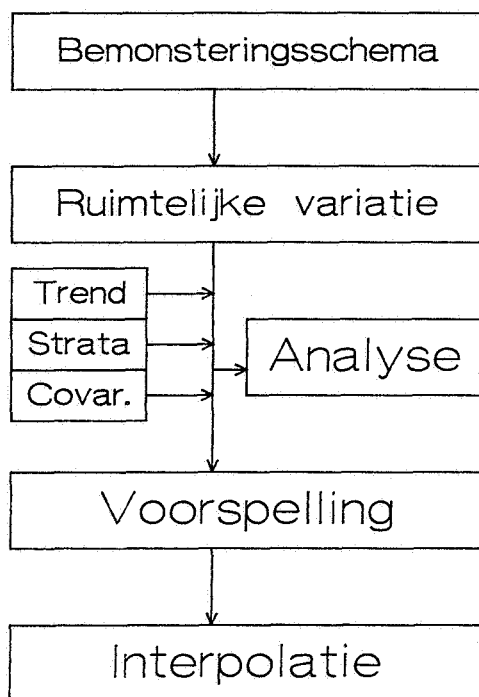
2 TWEE HOOFDTHEMA'S

In het algemeen zijn er bij de analyse van ruimtelijke variabiliteit twee hoofdthema's aan te geven, die complementair van karakter zijn (Fig. 1). We zullen deze hoofdthema's zowel tegenkomen bij het analyseren van het gebruik van ruimtelijke statistiek in omgevingsstudies, als gedurende de presentaties die in de loop van deze dag de revue passeren.

a. Bemonsteringsschema's

Bij het verzamelen van waarnemingen is een steeds terugkerende vraag "hoeveel" waarnemingen "waar" moeten worden uitgevoerd om een statistisch verantwoorde uitspraak te kunnen krijgen. Een antwoord op deze vraag is niet eenduidig te geven. Zo zullen bij het schatten van een semivariogram andere overwegingen een rol spelen dan bij het uitvoeren van ruimtelijke interpolatie. En weer andere overwegingen spelen een rol indien een statistische proefopzet wordt vastgelegd die de invloeden van parameters op bijv. de oogst moet achterhalen (De Gruijter, 1991). Met name de vraag naar het "aantal" waarnemingen is, gezien de daaraan gekoppelde economische consequenties, van groot belang. Een belangrijke operationele procedure wordt gevonden in het sequentieel samplen van ruimtelijke gegevens (Stein et al., 1989; Finke, 1991).

Deze vragen spelen in verschillende gebieden waarin de ruimtelijke statistiek zijn toepassingen vindt. Een belangrijke rol vinden we met name bij studies waarin aandacht wordt besteed aan het ontwerpen van 'optimale meetnetten' (Onderdelinden, 1991). Sommige oplossingen zijn gebaseerd op Kalman filtering (Van Geer, 1991). Maar ook in bodemkundige studies, waar al dan niet sprake is van ruimtelijke afhankelijkheid tussen de



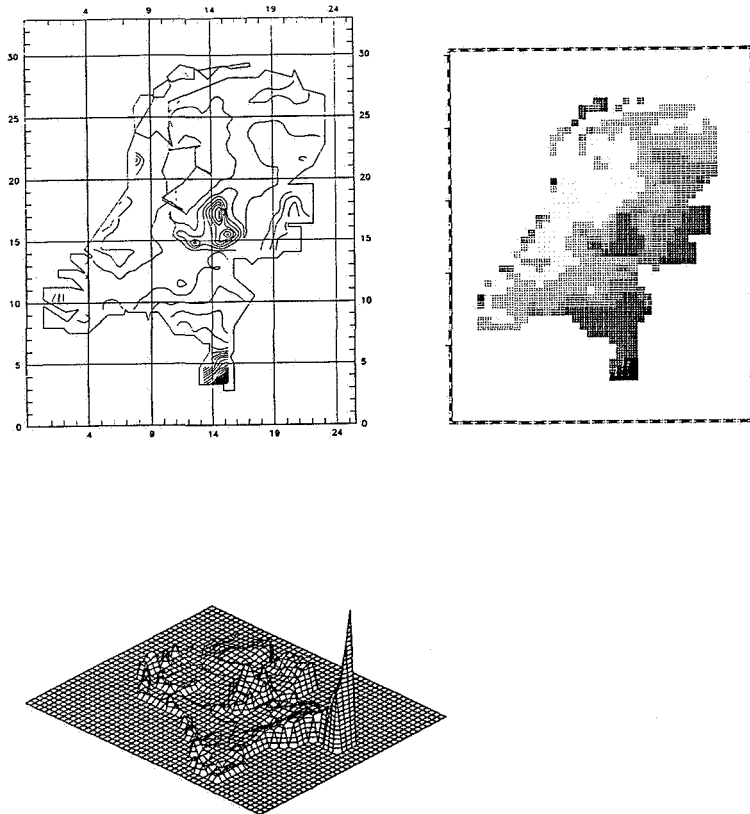
Figuur 1 Algemene procedure om te komen van puntwaarnemingen tot uitspraken voor vlakken

waarnemingen, keert deze vraag regelmatig terug.

b. Interpolatie

Het behoort tot het domein van de ruimtelijke statistiek om zich bezig te houden met ruimtelijke interpolatie. Het doen van uitspraken van punt naar vlak, met de daaraan gekoppelde onzekerheden, zijn van oudsher de vragen waar de ruimtelijke statistiek zich mee bezig heeft gehouden (Defize, 1991). Hieronder versta in de eerste plaats het schatten van de ruimtelijke structuur functies, al dan niet in de aanwezigheid van trends. De volgende stap betreft het uitvoeren van een voorspelling in een niet bezochte locatie. Door

deze voorspellingen uit te voeren in een groot aantal grid punten hebben we te maken met we van ruimtelijke interpolatie (Leenaers, 1991). Een dergelijk resultaat kan met behulp van een Geografisch Informatie Systeem worden gevisualiseerd (Burrough, 1991). Een voorbeeld van een dergelijke interpolatie is gegeven in Fig. 2.



Figuur 2 De hoogtekaart van Nederland in verschillende perspectieven

Belangrijke ontwikkelingen hebben zich de afgelopen jaren afgespeeld in dit gebied. Ik wil graag op deze plaats de ontwikkeling van Cokringing noemen en het gecombineerde gebruik van statistische interpolatie en choropleth kaarten (Stein et al., 1988 a, b). Een nieuwe ontwikkeling betreft interpolatie in de zin van simulatie van stochastische velden. Deze ontwikkeling kan waardevolle bijdragen gaan leveren bij het uitvoeren van modelberekeningen op niet bezochte lokaties, waarbij de onzekerheid van de invoer variabelen wordt gekwantificeerd (Witter et.al. 1991).

3 OMGEVINGSSTUDIES

In een aantal verschillende richtingen is inmiddels het gebruik van ruimtelijke statistiek doorgedrongen. In sommige disciplines is sprake van een zekere speerpunt functie, in andere disciplines is sprake van een wat meer geleidelijke ontwikkeling. Zonder in alle details te willen treden is het zinvol een volgend overzicht te geven.

a. Geologie en Mijnbouw

De ontwikkeling van de ruimtelijke statistiek is gestart vanuit de mijnbouw (Krige, 1954; Matheron, 1963). Gezien de grote economische belangen is het niet toevallig dat de eerste publikaties met name geënt zijn op vragen vanuit de mijnbouw. Oorspronkelijk waren het vooral de vragen vanuit de goudmijnen die de meeste aandacht kregen (Journel en Huijbregts, 1978). Dit verklaart onder andere een aantal termen die in de geostatistiek gehanteerd worden, zoals het Nugget Effect. Nog steeds vindt een snelle ontwikkeling plaats die met name gevoed wordt vanuit vragen van de olie industrie. Het is evident dat het gebruik van de statistische methoden uitstijgt boven de aard van het mineraal.

b. Bodem

In bodemkundige studies het is sinds het begin van de jaren tachtig gebruikelijk dat statistische hulpmiddelen gebruikt worden bij het uitvoeren van met name interpolatieproblemen (zie, bv. McBratney en Webster, 1981). Het kwantificeren van de onzekerheden die daaraan gekoppeld zijn is in veel studies cruciaal. De belangrijkste reden voor deze ontwikkeling is dat het in de meeste bodemkundige studies onmogelijk is om een overzicht te krijgen van de variatie van een kenmerk, in tegenstelling tot bijvoorbeeld hoogte gegevens. Bovendien is in de bodemkunde altijd sterk de behoefte gevoeld om een koppeling te leggen tussen punten en vlakken: het gebruik van kaarten vormt een cruciaal hulpmiddel bij het interpreteren van variatie in de bodem.

c. Hydrologie en hydraulica

Van groot belang in de ontwikkeling van methoden en modellen in de ruimtelijke statistiek is het gebruik binnen de hydraulica en de hydrologie. Het gebruik van statistische methoden

is daar als regel gecompliceerder, en bevat in zekere zin een extra uitdaging, omdat er naast ruimtelijke variatie zowel sprake is van variatie in de tijd, als van informatie die een deterministisch karakter heeft. Metingen kennen als regel herhalingen in de tijd, iets dat zich in bodem, geologie en mijnbouw niet voordoet. Om die reden is men ook in staat om een groter accent te leggen op analyse en beschrijving van de deterministische component. Wat dit laatste betreft kunnen we met name denken aan het gebruik van Darcy's wet of een andere differentiaal vergelijking.

d. Meteorologie

In de meteorologie zijn verschillende studies in het verleden uitgevoerd (o.a. Witter, 1985; Seo et al., 1989). De eerste studie behandelde de heterogeniteit van neerslaggegevens in Nederland, terwijl de tweede studie betrekking had op het gebruik van Cokriging van regenvalgegevens met radarmetingen. We merken hierbij op, dat deze tweede studie een interessante combinatie te zien geeft van eenvoudige, snel en overdadig beschikbare gegevens met een betrekkelijk kostbare variabele die voorspeld moet worden.

e. Overig

Onder de overige disciplines waarin, soms meer, soms minder sprake is van het gebruik van geostatistische methoden wil ik met name de vakgebieden Vegetatiekunde, Bosbouw, Sociologie, Geografie en Archeologie noemen. In al deze, uiteenlopende, vakgebieden is sprake van gegevens die een ruimtelijke verspreiding kennen. Het koppelen van deze gegevens aan coördinaten geeft daarmee de brug naar het gebruik van ruimtelijk statistische methoden. Het is mijn indruk dat er zich in deze gebieden voldoende belangstelling bevindt, en dat men de oplossingsmethoden die worden aangereikt door de ruimtelijke statistiek onderkent, maar dat tot op heden de accenten van onderzoek in een andere richting wijzen. Enkele publikaties vanuit deze vakgebieden hebben echter de aandacht getrokken. Het is mijn stellige overtuiging dat er binnen enkele van deze vakgebieden zich de komende jaren een duidelijker profilering met betrekking tot het gebruik van ruimtelijk statistische analyse technieken zal gaan afspelen.

4 STATISTIEK

In veel omgevingsstudies weten we dat waarnemingen die dicht bij elkaar liggen sterker op elkaar lijken dan waarnemingen die verder van elkaar verwijderd zijn. Dit gaat op als we kijken naar bodemgegevens, mijnbouwgegevens, neerslag-gegevens, etc. Een belangrijke rol wordt toegekend aan de interpretatie van de waargenomen correlatiestructuur. Deze structuur kan aanleiding geven voor beschouwing over ontstaansgeschiedenis, sedimentatiepatronen, landgebruik, etc. De correlatiestructuur dient als uitgangspunt van de statistiek zoals we die vandaag zullen beschouwen.

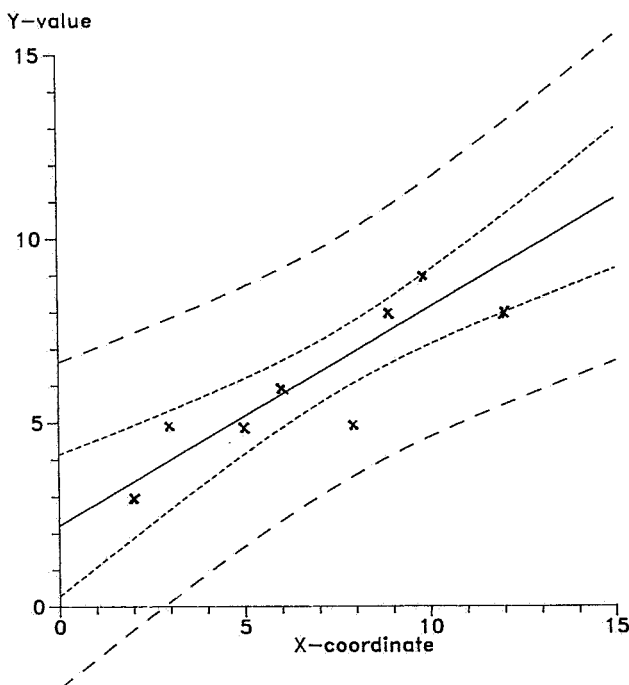
Centraal punt in alle studies is gelegen in de vraag hoe men omgaat met gegevens die ruimtelijk gecorreleerd zijn. Het is met name het "voorspellende" karakter hierin dat ons interesseert. Dit voorspellende karakter komt tot uiting in zijn meest simpele vorm als we een kaart willen vervaardigen van een kenmerk, en daarbij tevens willen dat die kaart, gegeven het waarnemingsmateriaal, zo goed mogelijk, d.w.z. optimaal, is. Tenslotte willen we ook graag weten hoe goed deze kaart voorspelt, en of zondig nieuwe waarnemingen moeten worden uitgevoerd. We spreken in dit geval van Kriging (Journel & Huijbregts, 1978; Corsten, 1989).

In een meer geavanceerde vorm wordt nieuwe informatie bij het uitvoeren van deze voorspellingen betrokken. Zo kunnen betrekkelijk dure waarnemingen hoog gecorreleerd zijn met goedkope waarnemingen. Als we deze covariabelen bij het uitvoeren van de voorspellingen gaan betrekken, spreken we van "Cokriging" (Stein en Corsten, 1990). Soms is er al een bodemkaart of een grondwaterkaart of een geologische kaart vervaardigd in het verleden, die mogelijk het uitvoeren van de voorspellingen verder kan verbeteren. We spreken dan van "stratificatie" van het studiegebied (Stein et al., 1988 b).

Een cruciaal punt betreft het benodigde aantal waarnemingen. Deze vragen komen naar voren op het moment dat een meetnet moet worden opgezet, of op het moment dat een bodeminventarisatie moet worden uitgevoerd. Hierbij kan gedacht worden aan sequentiële procedures, maar ook aan grids, random plans, etc. Hierbij komt nadrukkelijk een raakvlak naar voren tussen de ruimtelijke variabiliteitsstudies en het aloude ontwikkelen van sampling designs. Het is gebleken dat een genestelde procedure in een aantal opzichten bepaald niet is aan te bevelen (Corsten en Stein, submitted).

5 (CO)KRIGING ALS EEN REGRESSIE PROCEDURE

Bij het beschrijven van methoden in de ruimtelijke statistiek zullen we ons in eerste instantie concentreren op het begrip 'voorspelling'. Bij het uitvoeren van een voorspelling dienen we in acht te nemen dat we een stochastisch effect proberen te schatten, d.w.z. we voorspellen een toekomstige waarneming. In tijdreeksen is dat een gangbaar verschijnsel: op grond van waarnemingen in het verleden wordt een gooi gedaan naar een waarneming in de toekomst; echter, ook het 'voorspellen' van waarnemingen in het verleden en van tussenliggende waarnemingen is een legitieme activiteit. Het verschijnsel wordt het eenvoudigst beschreven aan de hand van lineaire regressie. Bij gewone lineaire regressie met ongecorrleerde waarnemingen, alle met dezelfde variantie, zijn de waarde van de voorspelling en van de beste schatting van de verwachting in een nieuw punt met regressorwaarden bevat in de vector x_0 , gelijk aan elkaar (Fig. 3).



Figuur 3 Het uitvoeren van een voorspelling in een niet-bezochte lokatie is onzekerder dan het schatten van de verwachting - vgl. brede streep met smalle streep; de waarden van voorspelling en schatting van de verwachting zijn hetzelfde (doorgetrokken lijn)

Echter, de variantie van de schatter van de verwachting is gelijk aan de restvariantie, vermenigvuldigd met $x_0'(X'X)^{-1}x_0$, terwijl voor de voorspeller, deze factor vervangen moet worden door $1+x_0'(X'X)^{-1}x_0$. Hierin is X de matrix waar ieder van de, zeg, k kolommen een n -vector is die de waarnemingen aan de regressor bevat in de n waarnemingspunten. Er kunnen aanzienlijke verschillen optreden.

Voor het beschrijven van waarnemingen in ruimtelijke zin is het gebruikelijk om te spreken van een 'geregionaliseerde variabele'. De term 'geregionaliseerde variabele', of 'geovariabele' wordt gebruikt om aan te geven dat een variabele waarden aanneemt in een 1, 2 of 3-dimensionale ruimte. Meestal wordt dit genoteerd door $y(x)$: het (stochastische) kenmerk y hangt af van de locatie x . De geregionaliseerde variabele waarvoor voorspellingen moeten worden uitgevoerd is de zg. doel-variabele, of de predictand. Deze is waargenomen in een vaste configuratie van, zeg, n_1 punten. Bovendien zijn er n_2 waarnemingen aan een andere variabele beschikbaar, in mogelijk andere punten. We beperken ons hier tot het gebruik van 1 covariabele. Het gebruik van meerdere covariabelen is een eenvoudige generalisatie, terwijl het ontbreken van een covariabele een vereenvoudiging is.

We zoeken een voorspeller \hat{t} voor de waarde y_0 van de predictand in een punt met gegeven coördinaten, gebruik makend van de n_1 waarnemingen aan de predictand, die bevat zijn in de stochastische n_1 -vector y_1 . Bovendien maken we gebruik van de n_2 waarnemingen aan de covariabele die bevat zijn in de stochastische n_2 -vector y_2 . Het totale aantal waarnemingen is n . In de praktijk zullen de waarnemingspunten van predictand en covariabele elkaar gedeeltelijk overlappen, maar in de formulering van het voorspellingsprobleem is dat niet cruciaal. Duidelijk is dat we te maken hebben met een configuratie van n_1+n_2+1 waarnemingspunten.

We veronderstellen dat de verwachting van ieder element van y_1 en y_2 een polynoom is in de coördinaten van de 1, 2 of 3-dimensionale ruimte waarin de waarnemingspunten, zich bevinden, met betrekking tot een willekeurige oorsprong:

$$E y_1 = X_1 \beta_1 \quad (1)$$

$$E y_2 = X_2 \beta_2 \quad (2)$$

Zowel $E y_1$ als $E y_2$ zijn lineaire combinaties van de regressorvectoren met tot op heden onbekende parameters β_1 en β_2 . De matrix X_1 bestaat uit n_1 rijen en p kolommen, waarin p afhangt van de graad van de polynomiale verwachting (de "trend", of de "drift") en de

dimensie van het gebied waarin de waarnemingen zijn uitgevoerd: bij een lineaire trend bestaat iedere rij van X_1 uit 1, ξ_1 en ξ_2 , waarin ξ_1 en ξ_2 de coördinaten van een waarnemingspunt van de predictand weergeven; als de graad van de trend gelijk is aan 2, d.w.z. als er sprake is van een kwadratische trend, dan bestaat iedere rij van X_1 uit 1, ξ_1 , ξ_2 , ξ_1^2 , $\xi_1\xi_2$, ξ_2^2 . Dit is analoog voor de matrix X_2 , die bestaat uit n rijen en q kolommen, waarbij iedere rij bestaat uit de q monomen van de waarden van de coördinaten tot de graad van de trend van een waarnemingspunt van de covariabele. In het algemeen kunnen p en q van elkaar verschillen. Het gebruik van de term 'Universal' bij Kriging en Cokriging geeft aan dat er sprake is, of kan zijn, van trend termen.

Het is duidelijk dat $E\mathbf{y}_0 = \mathbf{x}_0'\beta_1$, waarin de vector \mathbf{x}_0' bestaat uit een rij van p monomen van de coördinaten van het punt waar een voorspelling moet worden uitgevoerd. Deze monomen hebben dezelfde structuur als de p regressoren in de matrix X_1 .

In tegenstelling tot trend surface analyse, waar verondersteld wordt dat de waarnemingen onderling onafhankelijk zijn, gebruiken bij Kriging en Cokriging expliciet de afhankelijkheid van de waarnemingen. Deze wordt gemodelleerd met behulp van de covariantie tussen de waarnemingen onderling en de covariantie tussen de waarnemingen en de voorspellokatie.

We veronderstellen dat de afhankelijkheidsstructuur van de vector (y_1', y_2') wordt gegeven door de symmetrische covariantie matrix C, ter grootte $n_1 + n_2$,
$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$$

C_{11} is de covariantie matrix van de elementen van y_1 , C_{22} van de elementen van y_2 en C_{12} ($= C_{21}'$) is de covariantie matrix tussen de elementen van y_1 en de elementen van y_2 . Met betrekking tot y_0 geven c_{01} , c_{02} en c_{00} respectievelijk de covarianties tussen y_0 en de elementen van y_1 , y_2 en y_0 . Ieder element van C_{11} en van c_{01} en ook c_{00} , is gegeven als een isotrope functie $c_1(r)$, de zg. covariantie functie van de predictand, die enkel afhangt van de afstand r tussen een paar van waarnemingen. Evenzo veronderstellen we dat ieder element van C_{22} gegeven is als een isotrope functie $c_2(r)$, de zg. covariantie functie voor de co-variable, die enkel afhangt van de afstand r tussen waarnemingspunten van de covariabele, en dat ieder element van C_{12} en van c_{02} gegeven is als een isotrope functie $c_{12}(r)$ van de afstand r tussen een paar waarnemingspunten, één aan de predictand en een ander aan de covariabele, de zg. kruis-covariantiefunctie tussen predictand en co-variable. We zullen op deze plaats niet ingaan op het schatten van deze functies, of op positief-definietheids condities. Er zij opgemerkt dat onder een aantal condities het semivariogram equivalent is aan het tegenovergestelde van de covariantie functie.

We definiëren nu de volgende gepartitioneerde matrices en vectoren:

$$y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, X = \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix}, x_0^* = \begin{pmatrix} x_0 \\ 0 \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}. \quad (5)$$

Als we bovendien V definiëren als $V = (X'C^{-1}X)^{-1}$ vinden we dat de beste lineaire voorspeller, d.w.z. de voorspeller met minimale variantie van de voorspelfout, gelijk is aan

$$\hat{t} = x_0^{*\prime} \hat{\beta} + c_0' C^{-1} (y - X \hat{\beta}) \quad (6)$$

waarin $\hat{\beta} = VX'C^{-1}y$, de gegeneraliseerde kleinste kwadraten schatter voor β .

De voorspeller \hat{t} is de som enerzijds van de geschatte lokale verwachting van de predictand, en anderzijds van een lineaire combinatie van de waargenomen residuen, bevat in $y - X \hat{\beta}$, met de beste lineaire multivariate benadering van de predictand op grond van alle $m+n$ waarnemingen als coëfficiënt. Omdat $\hat{t} - x_0^{*\prime} \hat{\beta} = c_0' C^{-1} (y - X \hat{\beta})$, kunnen we deze procedure interpreteren als regressie van de residuen van \hat{t} met betrekking tot $x_0^{*\prime} \hat{\beta}$ op de residuen van de y 's met betrekking tot $X \hat{\beta}$. Op grond van deze uitdrukking is tevens een eenvoudige formule te vinden voor de gewichten λ in $\hat{t} = \lambda'y$:

$$\lambda = C^{-1} (c_0 + XVx_a) \quad (7)$$

De variantie van de voorspelfout is gelijk aan

$$\text{Var}(\hat{t} - y_0) = c_{00} - c_0' C^{-1} c_0 + x_a' V x_a \quad (8)$$

waar x_a is gedefinieerd als $x_a = x_0^* - XC^{-1}c_0$.

We merken op dat deze variantie vrij is van Lagrange multiplicatoren, in tegenstelling tot hetgeen vermeld staat in talrijke literatuur (Journal & Huijbregts, 1978). We kunnen de variantie als volgt interpreteren:

c_{00} is de variantie van de predictand;

$c_0' C^{-1} c_0$ is de reductie van de variantie als gevolg van het feit van de beste lineaire benadering door de andere waarnemingen, die we beschouwen als $m+n$ additionele multivariate kenmerken;

$x_a' V x_a = \text{Var}(x_a' \hat{\beta})$ is de variantie van de som van de geschatte verwachtingen van de predictand en van de covariabele, echter niet in de voorspellokatie, met regressoren bevat in

x_0^* , maar in een enigszins afwijkend punt met regressoren bevat in x_a .

In feite weerspiegelt de variantie van de voorspelfout de orthogonale ontbinding van de voorspelfout in respectievelijk $y_0 - c_0' C^{-1} y$ en $x_a' \hat{\beta}$.

De voorspeller \hat{t} is een exacte predictor: als een voorspelling wordt uitgevoerd in het i^e waarnemingspunt, dan is de vector c_0 gelijk aan de i^e kolom van C , zodat als de beste voorspelling de waarneming zelf wordt gegeven. In dit geval is de variantie van de voorspelfout gelijk aan 0.

Een van de aardige aspecten van de formulering zoals die hier wordt gegeven is dat als \hat{t} herhaaldelijk wordt berekend voor verschillende lokaties, dat alleen x_0 en c_0 herzien moeten worden, terwijl de rekenintensieve inversie van de matrix C slechts eenmaal hoeft te worden uitgevoerd.

Bij het ontbreken van een trend kunnen semivariogrammen gebruikt worden bij het uitvoeren van voorspellingen. Indien g_0 de vector is met semivariogram en cross-semivariogram tussen voorspellokatie en alle waarnemingslokaties, en G de matrix is met alle semi- en cross-semivariogrammen, analoog aan C , dan veranderen (6) en (8) in:

$$\hat{t} = \hat{\mu}_1 + g_0' G^{-1} (y - 1_n \hat{\mu}) \quad (9)$$

$$\text{Var}(\hat{t} - y_0) = g_0' G^{-1} g_0 - x_a' V x_a \quad (10)$$

met $x_a = x_0^* - XG^{-1}g_0$, $V = (X'G^{-1}X)^{-1}$ en $\hat{\mu} = VX'G^{-1}y$, waarbij $\hat{\mu}_1$ het eerste element is van $\hat{\mu}$. We merken op dat de matrix X een simpele structuur krijgt:

$X_1 = 1_{n_1}$ en $X_2 = 1_{n_2}$. Duidelijk is, dat een elegante en eenvoudige structuur wordt verkregen voor de belangrijkste voorspelformules. Voor meer details en generalisaties wordt verwezen naar Stein en Corsten (in press).

6 TOEKOMST

Er zijn verschillende gebieden aan te geven waar de komende periode ontwikkeling zal gaan plaats vinden. De ontwikkelingen in methodologische zin zullen vermoedelijk plaats vinden in de ruimte-tijd modellen. Deze worden op inmiddels toegepast in hydrologische en meteorologische studies, maar zullen ook in andere disciplines hun toepassingen gaan vinden. Onherroepelijk zal dit aanleiding gaan geven tot een verdere verfijning van de methodologie. Duidelijk is dat op het gebied van het kwantificeren en het interpreteren van

de ruimtelijke variabiliteit nog een wereld open staat.

Daarnaast is het evident dat, mede gezien de opkomst van Geografische Informatie Systemen, de drempel voor het gebruik van ruimtelijke statistiek aanzienlijk verlaagd is. Dit vergemakkelijkt het gebruik van dergelijke methoden binnen disciplines waar ze tot op heden niet of nauwelijks gebruikt zijn.

Het gebruik van gesimuleerde velden is bovendien aantrekkelijk om realistische invoergegevens voor modellen te krijgen op punten waar geen waarnemingen zijn uitgevoerd. Hiermee wordt modeluitvoer gekregen die de onzekerheden representeert die gekoppeld zijn aan de kenmerken zelf en aan de verspreiding van de waarnemingspunten in de ruimte.

LITERATUUR

- BURROUGH, P.A.; 1991. De rol van geostatistiek in intelligent GIS. CHO-TNO Rapporten en nota's 24.
- CHO-TNO.; 1987. Ruimtelijke variabiliteit van bodem en water. CHO-TNO Rapporten en nota's No. 18.
- CORSTEN, L.C.A.; 1989. Interpolation and optimal linear prediction. *Statistica Neerlandica*, 43: 69-84.
- CORSTEN, L.C.A., and STEIN, A.; 1991. Are nested sampling schemes recommendable for spatial semivariogram estimation? Subm. to JASA.
- DEFIZE, P.; 1991. De fundamente van Kriging en Cokriging. CHO-TNO Rapporten en nota's No. 24.
- FINKE, P.A.; 1991. Een sequentiële bemonsteringsprocedure voor het schatten van het gemiddelde van een ruimtelijk variërend kenmerk. CHO-TNO Rapporten en nota's No. 24.
- GEER, F.C. VAN; 1991. Kalman filtering bij ruimtelijke interpolatie van grondwaterstijghoogten. CHO-TNO Rapporten en nota's No. 24.
- DE GRUUTER, J.J.; 1991. Klassieke steekproeftheorie in de ruimtelijke statistiek. CHO-TNO Rapporten en nota's No. 24.
- JOURNAL, A.G. and HUIJBREGTS, C.J.; 1978. *Mining geostatistics*. Academic Press, London.
- KRIGE, D.G.; 1951. A statistical approach to some basic mine valuation problems on the Witwatersrand. *J. Chem., Metall. & Min. Soc. of S. Africa*, 52: 119-138.

- LEENAERS, H.; 1991. Het karteren van bodemverontreiniging in de het Geuldal door co-kriging met aanvullende hoogte-informatie. CHO-TNO Rapporten en nota's No. 24.
- MATHERON, G.; 1963. Principles of geostatistics. *Economic Geology*, 58: 1246-1266.
- MCBRATNEY, A.B., and WEBSTER, R.; 1983. Optimal interpolation and isarithmic mapping of soil properties. V. Co-regionalization and multiple sampling strategy. *J. Soil Sci.*, 34: 137-162.
- ONDERDELINDEN, D.; 1991. Ontwerp van meetnet voor luchtverontreiniging. CHO-TNO Rapporten en nota's No. 24.
- RIPLEY, B.D.; 1981. *Spatial Statistics*. Wiley, New York.
- SEO, D.J., KRAJEWSKI, W.F. and BOWLES, D.S.; 1990. Stochastic interpolation of rainfall data from rain gages and radar using Cokriging. 1. Design of Experiments. *Water Resour. Res.*, 26(3): 469-477.
- STEIN, A., BOUMA, J., DOOREMOLEN, W. VAN and BREGT, A.K.; 1988a. Cokriging point data on moisture deficits. *Soil Sci. Soc. Am. J.*, 52: 1418-1423.
- STEIN, A., HOOGERWERF, M.H. and BOUMA, J.; 1988b. Use of soil map delineations to improve (co-)kriging of point data on moisture deficits. *Geoderma*, 43: 163-177.
- STEIN, A., BOUMA, J., KROONENBERG, S.B. and COBBEN, S.; 1989. Sequential sampling to measure the infiltration rate within relatively homogeneous land units. *Catena*, 16: 91-100.
- STEIN, A. and CORSTEN, L.C.A.; 1990. Universal Kriging and Cokriging as a regression procedure. *Biometrics*, In press.
- WITTER, J.V.; 1984. Heterogeneity of dutch rainfall. Thesis, Agricultural University, Wageningen, The Netherlands.
- WITTER, J.V., EIJSBERGEN, A.C. VAN and MONTFORT, M.A.J. VAN; 1991. Covariantie structuur van kaarten. CHO-TNO Rapporten en nota's No. 24.

KLASSIEKE STEEKPROEFTHEORIE IN DE RUIMTELIJKE STATISTIEK

J.J. de Gruijter

SAMENVATTING

Een wijd verbreide gedachte onder geostatistici is dat de klassieke steekproeftheorie niet van toepassing is op ruimtelijke steekproefproblemen, omdat ruimtelijke gegevens in het algemeen afhankelijk zijn en de theorie eist dat ze onafhankelijk zijn. Echter, uit een vergelijking van de uitgangspunten van de klassieke steekproeftheorie met die van de geostatistische theorie blijkt dat dit een ongelukkig misverstand is. In het bijzonder geldt dat schattingen van ruimtelijke gemiddelden die zijn gebaseerd op klassieke steekproefopzetten juist minder vooronderstellingen vereisen, en daarom robuuster zijn dan die welke zijn gebaseerd op een geostatistisch model.

1 INLEIDING

"Classical techniques cannot be used because they are based on independence of sample data" (Yfantis et al., 1987). Deze en soortgelijke opmerkingen zijn keer op keer in de geostatistische literatuur verschenen (zie b.v. Russo and Bresler, 1981, Dahiya et al., 1985, Barnes, 1988). Veel geostatistici lijken dit nu als een elementair feit te beschouwen, dat niet meer steeds herhaald hoeft te worden. Het schijnt een soort geostatistisch paradigma te zijn geworden. Maar is het waar? Het antwoord op die vraag moet naar mijn mening ontkennend zijn, en dit heeft belangrijke praktische consequenties voor ruimtelijke steekproefname, zoals ik verderop zal laten zien.

Onafhankelijkheid is inderdaad een essentieel uitgangspunt in de klassieke schattings- en toetsingstheorie ontwikkeld door R. A. Fisher en anderen. Echter, in de klassieke *steekproef*theorie ontwikkeld door J. Neyman en anderen, worden géén vooronderstellingen over onafhankelijkheid gemaakt. Dit is als volgt toe te lichten.

Hedendaagse handboeken over steekproefname (b.v. Cassel et al., 1977, Krishnaiah and Rao, 1988) onderscheiden twee fundamenteel verschillende benaderingen die kunnen worden gevolgd bij schatting en toetsing omtrent ruimtelijke populatie-parameters. Särndal (1978) duidt deze aan als de "design-based" en de "model-based" benadering, en ik volg die terminologie. Synoniemen zijn, respectievelijk, "vaste populatie" en "superpopulatie" benadering. Zie ook Chauduri en Vos (1988).

In de design-based benadering wordt klassieke steekproeftheorie gebruikt. Het centrale begrip in die theorie is de populatie, gedefinieerd als de verzameling van alle eenheden die van belang zijn in een bepaald onderzoek. In ruimtelijke toepassingen is de populatie bijvoorbeeld een gebied A of, nauwkeuriger, de verzameling van alle mogelijke steekproeflokaties in het van belang zijnde gebied. In de design-based benadering wordt iedere eenheid (mogelijke steekproeflokatie) geacht een vaste, hoewel meestal onbekende waarde voor het te onderzoeken kenmerk te hebben. De steekproefname bestaat uit het selecteren van een deelverzameling van de eenheden van de populatie, waarna de waarden in die deelverzameling worden bepaald via waarneming. In het extreme geval van een steekproef die de gehele populatie omvat, en foutloze meting, zou er niets meer onbekend zijn. Afgezien van eventuele meetfouten, is de enige variatie die in deze benadering een rol speelt, die welke wordt veroorzaakt door de steekproefname. De wijze van schatten en toetsen wordt hier dan ook primair gebaseerd op de toegepaste steekproefopzet. Deze steekproefopzet, d.w.z. de wijze waarop de deelverzameling wordt geselecteerd, bepaalt namelijk voor elk willekeurig stel eenheden in de populatie de kans dat ze in de steekproef worden opgenomen. Hieruit zijn, zonder dat extra aannamen gemaakt hoeven te worden, valide schattings- en toetsingsprocedures af te leiden.

De model-based benadering, toegepast op ruimtelijke problemen, komt neer op het gebruik van geostatistische theorie, met name van hetgeen betrekking heeft op voorspellen van ruimtelijke gemiddelden. In deze benadering wordt de waarde van het kenmerk op een gegeven steekproeflokatie niet als vast, maar als stochastisch beschouwd. De verzameling

van waarden corresponderend met alle mogelijke steekproeflokaties in het te onderzoeken gebied vormen hier te zamen slechts één realisatie van een stochastisch proces. Van dit proces worden enkele aspecten verondersteld bekend te zijn, en deze vooronderstellingen worden geformaliseerd in een geostatistisch model. Dit model speelt de rol van wat in de steekproeftheorie een superpopulatie-model wordt genoemd. Indien, weer in het extreme geval, een steekproef het gehele gebied zou dekken, zou één realisatie compleet bekend zijn, maar er zou ook dan nog onzekerheid blijven bestaan over de parameters in het model. Typerend voor dit soort modellen is dat ze rekening houden met een zekere ruimtelijke structuur, wat het feit weerspiegelt dat metingen op dicht bij elkaar gelegen lokaties doorgaans waarden opleveren die minder van elkaar verschillen dan op verder van elkaar verwijderde lokaties.

In de model-based benadering hoeven de lokaties niet aselekt te worden bepaald; ze worden integendeel doorgaans juist gericht gekozen. De enige bron van stochasticiteit is in dat geval het gepostuleerde achterliggende model. Bij deze benadering wordt daarom de statistische verwerking (schatten, toetsen, voorspellen) primair gebaseerd op dat model. De aard van de stochasticiteit in de model-based benadering is dus fundamenteel verschillend van die in de design-based benadering waar deze, zoals we zagen, voortkomt uit een concreet fysisch proces van steekproefname. En dat laatste hebben we zelf in de hand. De design-based benadering vereist dus minder aannamen, is daarom robuuster, en kan om die reden de voorkeur verdienen.

Deze bijdrage, die grotendeels is ontleend aan De Gruijter en Ter Braak (1990), heeft tot doel het onderscheid tussen de genoemde benaderingen te verduidelijken. Als dit onderscheid niet duidelijk wordt gezien kan dat gemakkelijk tot misverstanden en drogredeneringen leiden. De materie wordt hier op vrij informele wijze besproken, waarbij gebruik wordt gemaakt van de terminologie en enkele van de belangrijkste begrippen van Cassel et al. (1977), die een theoretisch raamwerk hebben ontwikkeld dat beide benaderingen omvat. Als illustratie wordt steekproefname van punten in een vlak behandeld.

De beschouwing zal worden toegespitst op schatting van populatie-parameters zoals het ruimtelijk gemiddelde: het gaat dus niet over voorspelling van waarden op individuele lokaties of over het berekenen van niveaulijnen-kaarten. In de geostatistiek, bijvoorbeeld, kan het ruimtelijk gemiddelde worden bepaald via block-kriging met het gehele gebied als

één enkel blok.

2 DESIGN-BASED EN MODEL-BASED BENADERING IN EEN VOORBEELD

Om het verschil tussen de design-based en de model-based benadering verder te verduidelijken nemen we een denkbeeldig voorbeeld: het geval waarbij een steekproef van punten wordt genomen in een vlak. Stel dat een grootheid z wordt gemeten op n punten in een gebied A . De verkregen waarden worden gesymboliseerd met $z(x_i)$, waarbij x_i de vector van coördinaten voorstelt van het i de steekproefpunt. De waarden of de coördinaten (of beide) kunnen stochastische variabelen zijn. Om stochastische en deterministische grootheden van elkaar te onderscheiden worden de eerste met hoofdletters en de laatste met kleine letters aangeduid.

2.1 Design-based benadering

In de design-based benadering worden de n steekproefpunten aselekt gekozen, volgens een steekproefopzet p , en voor gegeven punten zijn de waarden vast. De betrokken variabelen kunnen dus worden geschreven als $z(X_i)$, waarin $i=1, \dots, n$. Zij zijn stochastisch omdat de bijbehorende lokaties (X_i) stochastisch zijn. Of $z(X_i)$ en $z(X_j)$ stochastisch onafhankelijk zijn of niet wordt volledig bepaald door de steekproefopzet en niet door de ruimtelijke variatie in A .

Dit simpele feit lijkt door sommige gebruikers van geostatistische technieken over het hoofd te worden gezien. In het bijzonder, als metingen worden gedaan op punten die aselekt en bovendien onafhankelijk van elkaar zijn gekozen, dan zijn de daarmee corresponderende stochastische variabelen altijd onderling onafhankelijk; niet bij wijze van aanname in een model maar in werkelijkheid. Het is het gevolg van het ontworpen waarnemingsproces. In het voorbeeld zou men dat proefondervindelijk kunnen laten zien door een enkelvoudige aselechte steekproef van omvang $n=2$ te trekken (d.w.z. door twee punten onafhankelijk van elkaar te trekken uit de uniforme verdeling over A), en dat een groot aantal keren te herhalen. Op die manier ontstaat een reeks tweetallen van waarden. Als de reeks lang genoeg is, zal blijken dat het spreidingsdiagram van de eerste tegen de tweede waarde geen

afhankelijkheid vertoont, ongeacht de ruimtelijke variatie in A^1 . In de design-based benadering hoeft men afhankelijkheid tussen variabelen dus niet te *negeren*: men kan het *vermijden*.

Stel dat de te schatten grootheid het ruimtelijk gemiddelde is:

$$m(A) = \frac{\int_A z(x) dx}{\int_A dx} \quad (1)$$

De gebruikelijke schatter van $m(A)$ in combinatie met enkelvoudige aselechte steekproefname (Engels: simple random sampling, verder aangeduid met srs) is het ongewogen steekproefgemiddelde,

$$T_{\text{um}} = \frac{1}{n} \sum_{i=1}^n z(X_i) \quad (2)$$

De steekproefstrategie (srs, T_{um}) wordt "p-zuiver" met betrekking tot het ruimtelijk gemiddelde genoemd, omdat de statistische verwachting van de schatter T_{um} over herhaalde steekproefname volgens de gegeven opzet srs gelijk is aan de doelparameter $m(A)$, d.w.z.

$$E_p(T_{\text{um}}) = m(A), \quad (3)$$

waarin E_p de statistische verwachting is over herhaalde steekproefname volgens de gegeven opzet. Ten gevolge van deze p-zuiverheid is de p-MSE van T_{um} , d.w.z. de Mean Squared Error onder opzet p, gelijk aan zijn p-variantie (zie b.v. Cochran, 1977):

$$E_p[T_{\text{um}} - m(A)]^2 = V_p(T_{\text{um}}) = v(A)/n, \quad (4)$$

waarin $v(A)$ de variantie is tussen punten in A , gedefinieerd als:

$$v(A) = \frac{\int_A [z(x) - m(A)]^2 dx}{\int_A dx} \quad (5)$$

¹Men hoeft dit experiment natuurlijk niet uit te voeren om het gestelde te bewijzen: het volgt rechtstreeks uit een stelling in de kansrekening die zegt dat als twee variabelen, b.v. X_i en X_j , onafhankelijk zijn, dan zijn ook functies daarvan, b.v. $z(X_i)$ en $z(X_j)$, onafhankelijk, welke de functie z ook is. Zie bijvoorbeeld stelling 2 in Ash (1970, p. 84) of stelling 6A in Parzen (1960, p. 295).

In de terminologie van de klassieke steekproeftheorie is de verzameling van alle lokaties in gebied A de te onderzoeken populatie. De lokaties worden de eenheden van deze (oneindige) populatie genoemd; het ruimtelijk gemiddelde $m(A)$ en de ruimtelijke variantie $v(A)$ zijn parameters van de populatie.

Om de efficiëntie te verhogen of het veldwerk te vergemakkelijken kunnen andere steekproefopzetten te hulp worden geroepen. De hier gegeven formules voor enkelvoudige aselechte steekproeven kunnen gemakkelijk worden gemodificeerd voor andere opzetten, die óók model vrije schatting en toetsing mogelijk maken, bijvoorbeeld gestratificeerde aselechte steekproeven (Cochran, 1977).

2.2 Model-based benadering

De model-based benadering gaat uit van een stochastische functie ξ , die wordt geacht een A overdekkend veld van stochastische waarden te genereren. Een steekproef op n vaste punten levert hier dus in onze notitie de stochastische variabelen $Z(x_i)$ op, met $i=1, \dots, n$. De hoofdletter Z benadrukt dat de waarden stochastisch zijn; de kleine letter x_i dat de lokaties vast zijn. Het bestaan en de mate van afhankelijkheid tussen $Z(x_i)$ en $Z(x_j)$ wordt bepaald door de stochastische functie ξ , die op zijn beurt voor een deel wordt gespecificeerd door het vooronderstelde geostatistische model.

Stel dat ξ in gebied A tweede-orde stationair is. Dat betekent dat $E_{\xi}Z(x)$ bestaat en niet afhangt van x , en dat de covariantie $\text{Cov}[Z(x), Z(x+h)]$ bestaat en alleen afhangt van h (zie Myers, 1989, voor een bespreking van deze en andere vormen van stationairiteit die in de geostatistiek worden gehanteerd):

$$E_{\xi}Z(x) = \mu \quad (6a)$$

$$\text{Cov}[Z(x), Z(x+h)] = E_{\xi}[Z(x) \cdot Z(x+h)] - \mu^2 = C(h) \quad (6b)$$

waarin E_{ξ} de statistische verwachting aangeeft over realisaties van ξ . Stel verder dat het model-gemiddelde μ onbekend is en dat de covariantie-functie $C(h)$ bekend is, met model-variantie $C(0)$. In de praktijk moet de covariantie-functie doorgaans uit de gegevens worden geschat, maar dat doet hier niet terzake.

In de model-based benadering kan het schatten van een centrale waarde worden gericht op ten minste twee doel-grootheden:

- het model-gemiddelde μ in Vergelijking (6a), of
- het ruimtelijk gemiddelde $M(A)$, gedefinieerd als:

$$M(A) = \frac{\int_A Z(x)dx}{\int_A dx} \quad (7)$$

Het model-gemiddelde is een vaste grootheid en wordt daarom, in statistische terminologie, geschat. Het ruimtelijk gemiddelde echter, zijnde het gemiddelde van een realisatie van ξ , is in deze benadering een stochastische variabele en wordt daarom voorspeld.

Het kwaliteitscriterium dat in de model-based benadering vaak wordt gehanteerd is de ξ -MSE, d.w.z. de Mean Squared Error over realisaties van ξ . Toegepast op een schatter $\hat{\mu}$ van het model-gemiddelde is dit criterium gedefinieerd als:

$$\xi\text{-MSE}(\hat{\mu}) = E_{\xi}(\hat{\mu} - \mu)^2 \quad (8)$$

Minimaliseren van dit criterium terwijl $\hat{\mu}$ wordt beperkt tot lineaire combinaties van de steekproefgegevens, levert de bekende ξ -BLU schatter $\hat{\mu} = \lambda_{\mu}^T Z_S$ waarin

$$\lambda_{\mu}^T = (1_n^T C_S^{-1} 1_n)^{-1} 1_n^T C_S^{-1} \quad (9)$$

Hierin zijn λ_{μ} , Z_S en 1_n respectievelijk de vector van optimale gewichten, de vector van Z -waarden op de steekproefpunten en de vector van enen (alle n -dimensionaal als de steekproefomvang n is), en C_S is de $n \times n$ matrix van covarianties tussen de steekproefpunten. De ξ -variantie van $\hat{\mu}$ is:

$$V_{\xi}(\hat{\mu}) = (1_n^T C_S^{-1} 1_n)^{-1} \quad (10)$$

De ξ -BLU voorspeller van $M(A)$ is evenzo een gewogen gemiddelde, $T_{wm} = \lambda_Z^T Z_S$, maar nu met gewichten

$$\lambda_Z = \lambda_{\mu} + (C_S^{-1} - \lambda_{\mu} 1_n^T C_S^{-1}) \bar{C}_{S,A} \quad (11)$$

waarin $\bar{C}_{S,A}$ de n-dimensionale vector is van gemiddelde covarianties tussen een steekproefpunt en alle punten in A. T_{wm} heeft variantie²

$$V_{\xi}(T_{wm}) = \bar{C}_{A,A} + \lambda_Z^T C_S \lambda_Z - 2\lambda_Z^T \bar{C}_{S,A} \quad (12)$$

waarin $\bar{C}_{A,A}$ de gemiddelde covariantie aangeeft tussen alle mogelijke puntenparen in A. Zowel $\bar{\mu}$ als T_{wm} zijn ξ -zuiver, omdat voor elk willekeurig stel steekproefpunten geldt dat

$$E_{\xi}(\bar{\mu} - \mu) = E_{\xi}[T_{wm} - \dot{M}(A)] = 0 \quad (13)$$

Indien men niet alleen de gewichten, gegeven de lokaties, maar ook de lokaties zelf wil optimaliseren met betrekking tot het ξ -MSE criterium, dan vloeit daar een gerichte steekproefname uit voort, i.p.v. een aselecte. De optimale opzet (p_{μ}) voor het schatten van μ zal in het algemeen verschillen van de optimale opzet (p_M) voor het voorspellen van $M(A)$. Het is van belang op te merken dat de strategie (p_M, T_{wm}) ξ -zuiver is maar niet p-zuiver, terwijl b.v. de strategieën (srs, T_{um}) en (srs, T_{wm}) beide zowel ξ -zuiver als p-zuiver zijn voor $M(A)$.

2.3 Discussie

Het besproken voorbeeld maakt enkele punten duidelijk die zeer relevant zijn voor de keuze van steekproefstrategieën.

2.3.1 Doelgrootheden

De model-based benadering kan worden gericht op twee soorten grootheden (in ons voorbeeld het model-gemiddelde en het ruimtelijk gemiddelde), die verschillen qua interpretatie, optimale steekproefopzet en optimale schatter c.q. voorspeller, en die

² Corsten (1989) laat zien dat de bovenstaande expliciete uitdrukkingen voor de BLU schatter en voorspeller, en hun varianties, verkregen kunnen worden als gegeneraliseerde kleinste kwadraten oplossingen van de corresponderende regressie problemen. Daarmee vermijdt men het gebruik van irrelevante Lagrange multiplicators. Dezelfde uitdrukkingen kunnen óók worden verkregen door de Lagrange multiplicators te substitueren in de gebruikelijke formules in b.v. Matheron (1971, § 3-4) en Journel en Huijbregts (1978, Ch. V).

verschillende varianties hebben (Vergelijking (10) en (12)). Modelparameters kunnen van belang zijn in studies die op processen zijn gericht, terwijl ruimtelijke gemiddelden e.d. juist relevant zullen zijn bij inventariserend onderzoek. Het is in elk geval noodzakelijk zowel bij opzet als rapportering volledige duidelijkheid te betrachten over het type grootheid waar het onderzoek op is gericht.

2.3.2 Stochastische onafhankelijkheid

Stochastische afhankelijkheid of onafhankelijkheid is geen eigenschap van een populatie of van een gebied. Het kan wèl een eigenschap zijn van een stel variabelen, en wordt ofwel geïnduceerd door een steekproefopzet, ofwel geïmpliceerd door een model. Het lijkt erop dat verwarring en misverstanden over onafhankelijkheid zich in een bedenkelijke mate hebben voortgeplant door de geostatistische literatuur. Misleidende opmerkingen zijn in dit verband bijvoorbeeld:

"The conventional statistical approach to describe variability of soil hydraulic properties treats the observations of a given property as being statistically independent regardless of their spatial position." (Russo and Bresler, 1981)

"In Part I (Dahiya et al., 1984a), spatial variability of some nutrient constituents, viz. NO_3 , K, Mg and organic C, of a loess soil field was evaluated by applying classical statistical analysis (i.e. probability density function, mean and variance). An implicit assumption in that analysis is that the observations of a given soil property are independent of one another regardless of their location in the field." (Dahiya et al., 1985)

"The classic development of nonparametric tolerance intervals begins with an assumption of independent, identically distributed random variables. This is unrealistic for the geologic environment - in general, geologic site characterization data are not independent." (Barnes, 1988).

De waarheid is dat men de design-based benadering met zijn klassieke steekproeftheorie niet terzijde kan leggen omdat daarin ongeldige model-aannamen zouden worden gemaakt (er zijn daar immers geen model-aannamen), maar dat men de model-based benadering juist wèl zou kunnen afwijzen om die reden.

2.3.3 Kwaliteitscriteria

Het kiezen van een steekproefstrategie houdt primair het kiezen van een kwaliteitscriterium in (Särndal, 1978). Zo'n criterium kan betrekking hebben op variatie van uitkomsten door de steekproefname, op variatie vanuit een stochastisch model, of op beide. (Termen als "kleinste variantie" en "zuiverheid" zijn daarom zonder nadere kwalificatie niet eenduidig.) Indien, bijvoorbeeld, het ξ -MSE criterium wordt geaccepteerd, dan vloeit daar een model-based strategie uit voort. Als daarentegen p-zuiverheid noodzakelijk wordt geacht, dan zal een design-based strategie met een vorm van aselechte steekproefname de consequentie zijn.

Het valt te betwijfelen of degenen die gebruik maken van resultaten van geostatistische methoden zich doorgaans bewust zijn van het feit dat ξ -zuiverheid slechts een "interne" garantie biedt: hij geldt alleen binnen het aangenomen model en is als zodanig niet gerelateerd aan de werkelijkheid in het te onderzoeken gebied. Hij biedt dan ook geen bescherming tegen gevolgen van, bijvoorbeeld, het kiezen van lokaties met verwachte hoge dan wel lage waarden, of van alleen de best toegankelijke lokaties, of van lokaties in een regelmatig patroon dat interfereert met een eventuele cyclische variatie in het gebied. Dergelijke steekproeven zullen veelal misleidende resultaten opleveren, ondanks de bovengenoemde zuiverheidsgarantie. Dit lijkt in het bijzonder van belang in de context van regelgeving en controle met betrekking tot het natuurlijk milieu.

2.3.4 Toetsen

In het voorbeeld is ingegaan op schatten of voorspellen van gemiddelden. Het voorgaande geldt mutatis mutandis echter ook voor toetsen. Het gehele scala van klassieke en parameter vrije statistische toetsen (die vrijwel steeds uitgaan van onafhankelijkheid) is, afgezien van eventuele extra aannamen, direct en exact van toepassing op steekproeven waarvan de eenheden aselechte en onafhankelijk van elkaar zijn gekozen. Als deze voorzorgen niet zijn genomen is correcte toetsing veel problematischer (zie Borgman, 1988, voor een recente verkenning van deze materie). Een voorbeeld van toetsing met ruimtelijke gegevens volgens de design-based benadering is te vinden in Finke (1991).

3 CONCLUSIES

Als men de klassieke steekproeftheorie afwijst omdat daarin een ongeldige aanname van onafhankelijkheid gemaakt zou worden, verengt men de keuze van een steekproefstrategie op oneigenlijke gronden tot die van de model-based benadering. Het is duidelijk dat de design-based en de model-based benadering elk hun eigen voor- en nadelen hebben. Design-based strategieën zijn niet uitvoerbaar als het om een of andere reden niet mogelijk is een aselechte steekproef te nemen. De model-based benadering is niet uitvoerbaar als het door gebrek aan gegevens niet mogelijk is op betrouwbare wijze een model vast te stellen.

Als beide benaderingen uitvoerbaar zijn, betreft de te maken keuze primair het kwaliteitscriterium waarmee mogelijke strategieën worden beoordeeld. Het ene type criterium heeft betrekking op de variatie in resultaten als de steekproefname herhaald zou worden in hetzelfde gebied volgens dezelfde opzet, maar uiteraard met steeds andere configuraties van steekproefpunten. Dit type criterium voert tot design-based strategieën. Het andere type heeft óók betrekking op variatie die zou optreden onder herhaalde steekproefname, maar nu met dezelfde puntenconfiguratie en, meer hypothetisch, in verschillende gebieden, met een variabiliteit die voldoet aan de wetmatigheden van het gehanteerde model. Dit type criterium leidt tot model-based strategieën. Geostatistische modellen lijken een natuurlijke basis te vormen voor schattingsprocedures in situaties waarin een deel van het gebied niet toegankelijk is, of als de waarnemingen meettechnisch zijn gecensureerd, dan wel zijn vertekend door systematische fouten (zie voor dit laatste Laslett and Sandland, 1989). Bovendien, als er gegevens uit de omgeving van het gebied beschikbaar zijn, dan kunnen die via de model-based benadering worden gebruikt bij het schatten, zoals gebeurt in block-kriging. Dat is natuurlijk speciaal van belang als de steekproef in het te onderzoeken gebied klein is.

Het inbouwen van randomisatie in een steekproefopzet kan, vergeleken met de beste model-based strategie, de efficiëntie verminderen wanneer althans het model correct en de covariantie-functie bekend is. Dit verlies van efficiëntie kan echter meestal worden beperkt door goed gekozen restricties op te leggen aan de randomisatie van de steekproeflokaties, b.v. door stratificering van het gebied. Een groot deel van de klassieke steekproeftheorie is daar aan gewijd. Het resterende verlies zou heel goed een verstandige "premie" kunnen zijn voor robuustheid tegen model-afwijkingen en voor p-zuiverheid. Bovendien werken model-

based strategieën in de praktijk met geschatte covarianties of variogrammen, en zijn dan dus óók niet optimaal.

In het landbouwkundig en bodemkundig onderzoek is de design-based benadering al tientallen jaren met succes toegepast om via steekproeven populatie-parameters te schatten. Webster (1977) bespreekt de toepassingsmogelijkheden van diverse steekproefopzetten, waaronder geneste, in het ruimtelijk bodemkundig onderzoek. Zie b.v. De Grijter en Marsman (1985) voor een onderzoek met steekproefpunten in aselechte raaien.

Deze beschouwing werd beperkt tot ruimtelijke steekproeven. Het is duidelijk dat op soortgelijke wijze óók modelvrije schattings- en toetsingsprocedures zijn te verkrijgen voor steekproeven in de tijd en in ruimte en tijd.

Tenslotte moet worden opgemerkt dat, behalve rechtstreeks als basis te dienen voor opzet en analyse van steekproeven, ruimtelijke modellen óók kunnen worden gebruikt om design-based strategieën te evalueren. In een dergelijke hybride benadering worden kwaliteitscriteria van het ξ_p -type gebruikt, die zijn gebaseerd op verwachtingen over herhaalde steekproefname volgens opzet p en over realisaties van ξ . Voorbeelden hiervan zijn al te vinden in Cochran (1946), Quenouille (1949), Das (1950) en Matérn (1960). Diggle en Ter Braak (1982) werkten specifieke gevallen uit voor de ecologie. De praktijk van het steekproefonderzoek zou er zeer bij gebaat zijn als er meer vergelijkende studies zouden worden gedaan naar de kwaliteiten van verschillende design- en model-based strategieën, onder verschillende modellen en afwijkingen daarvan.

LITERATUUR

ASH, R.B.; 1970. Basic probability theory. Wiley, New York, 337 pp.

BARNES, R.J.; 1988. Bounding the required sample size for geologic site characterization. *Mathematical Geology*, 20: 477-490.

BORGMAN, L.E.; 1988. New advances in methodology for statistical tests useful in geostatistical studies. *Mathematical Geology*, 20: 383-403.

CASSEL, C.M., SÄRNDAL, C.E. and WRETMAN, J.H.; 1977. Foundations of inference in survey sampling. Wiley, New York, 192 pp.

- CHAUDURI, A. and VOS, J.W.E.; 1988. Unified theory and strategies of survey sampling. North-Holland, Amsterdam, 414 pp.
- COCHRAN, W.G.; 1946. Relative accuracy of systematic and stratified random samples for a certain class of populations. *Annals of Mathematical Statistics*, 17: 164-177.
- COCHRAN, W.G.; 1977. Sampling techniques. Wiley, New York, 428 pp.
- CORSTEN, L.C.A.; 1989. Interpolation and optimal linear prediction. *Statistica Neerlandica*, 43: 69-84.
- DAHIYA, I.S., ANLAUF, R., KERSEBAUM, K.C. and RICHTER, J.; 1985. Spatial variability of some nutrient constituents of an Alfisol from loess. II. Geostatistical analysis. *Z. Pflanzenernaehr. Bodenk.*, 148: 268-277.
- DAS, A.C.; 1950. Two dimensional systematic sampling and the associated stratified and random sampling. *Sankhya*, 10: 95-108.
- DE GRUIJTER, J.J. and MARSMAN, B.; 1985. Transect sampling for reliable information on mapping units. In D.R. Nielsen and J. Bouma (Eds.), *Soil spatial variability*. Pudoc, Wageningen: 150-165.
- DE GRUIJTER, J.J. and TER BRAAK, C.J.F.; 1990. Model-free estimation from spatial samples: a reappraisal of classical sampling theory. *Mathematical Geology*, 22: 407-415.
- DIGGLE, P.J. and TER BRAAK, C.J.F.; 1982. Point sampling of binary mosaics in ecology. In B. Ranney (Ed.), *Statistics in theory and practice - Essays in honour of Bertil Matérn*. Swedish University of Agricultural Sciences, Section of Forest Biometry, S-901 83, Umeå, Sweden: 107-122.
- FINKE, P.A.; 1991. Een sequentiële bemonsteringsprocedure voor het schatten van het gemiddelde van een ruimtelijk variërend kenmerk. *CHO-TNO Rapporten en Nota's*, No. 24.
- JOURNAL, A.G. and HUIJBREGTS, CH.J.; 1978. *Mining geostatistics*. Academic Press, London, 600 pp.
- KRISHNAIAH, P.R. and RAO, C.R.; 1989. Sampling. *Handbook of statistics*, Vol. 6, North-Holland, Amsterdam, 594 pp.
- LASLETT, G.M. and SANDLAND, R.L.; 1989. Precision and accuracy of kriging estimators with inter-laboratory trial information. In M. Armstrong (Ed.), *Geostatistics*, Vol. 2, Kluwer, Dordrecht: 797-808.
- MATÉRN, B.; 1960. Spatial variation. *Medd. Statens Skogforskningsinst.*, 49: 1-144.
- MATHERON, G.; 1971. The theory of regionalized variables and its applications. *Ecole*

- Nationale Supérieure des Mines de Paris, Fontainebleau, 211 pp.
- MYERS, D.E.; 1989. To be or not to be...stationary? That is the question. *Mathematical Geology*, 21: 347-362.
- PARZEN, E.; 1960. *Modern probability theory and its applications*. Wiley, New York, 464 pp.
- QUENOUILLE, M.H.; 1949. Problems in plane sampling. *Annals of Mathematical Statistics*, 20: 355-375.
- RUSSO, D. and BRESLER, E.; 1981. Soil hydraulic properties as stochastic processes: I. An analysis of field spatial variability. *Soil Sc. Soc. Am. J.*, 45: 682-687.
- SÄRNDAL, C.E.; 1978. Design-based and model-based inference in survey sampling. *Scand. J. Statist.*, 5: 27-52.
- WEBSTER, R.; 1977. *Quantitative and numerical methods in soil classification and survey*. Clarendon Press, Oxford, 269 pp.
- YFANTIS, E.A., FLATMAN, G.T. and BEHAR, J.V.; 1987. Efficiency of kriging estimates for square, triangular, and hexagonal grids. *Mathematical Geology*, 19: 183-205.

EEN SEQUENTIËLE BEMONSTERINGSPROCEDURE VOOR HET SCHATTEN VAN HET GEMIDDELDE VAN EEN RUIMTELIJK VARIËREND KENMERK

P.A. Finke

SAMENVATTING

Bij het karakteriseren van ruimtelijke eenheden kan de klassieke steekproeftheorie worden gebruikt om populatieparameters zoals gemiddelde en standaardafwijking te schatten. De omvang van de steekproef bepaalt in het algemeen een belangrijk deel van de variabele kosten. Bij hoge kosten per waarneming kan sequentiële bemonstering worden toegepast om betrouwbare schattingen van deze populatieparameters te verkrijgen bij een minimale steekproefomvang. Sequentiële bemonstering is een procedure waarin een hypothese, gebaseerd op een aantal waarnemingen, wordt getoetst met behulp van additionele waarnemingen. De uitslag van de toets bepaalt, of verder bemonsteren noodzakelijk is. Het theoretisch kader, de randvoorwaarden en de procedure van sequentiële bemonstering worden toegelicht, mede aan de hand van een uitgewerkt voorbeeld.

1 INLEIDING

In het bodemkundig onderzoek wordt veel aandacht besteed aan het produceren van bodemkaarten en de interpretatie van deze kaarten en bijbehorende punt- of vlakinformatie in het kader van een bepaalde vraagstelling. De informatie die specifiek betrekking heeft op de vraagstelling kan niet altijd rechtstreeks uit de voorhanden gegevens worden afgeleid, zodat men terug moet "het veld in" om deze gegevens alsnog te verzamelen. Optimaal streeft men er dan naar, de ontbrekende gegevens zo te verzamelen, dat een koppeling met

de reeds aanwezige vlakinformatie kan worden gemaakt. Men zal er naar streven het aantal observaties te minimaliseren om de kosten te drukken, vooral als het gaat om complexe metingen of dure analyses.

Als het doel van de bemonstering is, de gebiedsgemiddelde waarde van een kenmerk te schatten, bijvoorbeeld als de vraagstelling betrekking heeft op dat gebied als geheel, dan kan sequentieel worden bemonsterd. Sequentiële bemonstering is gebaseerd op de klassieke steekproeftheorie (Wald, 1947), en wordt gebruikt om een schatting te verkrijgen van het populatiegemiddelde. Deze schatting is betrouwbaar en wordt verkregen met behulp van een minimaal aantal waarnemingen. Toepassing van deze techniek is vooral aantrekkelijk als waarnemingen arbeidsintensief en/of duur zijn. Een eerste toepassing van sequentiële bemonstering in de bodemkunde wordt gegeven door Stein et al. (1989).

Sequentiële bemonstering is naar de terminologie van Särndal (1978) en De Gruijter en Ter Braak (1990) een voorbeeld van de "design-based" benadering. Deze benadering stelt, dat de vereiste onafhankelijkheid van de waarnemingen wordt verkregen door de waarnemingslokaties onderling onafhankelijk te loten. Hoe een eventueel semi-variogram er uit ziet, doet bij deze benadering niet ter zake.

Als bij de sequentiële bemonstering de positie-coördinaten van de waarnemingen ook worden vastgelegd, en als na afloop van de bemonsterings-procedure het aantal waarnemingen voldoende groot is, om een betrouwbaar variogram te schatten, dan kan met behulp van blok-kriging een betrouwbare vlak-voorspelling worden verkregen (Journel and Huijbregts, 1978). Deze methode is een voorbeeld van de "model-based" benadering. Onderling onafhankelijke loting van meetlokaties is bij deze benadering geen randvoorwaarde. Nadeel van de blok-kriging methode ten opzichte van de sequentiële methode is, dat blok-kriging a priori eisen stelt aan het aantal waarnemingen, omdat een variogram moet kunnen worden geschat.

Samenvattend kan worden gesteld, dat toepassing van sequentiële bemonstering in de ruimtelijke statistiek zinvol is bij:

- (1) het karakteriseren van ruimtelijke eenheden (bijvoorbeeld/ eventueel na stratifikatie);
- (2) het karakteriseren van partijen (bijvoorbeeld verontreinigde grond), als de herkomst van het materiaal niet meer m.b.v. ruimtelijke coördinaten kan worden vastgelegd.

2 STATISTISCHE ACHTERGROND

2.1 Afleiding van acceptatie- en verwerpingscriteria

De statistische basis voor sequentiële bemonsteringsmethoden is gelegd door Wald (1947). Aanbevolen literatuur is Wetherill (1966). De Sequential Probability Ratio Test SPRT (Wald, 1947) is ontworpen om te beslissen welke van twee enkelvoudige hypotheses $H_0: \mu = \mu_0$ en $H_1: \mu = \mu_1$, kan worden geaccepteerd. De keuze voor H_0 of H_1 wordt gedaan op basis van de observaties x_1, x_2, \dots, x_n , met onbetrouwbaarheden α en β voor H_0 respectievelijk H_1 .

De optimale toetsingsgrootheid voor dergelijke problemen (Neyman and Pearson, 1928) is een waarschijnlijkheidsratio L_n , gegeven als:

$$L_n = \frac{\text{kans op observaties } x_1, \dots, x_n \text{ gegeven } H_1}{\text{kans op observaties } x_1, \dots, x_n \text{ gegeven } H_2} = \prod_{i=1}^n \frac{f(x_i, \mu_1)}{f(x_i, \mu_0)} \quad (1)$$

waarin $f(x, \mu_j)$ ($j=0,1$) is de kansdichtheid behorende bij observatie x , gegeven dat het gemiddelde gelijk is aan μ_j , en n is het aantal observaties. Bij de berekening van L_n uit produkten van kansdichtheden zoals in vergelijking 1 worden de observaties x_i stochastisch onafhankelijk verondersteld. Tevens moet voor de kansdichtheid van x een bepaalde verdeling worden aangenomen.

Bij sequentiële bemonstering worden drie gebieden onderscheiden in de L_n -n ruimte, elk gekoppeld aan een bemonsterings-scenario:

$$\text{zolang } B < L_n < A \quad : \text{ doorgaan met bemonsteren (indifferentie);} \quad (2a)$$

$$\text{zodra } L_n < B \quad : \text{ stoppen met bemonsteren en kiezen voor } H_0; \quad (2b)$$

$$\text{zodra } L_n > A \quad : \text{ stoppen met bemonsteren en kiezen voor } H_1; \quad (2c)$$

waarin A en B de acceptatiegrenzen voor H_1 respectievelijk H_0 .

Wald (1947) leidt af, dat de waardes van A en B kunnen worden bepaald met behulp van de onbetrouwbaarheden α en β door:

$$A \sim (1-\beta)/\alpha \tag{3a}$$

$$\text{en } B \sim \beta/(1-\alpha), \tag{3b}$$

waarin α de kans om H_1 te accepteren terwijl H_0 waar is, en β de kans om H_0 te accepteren terwijl H_1 waar is. De keuze van α en β berust dus op statistische criteria. Een kleinere α en β leiden tot een grotere A en kleinere B, dus tot een bredere zone van indifferentie.

Omdat de sequentiële bemonsteringsprocedure wordt gestopt zodra $L_n < B$ dan wel $L_n > A$, kunnen de ongelijkheden voor A en B als gelijkheden worden beschouwd. Wald (1947) toont aan, dat de fout die hierdoor wordt geïntroduceerd verwaarloosbaar klein is.

2.2 Samengestelde hypotheses: μ en σ onbekend

Als van een toevalsvariabele x wel het soort verdeling, maar niet de parameters van die verdeling bekend zijn, kan dit bij sequentiële toetsing leiden tot samengestelde hypotheses. Stel, x is normaal verdeeld, maar μ en σ zijn onbekend. Bij de keuze tussen $H_0: \mu = \mu_0$ en $H_1: |\mu - \mu_0|/\sigma > \delta$, waarbij δ een tolerantie marge aangeeft, is H_1 een samengestelde hypothese. Toetsing van dergelijke hypotheses is mogelijk door gebruik te maken van gewichtsfuncties. Bij normaal verdeelde x suggereert Wald:

$$\text{voor } \sigma \quad : \quad \phi(\sigma) = 1/2 \text{ als } 0 < \sigma \leq c, \tag{4a}$$

$$\text{en } \phi(\sigma) = 0 \text{ in alle andere gevallen;} \tag{4b}$$

$$\text{voor } \mu \quad : \quad W_0: \mu = \mu_0; \tag{5a}$$

$$W_1: \phi(\mu) = 1/2 \text{ als } \mu = \mu_0 \pm \delta \cdot \sigma \tag{5b}$$

$$\text{en } \phi(\mu) = 0 \text{ in alle andere gevallen.} \tag{5c}$$

De waarschijnlijkheid van observaties (x_1, x_2, \dots, x_n) wordt dan bij normaal verdeelde x :

$$\text{voor } H_0 : \int_0^c \frac{1}{c} \cdot \prod_{i=1}^n f(x_i, \mu_0) \cdot d\sigma; \quad (6)$$

$$\text{en voor } H_1 : \int_0^c \frac{1}{2c} \cdot \left[\prod_{i=1}^n f(x_i, \mu_0 + \delta \cdot \sigma) \cdot d\sigma + \prod_{i=1}^n f(x_i, \mu_0 - \delta \cdot \sigma) \cdot d\sigma \right] \quad (7)$$

als $c \rightarrow \infty$, dan kan de waarschijnlijkheidsratio worden geschreven als:

$$L_n = \frac{\int_0^{\infty} \left\{ \int_0^{\infty} \frac{1}{\sigma^n} \cdot \exp[-1/2\sigma^2 \cdot \sum_{i=1}^n (x_i - \mu_0 - \delta \sigma)^2] d\sigma + \int_0^{\infty} \frac{1}{\sigma^n} \cdot \exp[-1/2\sigma^2 \cdot \sum_{i=1}^n (x_i - \mu_0 + \delta \sigma)^2] d\sigma \right\}}{\int_0^{\infty} \frac{1}{\sigma^n} \cdot \exp[-1/2\sigma^2 \cdot \sum_{i=1}^n (x_i - \mu_0)^2] d\sigma} \quad (8)$$

Deze ratio kan worden berekend door numerieke integratie van vergelijking 8, onder substitutie van observaties x_i , de hypothetische μ_0 en de tolerantie δ .

Wetherill (1966) beschrijft tevens hoe bij onbekende μ een schatting kan worden verkregen van de populatie-standaardafwijking. Hierbij wordt $H_0: \sigma = \sigma_0$ sequentieel getoetst tegen $H_1: \sigma = \sigma_1$. Dit leidt tot een waarschijnlijkheidsratio L'_n , die met behulp van de bovengenoemde criteria A en B kan worden getoetst:

$$L'_n = \left\{ \frac{\sigma_0}{\sigma_1} \right\}^{(n-1)} \exp \left\{ (n-1) s^2 \left[\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2} \right] / 2 \right\}, \quad (9)$$

waarin s^2 de steekproefvariantie is.

2.3 Randvoorwaarden

Bij het toepassen van sequentiële bemonstering met behulp van vergelijkingen 2, 3, 8 en 9 wordt uitgegaan van 2 veronderstellingen:

- (1) de waarnemingen zijn onderling stochastisch onafhankelijk;
- (2) de waarnemingen zijn normaal verdeeld.

Stochastische onafhankelijkheid van de waarnemingen kan worden bereikt door bij de lokatiekeuze van de monsternamen een toevalsfactor te betrekken, bijvoorbeeld door "simple random sampling" (Cochran, 1963). Tevens is het noodzakelijk, dat de H_0 -hypothese niet is gebaseerd op de waarnemingen waarmee de toetsing plaatsvindt.

Er wordt van uitgegaan, dat de waarnemingen normaal zijn verdeeld. Als vooraf bekend is, dat de waarnemingen niet normaal verdeeld zullen zijn, dienen de waarnemingen getransformeerd te worden. Hierbij moet worden opgemerkt, dat Wetherill (1966) een waarschijnlijkheidsratio geeft voor binominaal verdeelde x .

Bij de berekening van de waarschijnlijkheidsratio L_n en de toetsing met behulp van de acceptatiegrenzen A en B moeten de parameters α , β en δ door de gebruiker worden opgegeven. De keuze van α en β wordt bepaald door statistische criteria. De tolerantie marge δ hangt af van wat door de gebruiker als een belangrijke afwijking van $H_0: \mu = \mu_0$ wordt beschouwd.

2.4 Procedure

De sequentiële bemonsteringsprocedure start met het definiëren van de onbetrouwbaarheden α en β en de tolerantie marge δ . Hierna wordt een aantal (N) waarnemingen gedaan volgens een bemonsteringscenario, waarin de lokatie door het toeval wordt bepaald. Met behulp van deze waarnemingen wordt de H_0 -hypothese opgesteld, bijvoorbeeld door μ_0 gelijk te stellen aan het gemiddelde van de N waarnemingen.

Vervolgens worden M additionele waarnemingen gedaan ($M=2$), waarmee H_0 wordt getoetst. Afhankelijk van het toetsingsresultaat zijn 3 vervolgcenario's mogelijk:

1. bij acceptatie van H_0 ($L_n < B$) wordt de bemonstering gestopt;
2. bij indifferentie ($B < L_n < A$) wordt een nieuwe waarneming gedaan en opnieuw getoetst, nu met behulp van $M+1$ waarnemingen. De uitslag bepaalt het vervolgsценario;
3. bij afwijzing van H_0 ($L_n > A$) dient een nieuwe H'_0 te worden geformuleerd, gebaseerd op N' (waarbij $N' > N$) waarnemingen. Met behulp van de reeds gedane waarnemingen $N'+1$ tot en met M' ($M' = N + M - N'$) kan H'_0 worden getoetst. De uitslag bepaalt het vervolgsценario.

Om een sequentiële bemonsteringsprocedure efficiënt te kunnen uitvoeren, moet aan enige logistieke voorwaarden tegemoet worden gekomen. Een personal computer is vereist voor de toetsing. Als de waarnemingen in het veld worden gedaan, is het efficiënt daar ook te toetsen. Als de waarnemingen laboratoriumanalyses vergen, kan het verzamelen van een groot aantal monsters en eventueel het doen van een voorbewerking opwegen tegen de verplaatsingskosten verbonden aan herhaalde monsternamen.

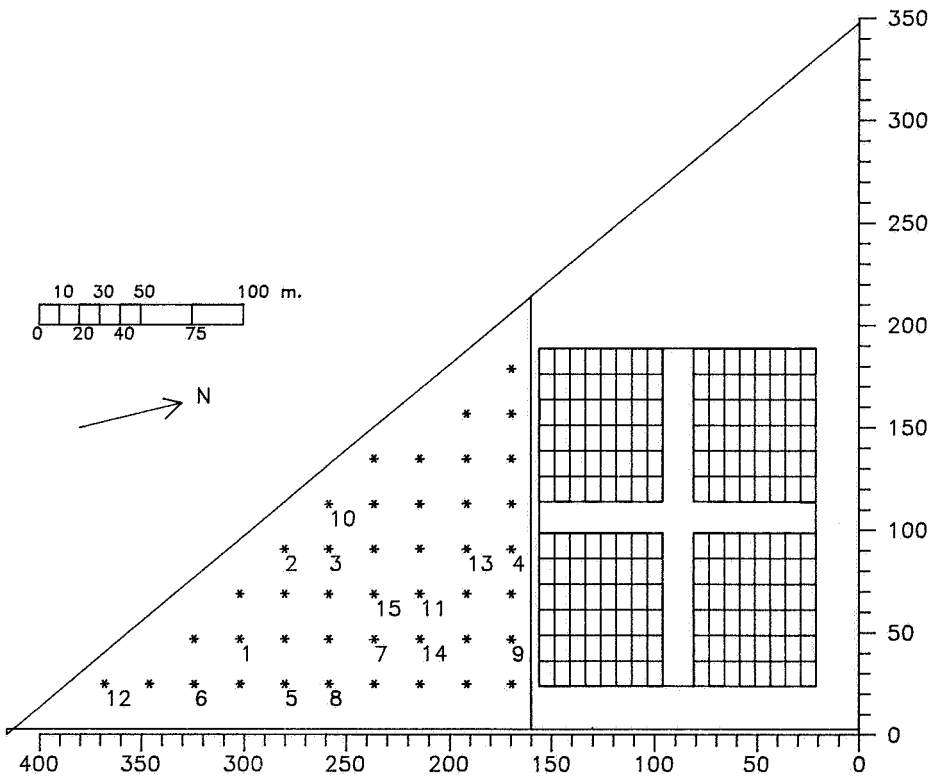
3 BEPALING VAN DE N-MINERAAL INHOUD VAN EEN LANDBOUWPERCEEL VOOR HET VERKRIJGEN VAN EEN BETROUWBAAR BEMESTINGSADVIES

3.1 Inleiding

In de akkerbouw is het gewoonte om elk jaar de bovengrond te analyseren op het gehalte aan minerale stikstof. Uit dit gehalte kan worden afgeleid hoeveel stikstof minimaal moet worden toegediend in de vorm van kunstmest om aan de theoretische behoefte van het gewas te voldoen. De adviesbemesting voor stikstof wordt berekend door de stikstofvoorraad af te trekken van een streefniveau gerelateerd aan de gewasbehoefte. De optimale hoeveelheid kunstmest is groot genoeg om samen met de reeds aanwezige stikstof in de gewasbehoefte te voorzien, en minimaliseert de kans op vervuiling van het grondwater door overbemesting. Aan deze twee voorwaarden kan in theorie het best worden voldaan door een combinatie van lokatiespecifiek bemestingsonderzoek en lokatiespecifieke bemesting. In de praktijk wordt gelijkmatige bemesting op basis van een gemiddeld bemestingsadvies voor het perceel nagestreefd, hoewel op experimentele basis in de USA

lokatiespecifieke landbouw wordt bedreven.

De standaardprocedure om te komen tot een bemestingsadvies is het analyseren van een mengmonster, bestaande uit submonsters afkomstig van willekeurige lokaties op het perceel. Nadeel van deze procedure is o.a. de gevoeligheid voor fouten geïntroduceerd door onvolledige homogenisatie van het mengmonster. Ook kan op deze manier geen indruk worden verkregen van de variabiliteit binnen het perceel.



Figuur 1 Perceelsindeling en lokaties voor N-mineraalbemonstering.

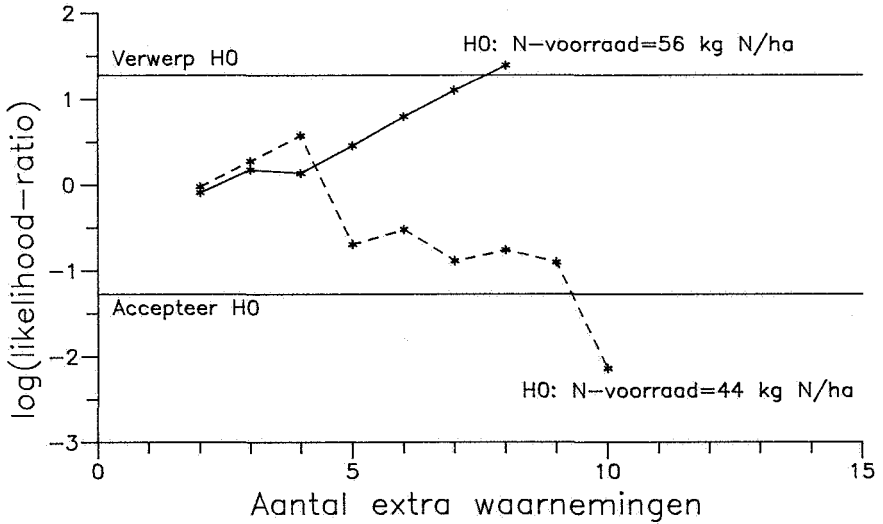
De procedure gevolgd in dit onderzoek bestond uit het bemonsteren van een perceel volgens een grid met een (arbitraire) maaswijdte van 22 meter met een door loting verkregen

startpunt (figuur 1). Dit leverde 43 monsters op, die in de diepvries werden bewaard om stikstofomzettingen te voorkomen. De aldus verkregen bodemonsters werden met behulp van toevalsgetallen in een willekeurige volgorde geplaatst, en in die volgorde ontdooid, geanalyseerd en sequentieel getoetst. De lokatie van de eerste 15 gelote bodemonsters is aangegeven in figuur 1.

3.2 Resultaten sequentiële bemonstering

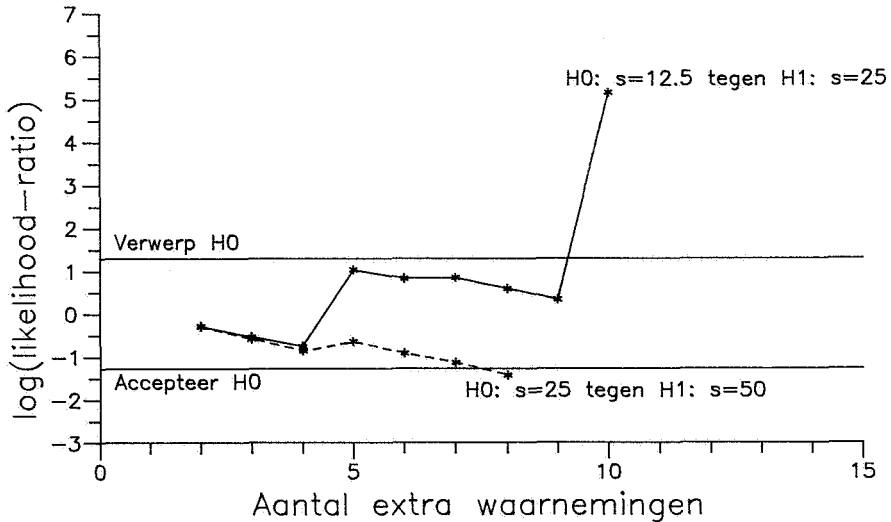
Voorafgaande aan de toetsing werden α en β gelijkgesteld aan 5%, en δ aan 1. De waarde van δ is vastgesteld op grond van het te verwachten effect van een te laag bemestingsadvies op de opbrengst. De responsiecurve van de gewasopbrengst op de stikstofvoorraad dichtbij de optimale voorraad kent in het algemeen een vrij vlak verloop (Neeteson, 1989). Het werd daarom verondersteld, dat een eventuele onder- of overbemesting ter grootte van 1 maal de standaardafwijking van het bemestingsadvies geen grote opbrengstderiving tot gevolg zou hebben. De standaardafwijking van het bemestingsadvies is gelijk aan die van de stikstofvoorraad, omdat beide variabelen een constant verschil hebben. Dit impliceerde, dat δ gelijk werd gesteld aan 1. Er werd dus getoetst $H_0: \mu = \mu_0$ tegen $H_1: |\mu - \mu_0| > \sigma$, beide met 95% betrouwbaarheid.

De eerste H_0 -hypothese omtrent de gemiddelde stikstofvoorraad werd gelijkgesteld aan het resultaat van de analyse van het mengmonster: $H_0: \mu = 56$ kg N/ha. De sequentiële toetsing leverde een afwijzing van deze hypothese op na 8 waarnemingen, blijkend uit het overschrijden van de afwijzingslijn door de likelihood-lijn in figuur 2. Als nieuwe hypothese werd de gemiddelde N-voorraad van de waarnemingen 1 tot en met 4 genomen: $H_0: \mu = 44$ kg N/ha. Na 9 keer toetsen (waarneming 5 tot en met 14) werd deze hypothese geaccepteerd (likelihood-lijn lager dan acceptatielijn in figuur 2). In totaal waren er dus 15 analyses nodig voor het verkrijgen van een betrouwbare schatting van de gemiddelde stikstofvoorraad.



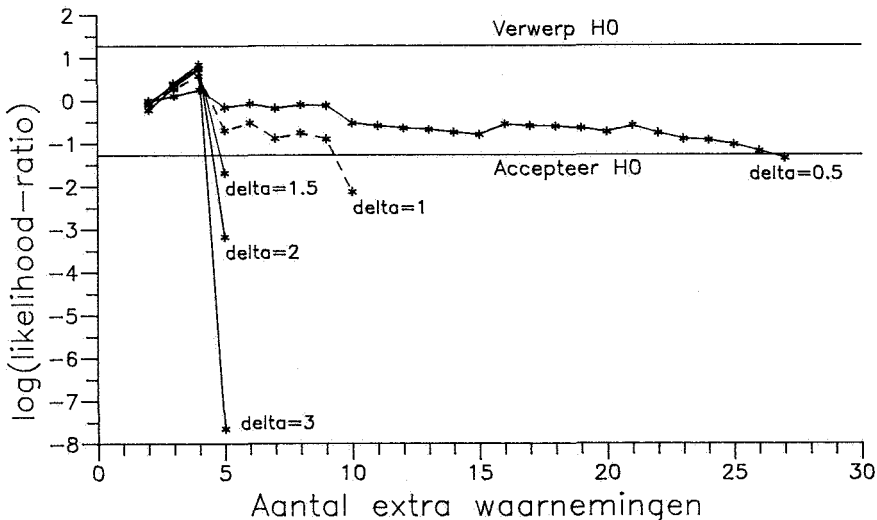
Figuur 2 Toetsing H_0 -hypotheses omtrent gemiddelde stikstofvoorraad.

Een sequentiële toetsing van de populatiestandaardafwijking werd uitgevoerd met een $H_0: \sigma = 12.5$ tegen $H_1: \sigma = 25$ kg N/ha. Deze H_0 , gebaseerd op de steekproefstandaardafwijking van de waarnemingen 1 tot en met 4, werd na 10 extra waarnemingen verworpen. Vervolgens werden de hypothesen $H_0: \sigma = 25$ en $H_1: \sigma = 50$ kg N/ha tegen elkaar getoetst. H_0 werd geaccepteerd na 7 keer toetsen (figuur 3).



Figuur 3 Toetsing H_0 -hypothesen omtrent standaardafwijking stikstof-voorraad.

Bij interpretatie van de toetsingsresultaten dient de invloed van de keuze van de tolerantie δ niet te worden onderschat. Een grotere waarde van δ bij gelijkblijvende H_0 kan leiden tot snellere acceptatie (bij minder waarnemingen), maar maakt verwerping minder waarschijnlijk vanwege de grotere tolerantiemarge. In figuur 4 is de gevoeligheid van het toetsresultaat voor verschillende δ weergegeven. Het is duidelijk, dat de keuze van δ moet berusten op het oordeel van de deskundige.



Figuur 4 Gevoeligheid van het toetsresultaat voor verschillende δ .

Samenvattend kan worden gezegd, dat een aantal van 14 waarnemingen (en 1 waarneming aan een mengmonster) voldoende was om een betrouwbare schatting te kunnen maken van zowel het gemiddelde als de standaardafwijking van de stikstofvoorraad in het onderzochte perceel (44 respectievelijk 25 kg N/ha). Na analyse van alle 43 bodemmonsters die waren verzameld, bleken het gemiddelde en de standaardafwijking 46.7 respectievelijk 27.1 kg N/ha te bedragen. Geostatistische analyse met behulp van het programmapakket AKRIP (Kafritsas and Bras, 1981) leverde een zuiver nugget model op in afwezigheid van trend. In dit geval (geen trend en een zuiver nuggetmodel) levert block-kriging een vlakvoorspelling op, die gelijk is aan het gemiddelde van de waarnemingen dus 46.7 kg N/ha.

4 CONCLUSIES

Als een betrouwbare uitspraak over het gebiedsgemiddelde van een ruimtelijk variërend kenmerk moet worden verkregen, en de kosten per afzonderlijke waarneming zijn hoog, dan

kan sequentiële bemonstering een goed alternatief zijn voor blok-kriging. Terwijl voor een zinvolle toepassing van blok-kriging een aantal van enkele tientallen waarnemingen noodzakelijk is om de ruimtelijke structuurfuncties te kunnen schatten, wordt in een sequentiële bemonsteringsprocedure het aantal benodigde waarnemingen juist geminimaliseerd.

De bruikbaarheid van het resultaat van een sequentiële toetsingsprocedure is in hoge mate afhankelijk van de tolerantiemargin δ . Deze dient door een deskundige te worden gekozen op basis van veldcriteria.

LITERATUUR

- COCHRAN, W.G.; 1963. Sampling techniques. John Wiley, New York.
- JOURNEL, A.G. and HUIJBREGTS, Ch.J.; 1978. Mining Geostatistics. Academic Press, London.
- GRUIJTER, J.J. De and TER BRAAK, C.J.F.; 1990. Model-free estimation from spatial samples: A reappraisal of classical sampling theory. Math. Geol. Vol. 22:407-415.
- KAFRITSAS, J. and BRAS, R.L.; 1981. The practice of Kriging. Department of civil engineering, Massachusetts Institute of Technology Report no. 263.
- NEETESON, J.J.; 1989. Assessment of fertilizer nitrogen requirement of potatoes and sugar beet. PhD Thesis Agricultural University, Wageningen.
- SÄRNDAL, C-E.; 1978. Design-based and model-based inference in survey sampling. Scand. J. Statist. Vol. 5:27-52.
- STEIN, A., BOUMA, J., KROONENBERG, S.B. and COBBEN, S.; 1989. Sequential sampling to measure the infiltration rate within relatively homogeneous soil units. Catena 16: 91-100.
- WALD, A.; 1947. Sequential analysis. John Wiley, New York.
- WETHERILL, C.B.; 1966. Sequential methods in statistics. John Wiley, New York.

ONTWERP VAN HET MEETNET VOOR LUCHTVERONTREINIGING

D. Onderdelinden

SAMENVATTING

Luchtverontreinigingsconcentraties worden in het algemeen bepaald met behulp van vaste meetstations; een meetnet. In Nederland werd ongeveer 20 jaar geleden een omvangrijk meetnet opgezet. Omdat meestal beoogd wordt een uitspraak te doen over het gehele gebied is het van groot belang na te gaan wat de ruimtelijke representativiteit van de meetresultaten is. In deze presentatie wordt ingegaan op de statistische eigenschappen en de onderlinge samenhang van de meetresultaten. De grootte van de interpolatiefouten wordt hierbij beschreven met een tweetal interpolatietechnieken. Door de grootte van de interpolatiefout te relateren aan de meetnetdichtheid wordt een praktische procedure verkregen voor het ontwerp van een meetnet met een voorgeschreven nauwkeurigheid. Het resultaat voor het meetnet van SO₂ wordt gegeven.

1 INLEIDING

Luchtverontreinigingsconcentraties worden in het algemeen bepaald op vaste meetstations. In Nederland werd ongeveer 20 jaar geleden een omvangrijk meetnet opgezet voor de componenten SO₂ (200), NO_x (90), CO (40) en O₃ (30). Het gedrag van uur- en dagwaarden uitgedrukt als 50-, 90- en 98-percentielen werd vergeleken met de beschikbare gezondheidsstandaarden. Nieuwe wettelijke regelingen in nationaal zowel als internationaal

verband maakten een nieuwe meetstrategie noodzakelijk. In deze presentatie ligt hierbij de nadruk op het ontwerp van een landelijk meetnet ter beschrijving van regionale niveaus. Voor de beschrijving van meer lokale concentraties is dit meetnet niet geschikt. Hiertoe werd gekozen voor een combinatie van een beperkt aantal meetstations in een goed gedefinieerde omgeving en berekeningen met een verspreidingsmodel.

2 UITGANGSPUNTEN

2.1 Doelstellingen Meetnet

De doelstellingen van een meetnet kunnen als volgt worden samengevat:

- * controle op overschrijding van in de wet gestelde grens- of richtwaarden;
- * informatie tijdens het optreden van episodes van verhoogde luchtverontreiniging (alarmeringsfunctie);
- * validatie van modelresultaten. Vergelijking van modelresultaten en metingen geeft de mogelijkheid tot herkennen van bronnen, veranderingen in de emissiesterkte en verdeling alsmede de vaststelling van trend;
- * bepaling van de atmosferische depositie. De meting van de concentratie in de lucht zowel als in regenwater gecombineerd met een depositiemodel maakt de bepaling van bijvoorbeeld de zure depositie over Nederland mogelijk.

Voor al deze doelstellingen wordt een uitspraak over een groter gebied dan alleen de lokatie van een meetstation verlangd. Van groot belang is daarom met welke nauwkeurigheid het ruimtelijk concentratieveld uit de waarnemingen van het meetnet kan worden bepaald.

2.2 Wettelijke Basis

In de wet op de luchtverontreiniging (Staatsblad, 1981) wordt de controle van de luchtkwaliteit toegekend aan het regionale bestuur, i.e. de provincies en de gemeenten. De centrale overheid definieert de grenswaarden en de nauwkeurigheid waarmee de concentraties moeten worden vastgesteld. Deze voorschriften worden per component opgegeven in zogenaamde AMvB's. Er zijn op dit moment regelingen voor SO₂. Zwarte

rook, NO₂, CO en Pb. Voor bijvoorbeeld SO₂ wordt de vaststelling van het 98-percentiel van daggemiddelde waarden voor elke 1*1 km² met een standaarddeviatie van 15% voorgeschreven. De provincies maken hiervoor gebruik van het door het RIVM beheerde landelijk meetnet.

2.3 Meetstrategie

Teneinde te voldoen aan de wettelijke eisen werd een meetstrategie ontwikkeld afhankelijk van het noodzakelijke ruimtelijk oplossend vermogen. Er werden daarom vier ruimtelijke schalen onderscheiden:

- * regionale schaal
oplossend vermogen 1*1 km², bronnen op een afstand > 5 km;
- * stadsachtergrond
oplossend vermogen 1*1 km², stad > 40.000 inwoners;
- * industriegebied
oplossend vermogen 100*100 m², bron op een afstand < 5 km;
- * stadsstraten
oplossend vermogen 100*10 m², in de straat passeren minstens 5.000 voertuigen per etmaal.

Voor de vaststelling van de niveaus op regionale schaal, het concentratieveld, wordt uitgegaan van statistische procedures. Voor de andere situaties wordt uitgegaan van geschatte emissies, een beperkt aantal metingen in geklassificeerde situaties en daaraan getoetste modelberekeningen. In deze presentatie wordt nader ingegaan op de procedures die in het jaar 1986 hebben geleid tot een nieuw gemoderniseerd landelijk meetnet op regionale schaal.

3 STATISTISCHE EIGENSCHAPPEN

3.1 Optimum Interpolatie

De analyse van concentratievelden kan op een vergelijkbare wijze worden benaderd als die

D. Onderdelinden

van meteorologische velden. Door Gandin (1965) is een zogenaamde optimum interpolatieschema voorgesteld waarbij de interpolatiefout wordt geminimaliseerd op basis van de statistische eigenschappen van de betreffende grootheid zoals waargenomen in het verleden.

De concentratie C_{ot} (verklaarde variabele) op het interpolatiepunt, dat wil zeggen een positie waar niet gemeten is, wordt geschat als een lineaire combinatie van gemeten concentraties op omliggende meetpunten (verklarende variabelen). Het verschil tussen de werkelijke (doch onbekende) en geschatte concentratie is de interpolatiefout E_{ot} . De op de omliggende meetpunten i gemeten waarden bestaan uit de werkelijke concentraties C_{it} , gemeten op het tijdstip t en de waarnemingsfouten H_{it} , waarin naast de meetfouten eveneens de invloed van de lokale effecten, zoals kleine bronnen in de onmiddellijke omgeving van het meetpunt is verdisconteerd (Van Egmond en Onderdelinden, 1981):

$$C_{ot} = P_o (C + H) + E_{ot} \quad (1)$$

hierbij is

- P_o de vector van de gewichtscoëfficiënten P_{io}
- C de matrix van concentraties op tijdstip t en lokatie i
- H de matrix van meetfouten op tijdstip t en lokatie i

De coëfficiënten P_{io} worden gevonden door minimalisatie van E_{ot}^2 over een reeks realisaties in het verleden. In matrix notatie volgt de oplossing voor P_o als:

$$P_o = (R + I\bar{h}^2)^{-1}r_o \quad (2)$$

waarin

- R de matrix van een genormeerde (deling door variantie ten opzichte van een gemiddeld niveau) kruisprodukten (correlaties) van concentraties op lokaties i en j
- I de eenheidsmatrix
- \bar{h}^2 de genormeerde meetfout gemiddeld over alle lokaties
- r_o de verwachte correlatie tussen interpolatiepunt en de meetpunten i

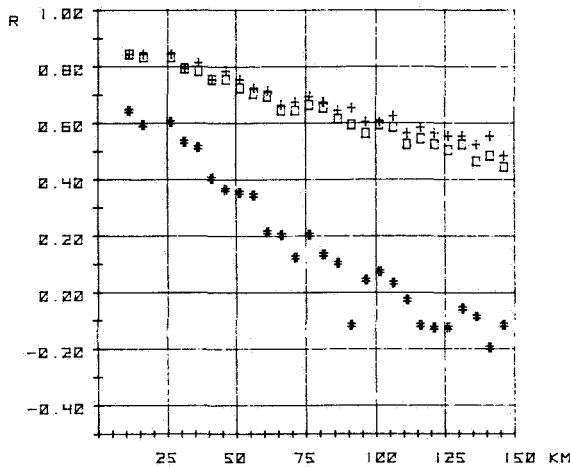
De correlaties worden berekend rondom een gemiddeld niveau. Dit niveau kan bestaan uit:

- het tijd en ruimte gemiddelde over het veld (meetstations);
- het ruimte gemiddelde op tijdstip;

- het individuele stationsgemiddelde.

In Figuur 1 zijn de correlaties voor de verschillende stationsafstanden weergegeven voor de verschillende gemiddelde niveaus. De correlaties zoals berekend uit individuele stationsparen blijkt een redelijk afstandsgedrag te vertonen. De correlatie ten opzichte van het momentane ruimte gemiddelde neemt daarbij het snelst af met de afstand.

Elke middeling over de tijd leidt zoals te verwachten valt uit bijvoorbeeld de optredende variabele windrichting tot een kleinere variatie over het veld.



Figuur 1 Correlatiefuncties voor SO₂-uurgemiddelde concentraties ten opzichte van tijd en ruimte gemiddelde (□), stationsgemiddelde (+) en ruimte gemiddelde (*).

De correlatiefuncties naderen voor kleine stationsafstanden niet naar de waarde 1; meetstations op afstanden van enkele honderden meters geven ten gevolge van meetfouten en de invloed van kleine lokale bronnen geen indentieke signalen. Deze "meetfout" wordt uit de correlaties geschat als het verschil tussen de naar afstand 0 geëxtrapolerde waarde r_{00} en de waarde 1:

$$h^2 = (1 - r_{00})/r_{00} \tag{3}$$

D. Onderdelinden

Uit figuur 1 volgt voor de ten opzichte van het ruimtevlak genormeerde meetfout een waarde van \underline{h}^2 van ca. 0.10.

De genormeerde interpolatiefout $\underline{\epsilon}^2$ wordt uiteindelijk gevonden als

$$\underline{\epsilon}^2 = 1 - r_0 p_0 \quad (4)$$

Het blijkt dat de interpolatiefout kleiner wordt naarmate de onderlinge correlaties groter worden.

3.2 Structuurfunctie

De hierboven beschreven optimum interpolatieprocedure is minder geschikt voor praktische toepassing in de luchtverontreiniging.

Afgezien van de rekentechnische inspanningen voor het inverteren van een matrix van hoge orde (vergelijking 2) kan bijna singulariteit optreden wanneer twee stations een hoge toevallige correlatie hebben.

Daarnaast voorziet de procedure niet in ruimtelijke variatie in de statistische eigenschappen van de beschouwde component. Een eenvoudiger procedure voor de vaststelling van het verband tussen statistische eigenschappen en interpolatiefout wordt hierna gegeven met behulp van de zogenaamde structuurfunctie. Samenhangend met het lognormale karakter van de frequentieverdelingen van luchtverontreinigingsconcentraties wordt daarbij uitgegaan van de logaritmen van de concentraties.

De structuurfunctie wordt nu gedefinieerd als:

$$\underline{B}_{ij}^m = \frac{1}{2} \overline{(C_{it} - C_{jt})^2} \quad (5)$$

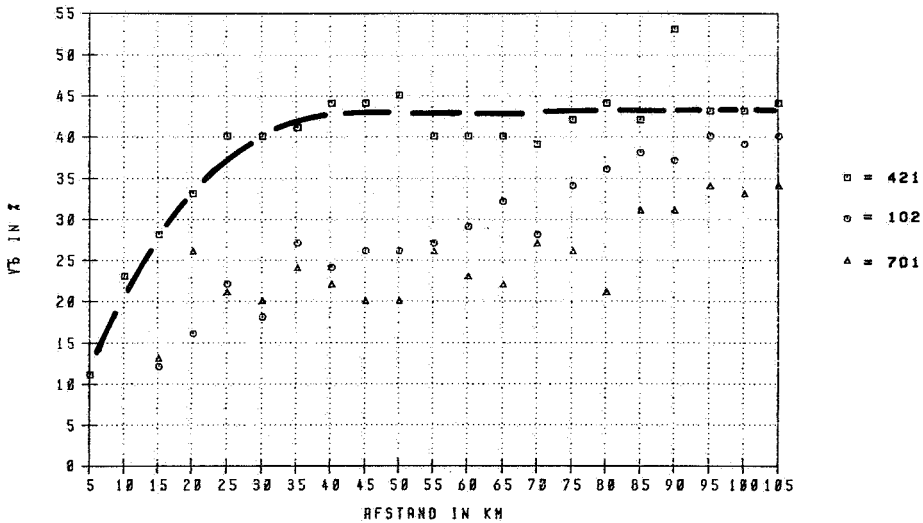
waarin c de logaritme van de gemeten concentratie
de bovenstreep middeling over de tijd

Een voorbeeld van een gemeten structuurfunctie voor 98-percentielwaarden van daggemiddelde SO_2 -concentraties wordt gegeven in Figuur 2. De gemeten structuurfunctie is $b^m(x) = b(x) + \underline{h}^2$. De relatieve lokale fout \underline{h} kan rechtstreeks uit de naar 0 geëxtrapoleerde waarde worden afgelezen, namelijk ca. 10%. Invulling van de

structuurfunctie in de gekwadrateerde, over de tijd gemiddelde vergelijking (1), levert een uitdrukking voor de relatieve interpolatiefout:

$$\overline{F^2} = 2 P_0 b_0 - P_0 B P_0 + h^2 P_0^2 \quad (6)$$

Voor de coëfficiënten P_{i0} wordt nu uitgegaan van een negatieve exponentiële weging (Van Egmond et al., 1984). Het verlies ten opzichte van het optimum interpolatieschema is slechts gering en de exponent kan verschillend gekozen worden voor delen van het veld, afhankelijk van de waargenomen regionale correlatie-afstand.

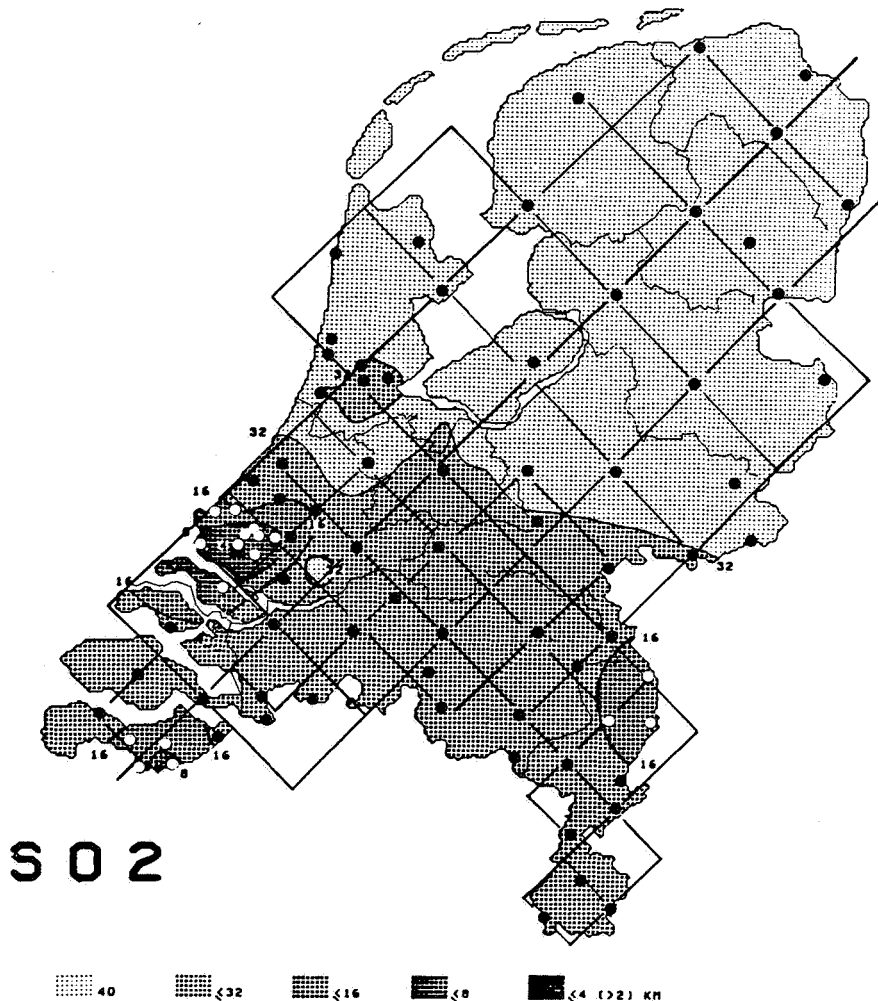


Figuur 2 Structuurfuncties voor 98-percentielwaarden van daggemiddelde SO_2 -concentraties
421 Rotterdam, 102 Arcen Noord-Limburg, 701 Uddel Gelderland

Deze correlatie-afstand wordt bepaald uit het afstandsgedrag van de structuurfunctie. Hiertoe wordt de structuurfunctie beschreven met de modelfunctie:

$$b^m(x) = h^2 + (b - h^2) (1 - \exp-x/x_s) \quad (7)$$

waarin



Figuur 3 Plaatsen van de SO_2 -meetstations in het Landelijk Meetnet Luchtverontreiniging.
De grijsntin geven de vereiste onderlinge stationsafstanden aan voor het realiseren van een standaardfout van hoogstens 15% voor 98-percentielwaarden van daggemiddelden.

- b de waarde van b_m op grote onderlinge stationsafstand
- x_s een specifieke afstand waarbinnen de variatie van meetresultaten nog gecorreleerd zijn

4 ONTWERP MEETNET

4.1 Procedure

De procedure voor het ontwerp van het nieuwe meetnet SO_2 van 1986 kan nu als volgt worden samengevat.

1. Voor een gekozen gedeelte van Nederland wordt de structuurfunctie bepaald uit een beperkt aantal stations.
2. De parameters van de modelfunctie van de structuurfunctie worden bepaald.
3. De interpolatiefout F wordt berekend voor het midden van een regelmatig grid van 12 stations, voor een serie gridgroottes x_o .
4. Uit de numerieke relatie tussen interpolatiefout F en x_o wordt de x_o bepaald waarvoor $F = 15\%$, de wettelijk vastgestelde nauwkeurigheid voor het 98-percentiel SO_2 .
5. Deze procedure wordt herhaald voor alle te onderscheiden gedeeltes van het land.

4.2 Resultaat voor SO_2

Een voorbeeld van een x_o -veld wordt gegeven in Figuur 3. Voor het noordelijk gedeelte van het land geldt de nauwkeurigheidseis niet omdat daar overschrijding van de grenswaarden onwaarschijnlijk is. In dit gebied wordt een x_o van 40 km aangehouden.

De inmiddels gerealiseerde stationsconfiguratie voor SO_2 is eveneens in de figuur aangegeven. Het aantal ruimtelijk representatieve SO_2 -stations bedraagt 77.

Voor de andere componenten zijn soortgelijke procedures toegepast.

5 DISCUSSIE

De hier gepresenteerde methode moet beschouwd worden als een in hoge mate praktische

ontwerpmethode. De benodigde informatie volgt uit de metingen van een eerder bestaand en (te) dicht meetnet. In principe zou dezelfde procedure gevolgd kunnen worden met behulp van concentratievelden zoals berekend met een verspreidingsmodel, jaarlijkse variatie van de meteostatistiek en veronderstelde emissies in Europa. In deze situatie lijkt een efficiënter methode gevormd te kunnen worden door directe combinatie van een beperkter aantal metingen en on-line uit te voeren modelberekeningen. Een voorbeeld hiervan wordt gegeven door Van Egmond en Kesseboom (1983) voor de interpolatie van uurvelden tijdens een winterepisode van verhoogde luchtverontreiniging.

LITERATUUR

- EGMOND, N.D., VAN en ONDERDELINDEN, D.; 1981. Objective analysis of air pollution monitoring network data; spatial interpolation and network density. *Atm. Env.* 15-6, pp. 1035-1046. Ruimtelijke betekenis van luchtverontreinigingsmeetresultaten. RIVM rapport 227905035, Postbus 1, Bilthoven.
- EGMOND, N.D., VAN en KESSEBOOM, H.; 1983. The air pollution monitoring dilemma; measuring or modelling. Proc. 6th Clean Air Congress, Paris, France.
- EGMOND, N.D., VAN, ONDERDELINDEN D. en KESSEBOOM, H.; 1984. An objective procedure for air pollution monitoring network design. In H.J. van der Wiel, K.-H Zierock and N.D. van Egmond (eds.). Monitoring sulphur dioxide, black smoke and suspended particles, Proc. EG workshop Bilthoven, pp. 234-240.
- GANDIN, L.S.; 1965. Objective analysis of meteorological fields. Transl. from Russian: *Isr. Progr. Sci. Transl.*, Jeruzalem, 242.
- MUNN, R.E.; 1981. The design of air quality monitoring networks. MacMillan Publ. Ltd., London, ISBN 0 333 30460 8.

DE FUNDAMENTEN VAN KRIGING EN CO-KRIGING

P.R. Defize

SAMENVATTING

Het doel van ruimtelijke interpolatie in het algemeen is het zo goed mogelijk in kaart brengen van een ruimtelijk verschijnsel, waarover men slechts in beperkte mate informatie bezit. Veelal zullen dit (punt)metingen zijn op een beperkt aantal lokaties. Een voorbeeld hiervan zijn metingen aan bodemmonsters op een mogelijk verontreinigd terrein. Aan de hand van deze metingen zal men pogen om de totale verontreiniging in kaart te brengen, zodat een afschatting kan worden gemaakt van de saneringskosten.

Een ruimtelijke interpolatietechniek die zich in een nog steeds toenemende belangstelling mag verheugen is "kriging". Deze techniek is gebaseerd op een statistisch model dat de eigenschappen van een ruimtelijke variabele beschrijft in termen van verwachtingswaarde en ruimtelijke correlatie. Een uitbreiding van kriging is co-kriging, dat kan worden toegepast als ruimtelijk meerdere variabelen zijn gemeten.

Kriging en co-kriging zijn optimaal in statistische zin. Hierbij moet worden opgemerkt dat deze optimaliteit alleen geldig is als aan de modelveronderstellingen wordt voldaan. Deze zijn in het algemeen moeilijk te controleren.

In dit artikel worden de veronderstellingen waarop kriging en co-kriging zijn gebaseerd nader toegelicht. Tevens worden formules afgeleid voor de interpolator en de interpolatiefout. In hoeverre co-kriging een verbetering kan zijn ten opzichte van kriging wordt onderzocht aan de hand van enige theoretische gevallen, waarvan één afkomstig is uit

de geohydrologie. Uit deze voorbeelden blijkt dat de keuze van de meetpunten een belangrijke factor is.

1 INLEIDING

Kenmerkend voor vele studies op het gebied van water- en bodemkunde is dat gegevens worden verzameld op plekken of lokaties in de 2- of 3-dimensionale ruimte. Voorbeelden hiervan zijn het vochtgehalte van bodemmonsters op een bepaalde diepte ten opzichte van maaiveld, gehalten aan zware metalen in bodemmonsters, seismische reactiesnelheden in een petroleumreservoir, grondwaterstanden of stijghoogtes, hydrologische bodemparameters als K_d -waarde. Aan de hand van dergelijke puntgegevens wil men een ruimtelijk totaalbeeld verkrijgen van het gebied onder studie. Dit doet men door middel van ruimtelijke interpolatie. Er zijn echter vele methoden van ruimtelijke interpolatie en men zal dus een keuze moeten doen. In vele gevallen zal het niet zoveel uitmaken welke methode men kiest, bijvoorbeeld als de waarnemingspunten volgens een vierkant of rechthoekig rooster zijn verdeeld. In andere gevallen, bijvoorbeeld bij clustering van waarnemingspunten, wordt een weloverwogen keuze belangrijk. Tevens is het van belang om de nauwkeurigheid van de interpolatie te kunnen kwantificeren (bijvoorbeeld door middel van de variantie).

De afgelopen 10-15 jaar heeft de kriging-methode erg veel aandacht en toepassingen gekregen. De methode is gebaseerd op statistische principes en is in een bepaalde zin optimaal. Een uitbreiding van kriging is co-kriging. Deze methode kan worden toegepast indien meerdere variabelen zijn gemeten. Door gebruik te maken van de onderlinge samenhang van deze variabelen kan de nauwkeurigheid van de ruimtelijke interpolatie verder worden verbeterd. In dit artikel wordt de theorie van kriging en co-kriging behandeld. Aan de hand van twee theoretische voorbeelden wordt een vergelijking tussen beide methoden gemaakt.

2 RUIMTELIJKE CORRELATIE

Kriging en co-kriging zijn interpolatietechnieken die gebaseerd zijn op een statistisch model. Statistische modellen zijn gebaseerd op veronderstellingen. Of deze veronderstellingen bij

een bepaalde toepassing redelijk zijn zal men dan ook moeten checken alvorens de techniek toe te passen. We behandelen eerst de veronderstellingen die ten grondslag liggen aan de eenvoudigste vorm van kriging (Eng. simple kriging). Daartoe worden eerst enige notaties ingevoerd. Met $Z(\cdot)$ wordt de te interpoleren ruimtelijke variabele aangeduid, met de vector x de lokatie en met h de verschilvector tussen twee lokaties x en $x+h$. De grootheid Z is een stochastische variabele en wordt daarom genoteerd met een hoofdletter.

In formule luiden de veronderstellingen als volgt:

$$E (Z(x+h) - Z(x)) = 0 \quad (1a)$$

$$\text{Var}(Z(x+h) - Z(x)) = 2\gamma(h) \quad (1b)$$

(Het symbool E staat voor verwachtingswaarde).

Wat betekenen nu deze veronderstellingen in concreto? Uit definitie 1a volgt dat $E(Z(x+h)) = E(Z(x))$ voor iedere x en h en dus voor elke lokatie in het 2-dimensionale vlak. Dit is een vrij zware veronderstelling. In het geval van grondwaterstanden bijvoorbeeld zal door eventuele kennis over lokaties van onttrekkingen of ligging van sloten bovenstaande veronderstelling niet meer opgaan.

Veronderstelling 1b zegt iets over de variantie van verschillen. De variantie van het verschil tussen twee Z -waarden (uitgedrukt in het variogram $\gamma(h)$) hangt alléén af van de verschilvector h en niet van de lokatie x . In vrijwel alle praktische toepassingen wordt veronderstelling 1b nog aangescherpt door h te vervangen door $|h|$ (de afstand). Ook deze veronderstelling zal in vele gevallen niet opgaan. Het proces kan bijvoorbeeld in het ene deelgebied minder variabel zijn dan in het andere. Of er zijn verschillen in variabiliteit in de NZ-richting in vergelijking met de OW-richting.

De zwaarte van de veronderstellingen die ten grondslag liggen aan simple kriging zijn in de literatuur ook wel onderkend en er zijn dan ook andere modellen voorgesteld met minder zware veronderstellingen, zoals bijvoorbeeld universal kriging, waarbij een polynomiale trend wordt toegestaan en intrinsic kriging waarbij de rol van het variogram wordt overgenomen door de zogenaamde gegeneraliseerde covariantiefunctie $K(h)$. De theorie van deze uitbreidingen is vanuit mathematisch oogpunt bezien bijzonder elegant en vernieuwend, hetgeen de voornaamste reden lijkt te zijn voor het verschijnen van zoveel theoretische artikelen over kriging en co-kriging.

Met de komst van meer geavanceerde kriging en co-kriging procedures, die ook zijn geïmplementeerd in softwarepakketten als GEO-EAS, BLUEPACK en ARKI, neemt de doorzichtelijkheid van de techniek echter af. De veronderstellingen, waaronder de geavanceerde versies mogen worden toegepast kunnen namelijk alléén door middel van tamelijk gecompliceerde wiskundige formules worden weergegeven en laten zich niet meer door middel van eenvoudige bewoordingen uitleggen. Het gevolg is dat de techniek in toenemende mate wordt toegepast, omdat deze immers 'optimaal' is en geïmplementeerd in softwarepakketten, zonder dat gecontroleerd wordt of aan de (ondoorgroendelijke) onderliggende veronderstellingen (bij benadering) wordt voldaan. Het laatste is in feite ook (nog) niet goed mogelijk. Onderzoek naar en ontwikkeling van nieuwe bruikbare diagnostische middelen (zoals bijvoorbeeld bij regressie) is dan ook belangrijk.

De meest populaire methode voor modelcontrole is cross-validatie. Door op alle meetlokaties de Z-waarde te voorspellen uit metingen of waarnemingen op naburig gelegen lokaties en de voorspelde waarde te vergelijken met de werkelijke waarde kan worden nagegaan of de gekozen kriging-methode 'werkt'.

Met cross-validatie kunnen ook verschillende aanpakken met elkaar worden vergeleken. Of de methode 'werkt' is echter niet equivalent met 'de modelveronderstellingen kloppen'.

3 FORMULES

3.1 Kriging

Deze paragraaf geeft de formules voor de kriging-interpolator of kriging-voorspeller ¹⁾ onder modelveronderstelling 1a en 1b. De kriging-interpolator is een lineaire interpolator: $\hat{Z}(x_0) = \sum \lambda_i Z(x_i)$. Hierbij zijn $Z(x_i)$ ($i=1 \dots n$) de waarnemingen op lokatie i en $\hat{Z}(x_0)$ is de voorspelde waarde op lokatie x_0 . De gewichten λ_i worden zodanig bepaald dat de voorspeller:

- a) zuiver is
- b) een minimale variantie heeft.

¹⁾ Vanuit statistisch oogpunt is de term 'voorspeller' aan te bevelen, omdat het om het schatten van een realisatie van een stochastische variabele gaat. Uit praktisch oogpunt is echter de term 'interpolator' meer op zijn plaats. In dit artikel worden beide termen gebruikt.

Dit resulteert in de volgende kriging-vergelijkingen:

$$\sum_j \lambda_j \gamma(x_i, x_j) + \Psi = \gamma(x_i, x_0) \quad i=1 \dots n$$

$$\sum_j \lambda_j = 1 \quad (2)$$

Dit is een lineair stelsel van $(n+1)$ vergelijkingen met $(n+1)$ onbekenden (n gewichten λ_j + de Lagrange factor Ψ).

De kriging-variantie (de variantie van de voorspeller $\hat{Z}(x_0)$) is gelijk aan:

$$\text{var}(\hat{Z}(x_0) - Z(x_0)) = \sum_i \lambda_i \gamma(x_i, x_0) + \Psi \quad (3)$$

3.2 Kruiscorrelatie en co-kriging

Stel dat men niet alleen beschikt over waarnemingen aan de variabele Z maar ook over waarnemingen aan een 2^e ruimtelijke variabele Y . Met behulp van co-kriging kan de voorspelling van Z worden verbeterd door gebruik te maken van de covariantie tussen Z en Y :

$$C_{1,2}(h) = E[Z(x+h)Y(x)] - E(Z(x+h))E(Y(x)) \quad (4)$$

De covariantiefunctie is direct gerelateerd aan het kruisvariogram:

$$\gamma_{1,2}(h) = \frac{1}{2} E[(Z(x+h) - Z(x))(Y(x+h) - Y(x))] \quad (5)$$

Uitgaande van isotropie en de afwezigheid van trend kan worden afgeleid dat:

$$\gamma_{1,2}(h) = C_{1,2}(0) - C_{1,2}(h) \quad (6)$$

De co-kriging voorspeller op een onbekende lokatie x_0 is een lineaire combinatie van de waarnemingen aan Z op lokaties x_{1i} ($i=1 \dots n$) en Y op lokaties x_{2j} ($j=1 \dots m$):

$$\hat{Z}_c(x_0) = \sum_i \lambda_i Z(x_{1i}) + \sum_j \mu_j Y(x_{2j}) \quad (7)$$

Onder dezelfde eisen als bij kriging kunnen λ_i en μ_j worden bepaald uit de volgende co-kriging vergelijkingen:

$$\sum_i \lambda_i = 1$$

$$\sum_j \mu_j = 0$$

$$\sum_{r=1}^n \lambda_r \gamma_1(x_{1i}, x_{1r}) + \sum_{s=1}^m \mu_s \gamma_{1,2}(x_{1i}, x_{2s}) + \Psi_1 = \gamma_1(x_0, x_{1i}) \quad i=1 \dots n$$

$$\sum_{r=1}^n \lambda_r \gamma_{1,2}(x_{1r}, x_{2j}) + \sum_{s=1}^m \mu_s \gamma_2(x_{2j}, x_{2s}) + \Psi_2 = \gamma_{1,2}(x_0, x_{2j}) \quad j=1 \dots m \quad (8)$$

In deze formules zijn γ_1 en γ_2 de variogrammen van respectievelijk Z en Y. Er zijn $n+m+2$ vergelijkingen met evenveel onbekenden. Ψ_1 en Ψ_2 zijn weer Lagrange-factoren.

De co-kriging variantie is gelijk aan:

$$\begin{aligned} \text{Var}(\hat{Z}_c(x_0) - Z(x_0)) = \\ \sum_i \lambda_i \gamma_1(x_0, x_{1i}) + \sum_j \mu_j \gamma_{1,2}(x_0, x_{2j}) + 2 \sum_j \mu_j \gamma_2(x_0, x_{2j}) + \Psi_1 \end{aligned} \quad (9)$$

3.3 Blok-kriging

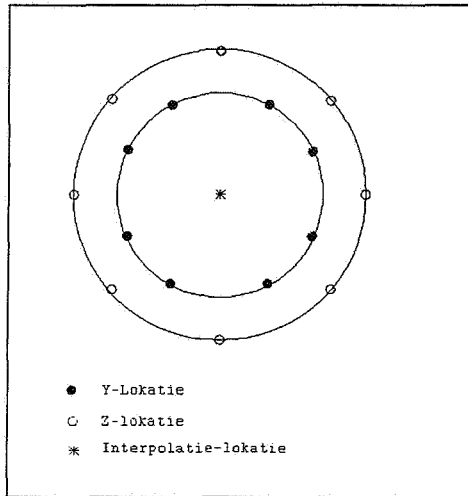
In sommige toepassingen is men niet geïnteresseerd in voorspellingen op specifieke lokaties, maar is de gemiddelde waarde van Z over een bepaald gebied van belang. De techniek die men hiervoor kan gebruiken heet blok-kriging (of blok-co-kriging). De vergelijkingen uit 3.1 en 3.2 dienen dan enigszins te worden aangepast. De variantie van de interpolatiefout zal afnemen (afhankelijk van de blok grootte en de correlatiestructuur).

4 VERGELIJKING KRIGING VERSUS CO-KRIGING

Op de vraag in hoeverre co-kriging een verbetering kan geven t.o.v. kriging kan moeilijk in algemene zin een antwoord worden gegeven. Uiteraard zal de functie $\gamma_{1,2}(h)$ een belangrijke invloed hebben, maar evenzo speelt het aantal lokaties waar Y gemeten is en de ligging van deze lokaties een belangrijke rol. De beste methode om e.e.a. vast te stellen is uitproberen. Leenaers (1991) geeft hiervan een voorbeeld. In deze bijdrage behandelen we 2 theoretische situaties.

4.1 Z- en Y-lokaties symmetrisch op een cirkel met de te interpoleren lokatie als middelpunt

De situatie is geschetst in figuur 1. Als men aan dit voorbeeld gaat rekenen, komt men tot de op het eerste gezicht verrassende conclusie dat co-kriging géén enkele verbetering geeft t.o.v. gewoon kriging, onafhankelijk van het veronderstelde kruisvariogram. Uit symmetrie-overwegingen volgt namelijk dat $\mu_1 = \mu_2 = \dots = \mu_m$. Gecombineerd met de co-kriging-eis dat $\sum \mu_j = 0$ volgt $\mu_1 = \mu_2 = \dots = \mu_m = 0$, dus géén bijdrage van de Y-waarnemingen. Het uitzoeken van andere theoretische situaties zou tot nog meer verrassende resultaten kunnen leiden en tevens bruikbare 'vuistregels' kunnen opleveren voor het opzetten van meetnetten. Zoals uit het volgende voorbeeld zal blijken is het bijvoorbeeld erg belangrijk dat lokaties voorkomen, waar Z en Y beiden gemeten worden.



Figuur 1 Meetlokaties op een cirkel

4.2 Simultane interpolatie van de stijghoogte in meerdere watervoerende pakketten

Defize (1987) heeft een pilotstudie verricht naar de mogelijkheden van co-kriging bij het interpoleren van de stijghoogte in meerdere watervoerende pakketten. Uitgangspunt van de studie was een artikel van Hoeksema en Kitanidis (1984), waarin co-kriging is toegepast op transmissiviteit- en stijghoogtemetingen. Het bijzondere van deze toepassing is dat een statistisch en een deterministisch model worden gecombineerd. Het variogram van de stijghoogte en het kruisvariogram tussen stijghoogte en transmissiviteit wordt namelijk afgeleid uit het deterministisch verband dat tussen beide grootheden bestaat (via de wet van Darcy en de continuïteitsvergelijking). Een vergelijkbare aanpak is die van Van Geer (1991).

Defize heeft de methode uitgewerkt voor 2 watervoerende pakketten met bekende eigenschappen (transmissiviteit, laagdikte) van de scheidende laag. Voor de ruimtelijke

verdeling van de stijghoogte in het bovenste pakket is het volgende variogram verondersteld

$$\gamma(h) = \alpha(1-\exp(\Theta|h|)) \quad (10)$$

De parameter α is een schaalparameter en is voor het gemak op 1 gesteld. Voor Θ geldt: $\Theta < 0$.

De volgende vergelijking beschrijft het verband tussen de stijghoogtes in beide pakketten:

$$kD \left(\frac{\delta^2 \delta_0}{\delta^2 u^2} + \frac{\delta^2 \delta_0}{\delta^2 v^2} \right) = \frac{\delta_1 - \delta_0}{c} \quad (11)$$

waarbij :

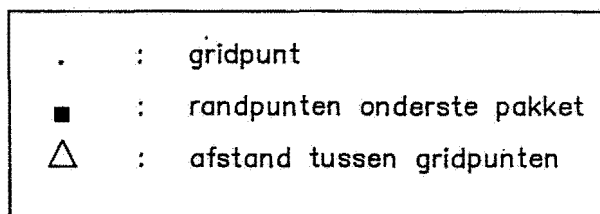
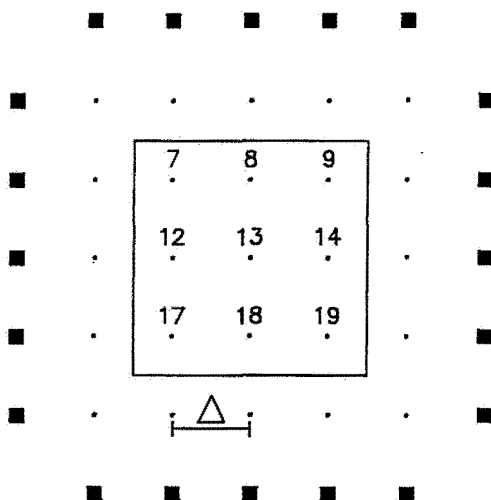
- δ_0 : stijghoogte in onderste pakket
- δ_1 : stijghoogte in bovenste pakket
- kD : transmissiviteit
- c : weerstand in de semi-permeabele laag
- u, v : coördinaten

Met behulp van deze vergelijking is het semivariogram van ϕ_0 en het kruisvariogram van δ_0 en δ_1 vastgesteld.

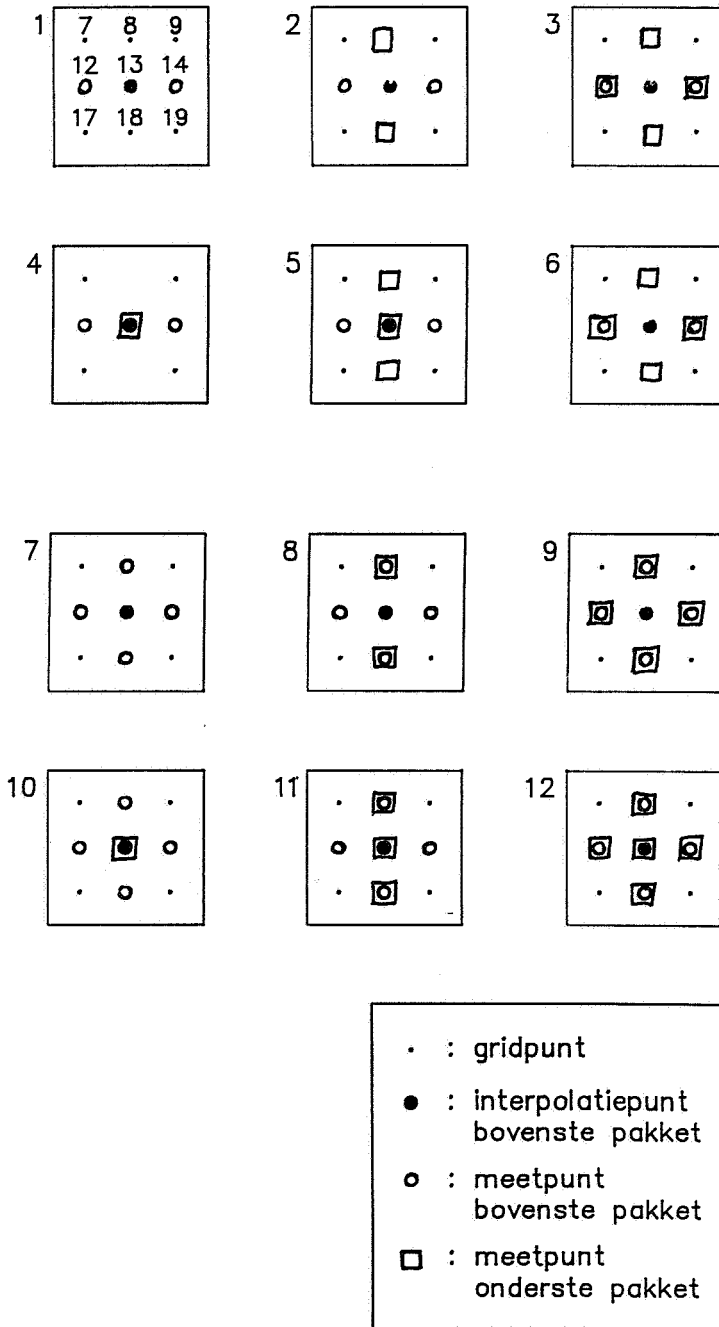
4.2.1 Co-kriging in het bovenste pakket

Als uitgangspunt is een vierkant netwerk met 25 roosterpunten gekozen (figuur 2). De stijghoogtes in het onderste pakket op de 20 randpunten worden bekend verondersteld. Verder wordt steeds geïnterpoleerd in het middelste punt van het rooster (nr. 13). Een 12-tal schema's is in de studie geanalyseerd. Uitgaande van de nummering in figuur 2 zijn deze schema's opgenomen in figuur 3.

BASIS NETWERK



Figuur 2 Basis netwerk



Figuur 3 Meetschema's interpolatie bovenste pakket

De schema's laten zich als volgt beschrijven:

Schema's 1 t/m 6 hebben alle 2 meetpunten in het bovenste pakket (lokatie 12 en 14), terwijl de plaats van en het aantal meetlokatie in het onderste pakket is gevarieerd. De schema's 7 t/m 12 bevatten 2 extra meetpunten (nrs 8 en 18) in het bovenste pakket in vergelijking met de schema's 1 t/m 6. Schema's 1 en 7 hebben géén meetpunten in het onderste pakket, hetgeen inhoudt dat interpolatie met dit schema neerkomt op kriging in plaats van co-kriging.

Voor de roosterafstand is een waarde van 600 m gekozen.

De waarden van kDc (het product van kD en c , ofwel het kwadraat van de karakteristieke lengte) en Θ zijn gevarieerd en wel als volgt:

kDc : 25000, 50000, 100000, 200000 m^2

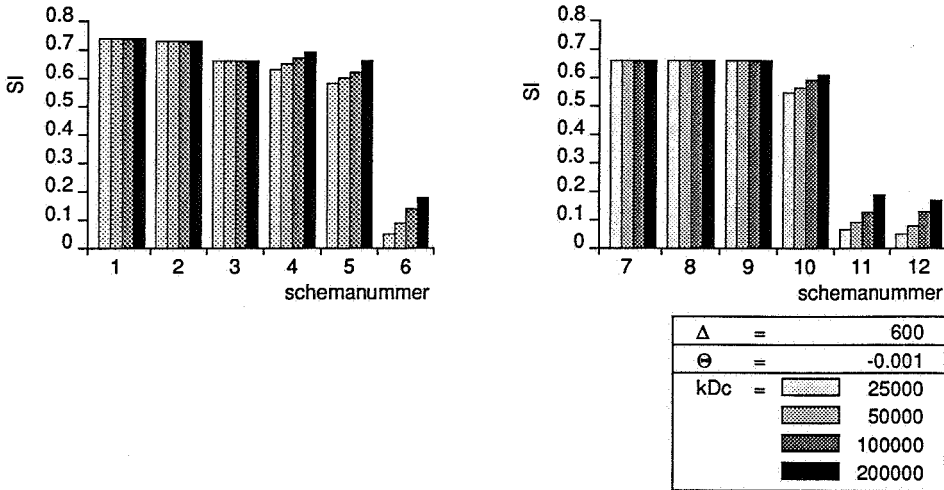
Θ : -0.0005, -0.001 en -0.002 m^{-1}

Als criterium voor de nauwkeurigheid van de interpolatie is de standaardafwijking van de interpolatiefout (SI) gehanteerd.

In figuur 4 is de SI uitgezet tegen het schanummer bij een vaste waarde van Θ (-0.001) voor vier verschillende waarden van kDc . De volgende zaken vallen op.

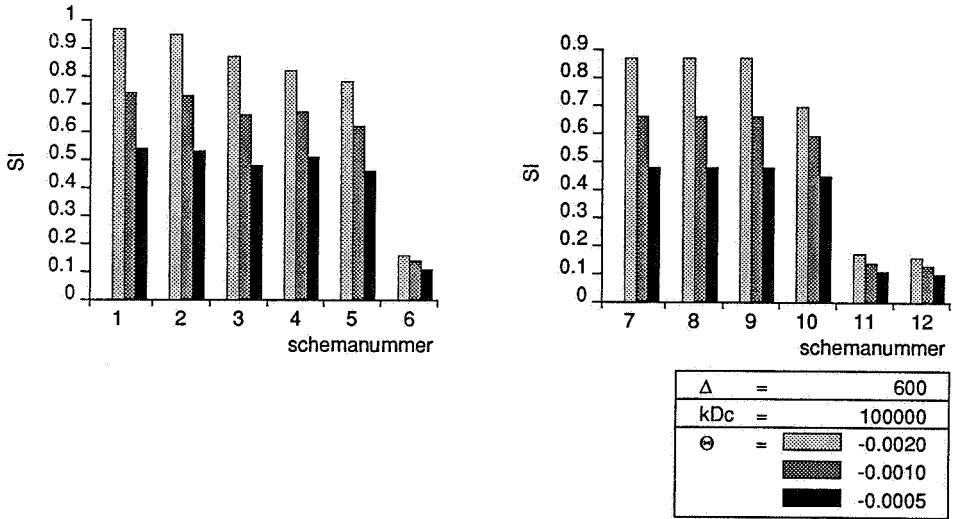
- De SI is relatief klein bij de schema's 6, 11 en 12. Uniek voor deze schema's t.o.v. de andere is dat er naast een meetpunt op locatie 13 in het onderste pakket, ook locaties voorkomen met meetwaarden in beide pakketten.

Dat de interpolatie nauwkeuriger wordt is ook logisch aangezien a) het min of meer constante niveauverschil in stijghoogte tussen beide pakketten goed kan worden geschat en b) dit niveauverschil bij de waarneming op locatie 13 kan worden opgeteld om een goede schatting van de stijghoogte in het bovenste pakket te krijgen.



Figuur 4 Resultaten co-kriging bovenste pakket, vaste waarde voor Θ

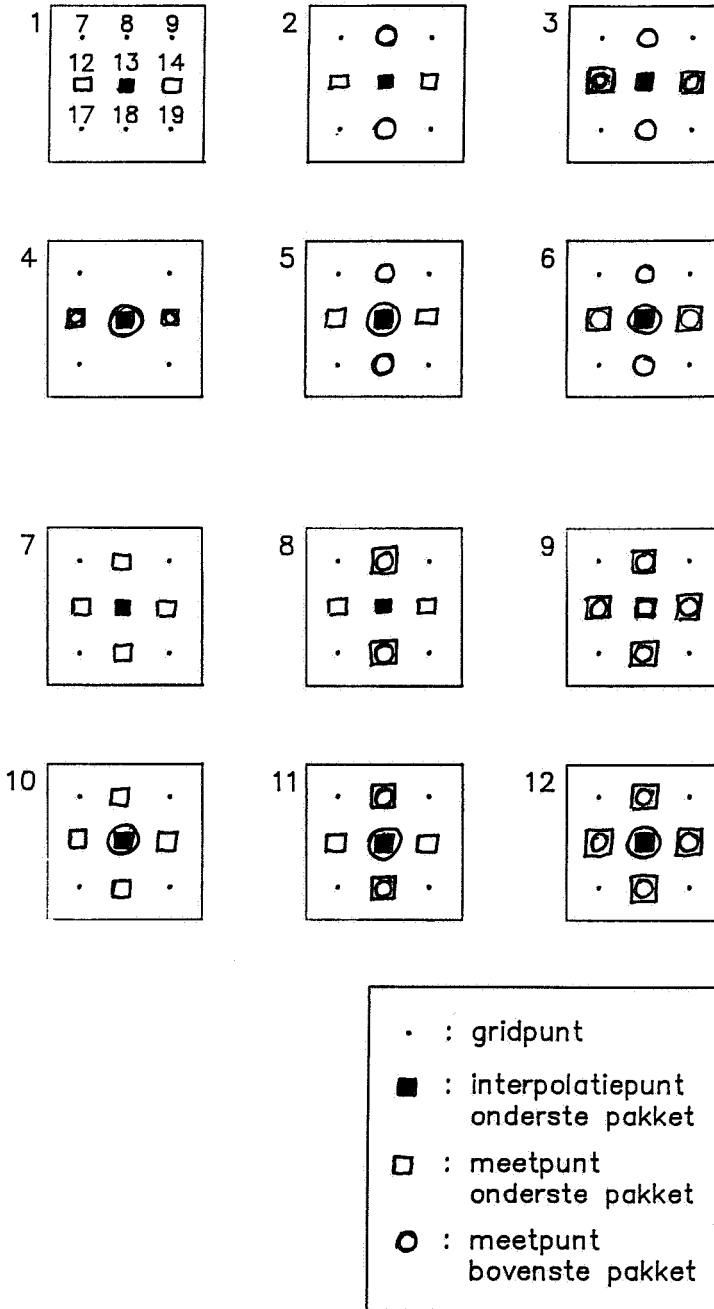
- Uiteraard neemt de SI af bij de schema's 2 t/m 6 t.o.v. schema 1 (en schema's 8 t/m 12 t.o.v. schema 7), maar deze afname is slechts van enige importantie voor die schema's waarbij op lokatie 13 een meting is gedaan in het onderste pakket (schema's 4, 5, 6, 10, 11, 12).
- De SI wordt groter naarmate de waarde van kDc toeneemt. Ook dit is logisch aangezien bij een toenemende kDc-waarde de doorlatendheid van de watervoerende laag toeneemt t.o.v. de doorlatendheid van de scheidende laag en informatie uit het onderste pakket minder informatief wordt. Hierbij dient wel te worden opgemerkt dat in het statistische model de mate van onzekerheid volledig bepaald wordt door de modellering van de onzekerheid in het bovenste pakket.
- De invloed van de waarde van Θ is af te lezen uit figuur 5. Naarmate Θ kleiner wordt (in absolute waarde) is er sprake van meer ruimtelijke samenhang. Dat de SI afneemt met een afnemende Θ (in absolute waarde) is dan ook logisch.



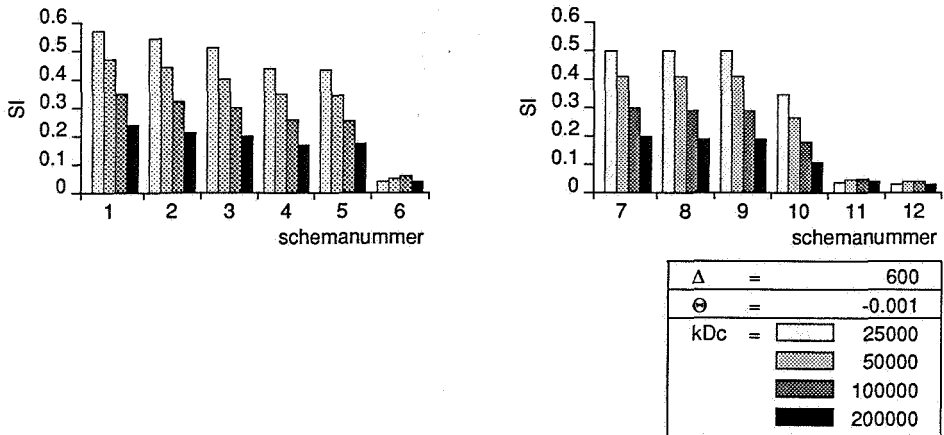
Figuur 5 Resultaten co-kriging bovenste pakket, vaste waarde voor kDc

4.2.2 Co-kriging in het onderste pakket

De aanpak is dezelfde als bij de interpolatie in het bovenste pakket met dit verschil dat de liggingen van de meetlokaties in het bovenste en onderste pakket zijn verwisseld. Op dezelfde wijze als bij de interpolatie in het bovenste pakket zijn de resultaten samengevat in de figuren 6 en 7.

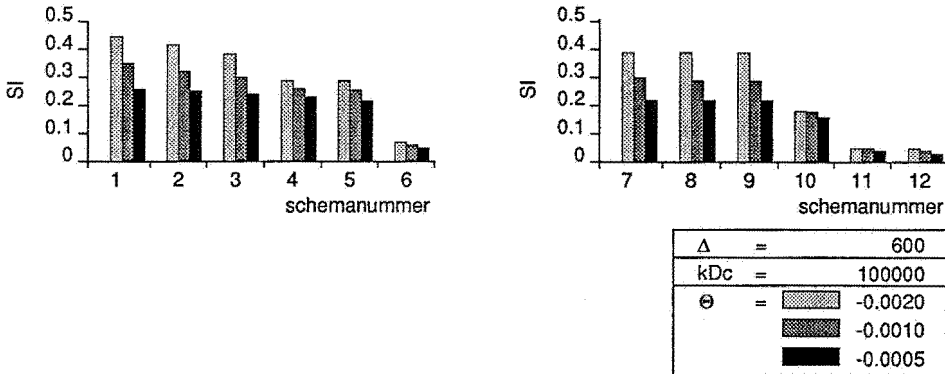


Figuur 6 Meetschema's interpolatie onderste pakket



Figuur 7 Resultaten co-kriging onderste pakket, vaste waarde voor Θ

Er treden opvallende verschillen op in de resultaten. Dat was ook niet op voorhand te verwachten, aangezien de rol van onderste en bovenste pakket niet symmetrisch is. Ten eerste is vergelijking 11 niet symmetrisch en ten tweede is een ruimtelijk stochastisch model verondersteld voor het bovenste pakket en niet voor het onderste : dat is immers afgeleid uit vergelijking 11. Een gevolg hiervan is dat over het algemeen de SI juist afneemt met een toenemende waarde van kDc. Dit volgt direct uit het veronderstelde model. Bij een zeer grote waarde van kDc (ondoorlatende laag) is het model voor het onderste pakket volledig deterministisch en zal de interpolatiefout 0 worden. Merk overigens op dat voor de schema's met een meting in het bovenste pakket op lokatie 13 (schema's 4, 5, 6, 10, 11 en 12) de interpolatiefout op lokatie 1 ook 0 wordt bij een kDc-waarde 0. In dat pakket zijn de stijghoogtes in beide pakketten immers gelijk. Hiermee is een verklaring gegeven voor het verloop van de SI bij de schema's 6, 11 en 12. Bij deze schema's loopt de SI namelijk eerst op en dan weer af. De maximale SI ligt bij een kDc-waarde tussen 25000 en 100000. De SI neemt ook hier af bij een toenemende ruimtelijke samenhang (figuur 8).



Figuur 8 Resultaten co-kriging onderste pakket, vaste waarde voor kDc

5 CONCLUSIE

Kriging en co-kriging zijn geschikte technieken voor ruimtelijke interpolatie. Ten gevolge van de complexiteit van het onderliggende statistisch model zijn modelveronderstellingen moeilijk te controleren. In theorie is de interpolatiefout te kwantificeren door middel van de kriging- of co-kriging-variantie. Of de uitkomsten van variantieberekeningen 'kloppen' zal men in praktische toepassingen echter altijd moeten nagaan via cross-validatie.

Co-kriging kan een verbetering opleveren ten opzichte van co-kriging indien

- a) er een goede kruiscorrelatie bestaat tussen de te interpoleren variabele en de co-variabele
- b) een slimme keuze van de meetlokaties is gedaan.

LITERATUUR

- DEFIZE P.R.; Simultane interpolatie van de stijghoogte in meerdere watervoerende pakketten met behulp van co-kriging. DGV-TNO. Rapportnr. 87 GM 014.
- GEER, F.C. VAN; 1991. Kalman filtering bij ruimtelijke interpolatie van grondwaterstijghoogten. CHO-TNO Rapporten en nota's No. 24.
- HOEKSEMA, R.J. and KITANIDIS, P.K.; 1984. An application of the Geostatistical Approach to the Inverse Problem in Two-Dimensional Groundwater Modelling. Water Resources Research, p. 1003-1020.
- LEENAERS, H.; 1991. Het karteren van bodemverontreiniging in het Geuldal door co-kriging met aanvullende hoogte-informatie. CHO-TNO Rapporten en nota's No. 24.

HET KARTEREN VAN BODEMVERONTREINIGING IN HET GEULDAL DOOR CO-KRIGING MET AANVULLENDE HOOGTE-INFORMATIE

H. Leenaers

SAMENVATTING

De oevergronden van de Geul zijn in hoge mate verontreinigd met lood, zink en cadmium door de aanvoer van verontreinigd slib tijdens overstromingen. De mate van verontreiniging vertoont een sterke samenhang met de geomorfologie en de hoogte ten opzichte van de rivierbedding. Deze laatste parameter werd gebruikt als co-variabele bij het karteren van de zinkverontreiniging met behulp van co-kriging. De co-kriging-voorspellingen van het zinkgehalte in de bodem bleken aanmerkelijk beter dan de kriging-voorspellingen zonder gebruik van de hoogtegegevens. Bovendien bleek de bijbehorende co-kriging-variantie niet alleen stabiel van aard, maar ook veel lager (10-30 %) dan de kriging-variantie.

1 INLEIDING

Metaalmijnbouw in het Belgische deel van het stroomgebied van de Geul is de oorzaak van de verontreiniging van haar oevergronden met lood, zink en cadmium. Ondanks het feit dat de mijnen enige tientallen jaren geleden werden gesloten, gaat de inbreng van zware metalen in dit riviersysteem nog steeds door. Uitloging van de afvalbergen en aantasting van de verontreinigde oevers door de meanderende beek zijn op dit moment de belangrijkste bronnen van zware metalen. De op die manier in de bedding gebrachte verontreinigde sedimenten worden tijdens hoogwaterperioden weer op de oevers afgezet.

De verontreiniging van de oevergronden kan leiden tot het nemen van beperkende

maatregelen t.a.v. het landgebruik. Om die reden is het van belang verontreinigingskaarten te vervaardigen die een nauwkeurig beeld geven van de ligging van de meest verontreinigde delen van de dalbodem. Bij het karteren van bodemverontreiniging in rivierdalen werd in het verleden vaak gebruik gemaakt van de relaties tussen enerzijds de bodemkundige en geomorfologische kenmerken van een gebied, en anderzijds de mate van verontreiniging. Bodemkaarten en geomorfologische kaarten geven immers een beeld van de mate van sedimentatie en het type sediment, en daarmee ook van het ruimtelijk patroon van de verontreiniging met zware metalen of andere sedimentgebonden componenten (Rang et al., 1987; Leenaers & Okx, 1989). De op die wijze vervaardigde choropletenkaarten doen echter geen recht aan het vaak geleidelijk variërend karakter van bodemverontreiniging. De genoemde terreineigenschappen (bodemtype, geomorfologie, maar ook overstromingsfrequentie) zijn allen gerelateerd aan de hoogteligging t.o.v. de rivierbedding. Het verzamelen van hoogte-informatie is betrekkelijk eenvoudig en goedkoop en om die reden werd getracht bij het karteren van de geleidelijke variaties van de bodemverontreiniging gebruik te maken van aanvullende hoogte-informatie. De in het vorige hoofdstuk besproken interpolatietechniek co-kriging (Defize, 1991) maakte het mogelijk deze aanvullende informatie te gebruiken bij het voorspellen van zinkgehalten in de bodems van het Geuldal. De resultaten worden o.a. vergeleken met die van point kriging zonder hoogte-informatie.

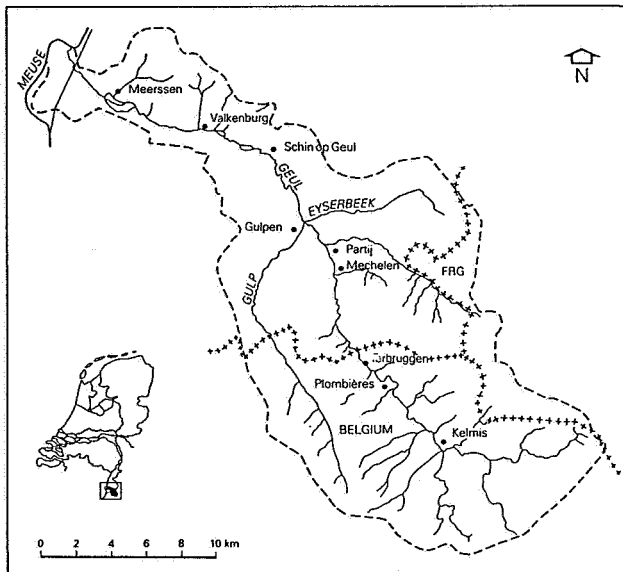
2 HET STROOMGEBIED VAN DE GEUL

2.1 Hydrologische kenmerken

De Geul ontspringt in de gemeente Eynatten (België), nabij de Duits-Belgische grens, en mondt 56 km stroomafwaarts en 242 m lager uit in de Maas nabij Bunde (zie Figuur 1). Van de 350 km² die het stroomgebied omvat, liggen er 180 in Nederland. De breedte van de dalbodem varieert van minder dan 100 m in België tot meer dan 600 m in Nederland. Het oppervlak van de fluviaal beïnvloede gronden in Nederland bedraagt ca. 12 km². De gemiddelde afvoer van de Geul bedraagt 1 m³/s bij de grensovergang tussen Nederland en België, en neemt toe tot 3 m³/s in Meerssen.

Het optreden van hoogwater hangt sterk samen met de hoeveelheid neerslag in het stroomgebied. Binnen enkele uren na een regenbui neemt de afvoer toe en de maximale waterstand kan vaak binnen 24 uur worden geregistreerd. De hoogste afvoeren die ooit

werden gemeten variëren van 30 m³/s bij de grens tot 60 m³/s bij de monding. Gedurende hoogwaterperioden is de erosieve kracht van het snelstromende water groot, zodat de oevers van de Geul plaatselijk worden aangetast. De sedimentlast kan daardoor toenemen tot enkele grammen per liter. Op plaatsen waar het water wordt opgestuwd, bijv. in de Kapolder waar de Gulp en de Eijserbeek uitmonden in de Geul, treedt de beek buiten haar oevers. Tijdens overstromingen neemt de stroomsnelheid van het water sterk af en wordt het zwevend slib op de oevers afgezet.



Figuur 1 Het stroomgebied van de Geul

2.2 Gevolgen van de metaalmijnbouw

De eerste vermelding van de winning van lood- en zinkerts in het Belgische deel van het stroomgebied van de Geul dateert uit 1270. De belangrijkste mijnen lagen bij de plaatsen Plombières en Kelmis (zie Figuur 1), waar de hoogtijdagen van de mijnactiviteit werden bereikt in de periode 1820-1880. In de periode 1855-1881 werd in Plombières 97.543 ton galeniet, 100.226 ton blende en 502 ton zinkerts gewonnen. De ertswinning in de dagbouw-mijnen van Kelmis bedroeg ca. 135.000 ton per jaar en waren er 1300 mensen werkzaam in de mijnindustrie. Vanaf 1837 werd hier ca. 1.4 miljoen ton zinkerts gewonnen. De laatste

mijn sloot in 1938.

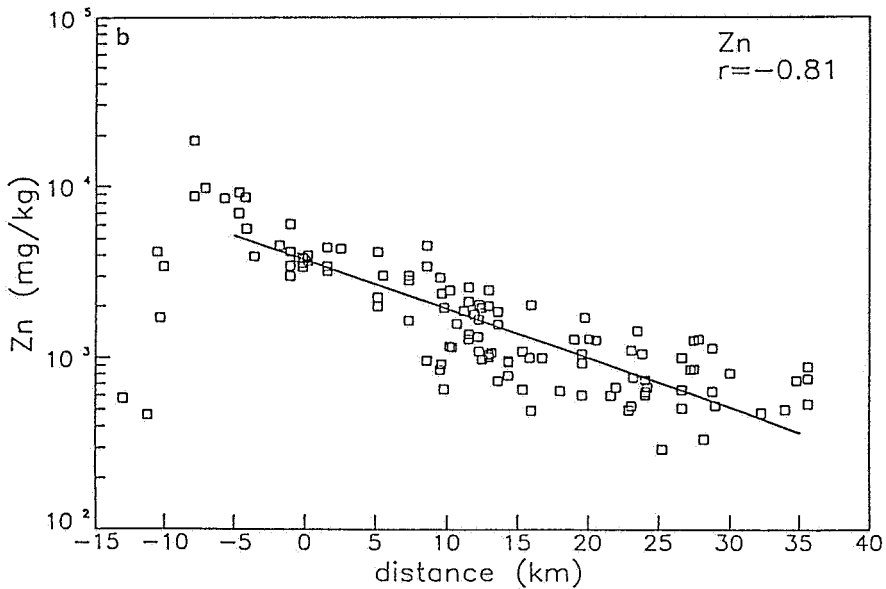
Door de aanwezigheid van de mijnen en de ertsverwerkende industrie zijn in de loop der tijd grote hoeveelheden lood, zink en cadmium in de Geul terechtgekomen als gevolg van (1) lozingen van drainagewater en afvalwater met hoge concentraties opgeloste metalen en ertsdeeltjes, (2) erosie en uitspoeling van de - nog steeds aanwezige - afvalbergen op de oevers en (3) de verwerking van zwavelhoudende metaalverbindingen door blootstelling aan de lucht.

De zware metalen worden - deels opgelost, deels gebonden aan het slib en deels in minerale vorm - door de Geul in stroomafwaartse richting getransporteerd en tijdens overstromingen afgezet op de oevers. De oevergronden, die de Geul in de huidige situatie aantast tijdens het meanderen in haar eigen afzettingen, bestaan dan ook uit verontreinigde sedimenten. De oorspronkelijke puntbron van zware metalen, het mijngebied in België, is verworpen tot een min of meer diffuse lijnbron, de huidige oevergronden.

2.3 Verontreinigingspatronen

In het verleden uitgevoerd onderzoek heeft inzicht verschaft in de globale patronen van bodemverontreiniging in het Geuldal. Het in dit artikel beschreven onderzoek is het logisch vervolg daarop en om die reden volgt hieronder een korte samenvatting.

Op de schaal van het gehele rivierdal kan in stroomafwaartse richting een sterke afname in de mate van verontreiniging met zware metalen worden waargenomen (Figuur 2). Dit patroon kan het best worden beschreven met een exponentiële functie en kan worden verklaard uit het optreden van 'verdunning' met relatief schone sedimenten tijdens het transport (Leenaers et al., 1988). Indien we de schaal van observatie vergroten naar die van een overstromingsgebied, dan blijken er lokale afwijkingen van dit globale patroon op te treden die het gevolg zijn van variabele stromingscondities tijdens een overstroming en het lokale reliëf van de overstromingsvlakte (Figuur 3). Door gebruik te maken van een Digitaal Terrein Model werd een overstromingsfrequentiekaart gemaakt van een deel van het Geuldal. Deze kaart vormde de basis voor choropletenkaarten van de bodemverontreiniging met hoge percentages verklaarde variantie (bijvoorbeeld voor Zn: 65 %) (Leenaers & Okx, 1989). Dit resultaat illustreert het belang van het verband tussen de hoogteligging en de mate van verontreiniging, en vormde de aanleiding tot het onderhavige onderzoek waarin het continue verloop van de bodemverontreiniging wordt gekarteerd met gebruikmaking van aanvullende hoogtegegevens.

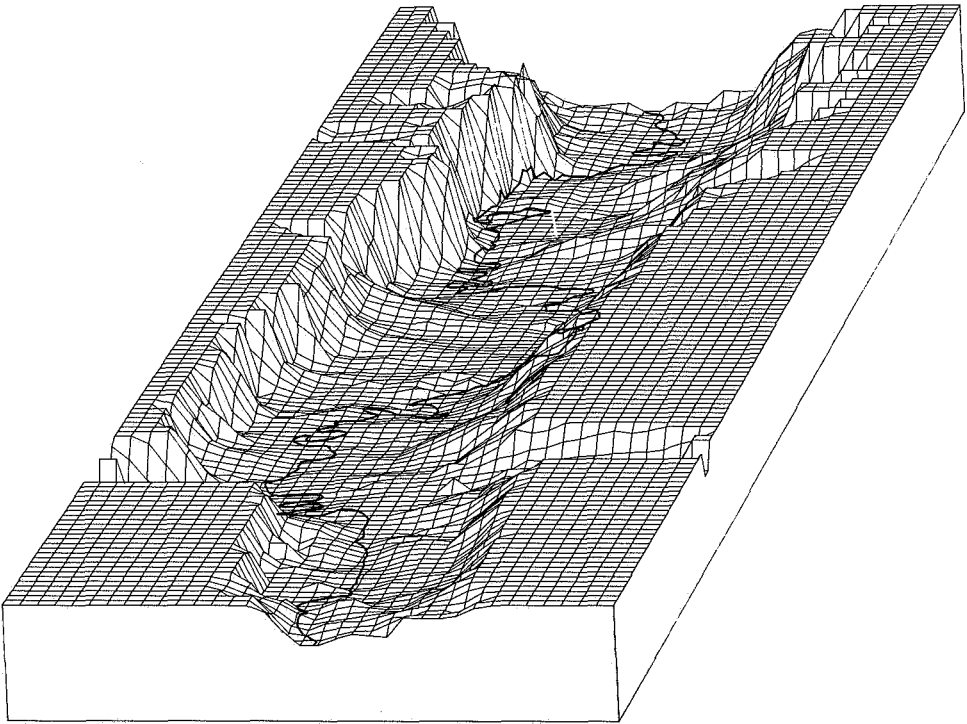


Figuur 2 Stroomafwaartse veranderingen in de Zn-gehalten van recent afgezet Geulslib

3 WERKWIJZE

3.1 Bemonstering en analyse

In een deel van het Geuldal met een omvang van ca. 2,5 km² werden op 199 lokaties bodemonsters genomen van de toplaag van de bodem (0-10 cm diepte). De bemonsteringsafstand varieerde van 50-100 m loodrecht op de dalrichting tot 200-400 m evenwijdig aan de dalrichting. De lokatiekeuze vond plaats door 'stratified random sampling', waarbij de bodemkaart van Van de Westeringh (1980) als uitgangspunt werd gebruikt. Monsterpunten in het deel van het onderzoeksgebied waarvoor ook verontreinigingskaarten zullen worden gepresenteerd zijn weergegeven in Figuur 4. In het laboratorium werden de monsters gedroogd en vermalen, ontsloten met 30 % HNO₃ en werden de zinkgehalten gemeten op de A.A.S.



Figuur 3 Digitaal Terrein Model van een deel van het Geuldal

3.2 Gegevensverwerking

Van alle monsterpunten was de hoogteligging ten opzichte van N.A.P. bekend. De hoogten opzichte van de bedding werd berekend door een correctie te maken voor het lengteprofiel. De trend in het zinkgehalte was over het gebruikte afstandsinterval niet statistisch significant. In aanvulling op deze 199 relatieve hoogtes werden 309 hoogtepunten van de Topografische Kaart (1 : 10.000) gedigitaliseerd en omgezet naar relatieve hoogtes.

Het gegevensbestand van 199 punten waar zowel het zinkgehalte (Zn) en de relatieve hoogte (RE) bekend waren, werd in twee delen gesplitst. Het grootste deel (154) werd gebruikt voor het berekenen van de variogrammen, de kartering van Zn m.b.v. kriging en het berekenen van een regressievergelijking tussen Zn en RE. Voor het karteren van Zn m.b.v. co-kriging werd dit bestand aangevuld met de beschikbare 309 hoogtepunten.

Tabel 1 Centrum- en spreidingsmaten van zinkgehalten in de bodem (mg/kg) en relatieve hoogten (cm) (n=monsteraantal)

	n	gem.	variantie	mediaan	range
zink concentratie (voor interpolatie)	154	741	292877	543	114-2270
zink concentratie (voor validatie)	45	780	276497	67	117-1437
relatieve hoogte (voor interpolatie)	463	42	5112	411	224- 791
relatieve hoogte (voor validatie)	45	432	4366	427	260- 609

De overige 45 punten werden gebruikt voor het valideren van de ruimtelijke modellen. Bij het karteren van Zn m.b.v. kriging en co-kriging werd geïnterpoleerd naar een raster van 1,4 x 5,0 km met cellen van 40 x 40 m.

Tabel 1 geeft een aantal beschrijvend statistische maten voor Zn en RE en Tabel 2 geeft de parameters van de regressievergelijkingen tussen Zn en RE.

Tabel 2 Correlatie coëfficiënten en parameters van de lineaire relaties ($Y=a+bX$) tussen het zinkgehalte (Zn) en de relatieve hoogte (RE) (n=154)

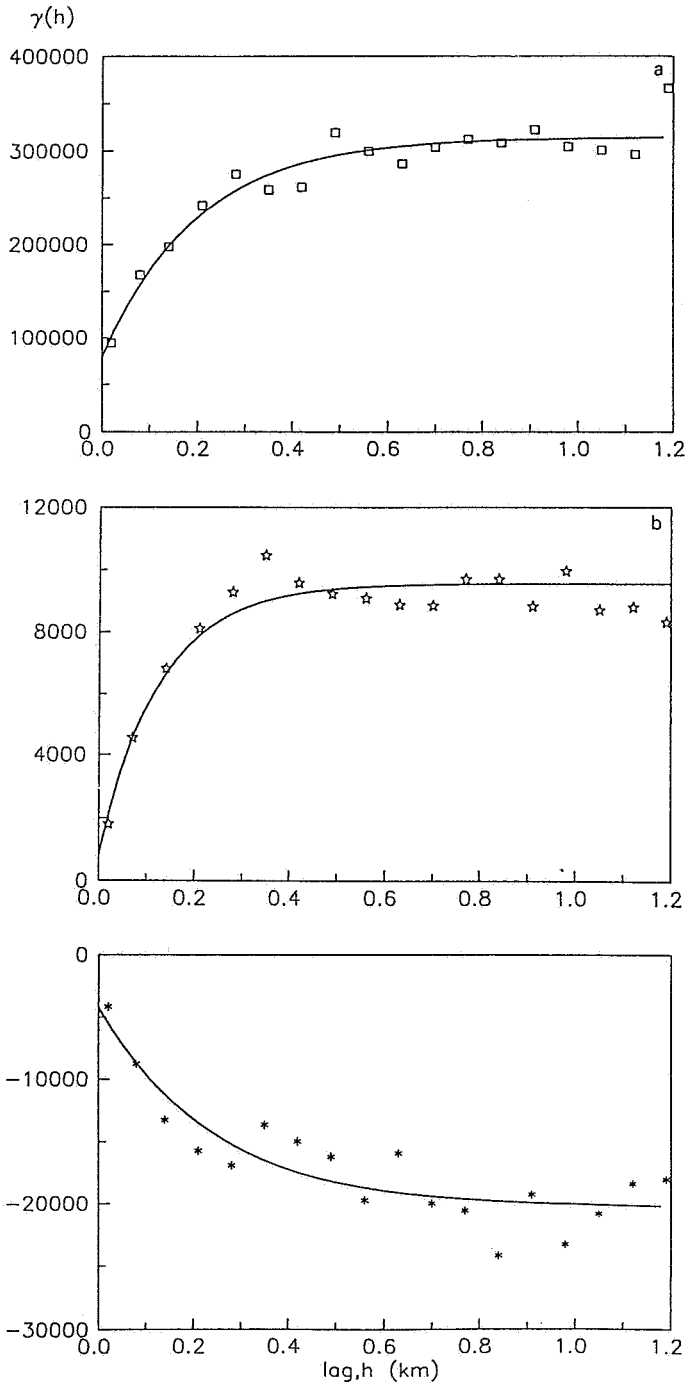
Y	X	a	b	r
Zn	RE	1617	-206	-0.38
$\log_{10}(\text{Zn})$	RE	3.48	-0.17	-0.48
Zn	$\log_{10}(\text{RE})$	5986	-2004	-0.36
$\log_{10}(\text{Zn})$	$\log_{10}(\text{RE})$	7.19	-1.70	-0.46

4 HET RUIMTELIJK VERBAND TUSSEN Zn-GEHALTE EN HOOGTELI- GING

4.1 De variogrammen

De experimentele variogrammen voor Zn en RE zijn weergegeven in Figuur 5. Ieder symbool vertegenwoordigt een afstandsinterval van 0.07 km. Het fitten van exponentiële modellen gebeurde met behulp van een gewogen kleinste kwadraten-benadering en een kruisvalidatieprocedure zoals beschreven door Cooper en Istok (1988). De parameters van de gefitte variogrammodellen staan weergegeven in Tabel 3. De variabiliteit van het zinkgehalte is erg groot en leidt tot een sill-waarde van ca. 300.000. Het nugget-effect bedraagt ca. 25 % van die waarde en dat is aanmerkelijk meer dan de 9 % nugget voor de relatieve hoogte (RE). De range van beide variogrammen ligt in dezelfde orde van grootte en bedraagt ca. 500 m. Het kruisvariogram van Zn en RE is negatief vanwege het inverse verband tussen beide eigenschappen (zie Tabel 2).

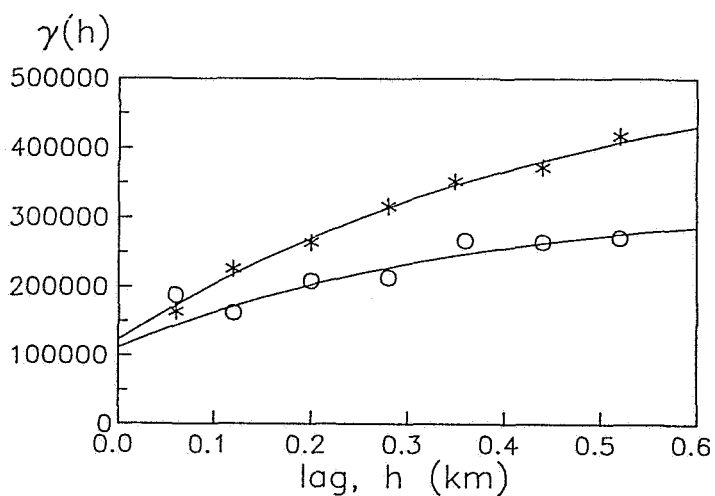
De variogrammen in Figuur 5 beschrijven isotope variaties, d.w.z. de ruimtelijke correlatiestructuur vertoont geen richtingseffecten. In een rivierdal waar stroomrugronden en komgronden voorkomen parallel aan de dalrichting, lijkt het waarschijnlijk dat de mate waarin bodemeigenschappen variëren wel degelijk een richtingsafhankelijkheid zal vertonen: grote korte afstandsvariaties loodrecht op het dal en geleidelijke overgangen parallel aan het dal. Om die reden werden eveneens variogrammen berekend voor deze twee hoofdrichtingen. Het anisotopie variogram voor zink staat weergegeven in Figuur 6 en bevestigt het geschetste beeld. Het variogram berekend loodrecht op de dalrichting vertoont een veel steiler verloop dan dat berekend parallel aan de dalrichting. Bovendien blijkt de sill-waarde voor het eerstgenoemde variogram veel hoger te liggen. Deze vorm van anisotropie wordt in de literatuur beschreven als 'zonale' anisotropie (Journel & Huijbregts, 1978). McBratney & Webster (1986) merken op dat deze vorm van anisotropie nog nooit in bodemstudies is vermeld en voor zover ons bekend is er dan ook geen programmatuur aanwezig die de benodigde berekeningen correct kan uitvoeren. Vanwege het feit dat de isotope variogrammen valide zijn in termen van de kruisvalidatie-procedure, d.w.z. ze leiden tot zuivere 'kriging' schattingen waarvan de variantie consistent is met de geschatte variantie, werd besloten de richtingseffecten tijdens de interpolatie buiten beschouwing te laten.



Figuur 5 Experimentele variogrammen voor (a) Zn en (b) RE, en (c) kruisvariogram voor Zn en RE (isotoop)

Tabel 3 Parameters van de exponentiële variogrammodellen

attribuut	nugget C_0	sill C_0+C	range a
Zn	79262	314958	0.59
RE	849	9564	0.38
Zn-RE	-4254	-20351	0.73



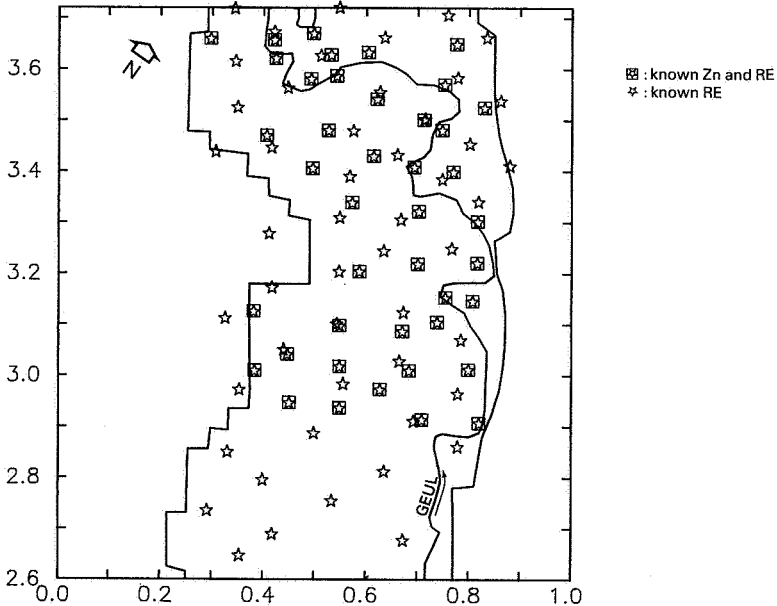
Figuur 6 Experimentele variogrammen voor Zn (* = loodrecht op de dalrichting, o = parallel aan de dalrichting)

4.3 De kartering

De variaties van de zinkgehalten in de bodem van het studiegebied werden op drie manieren in kaart gebracht:

- (1) lineaire regressie, gebruikmakend van de 45 testpunten en de regressievergelijking met de grootste correlatiecoëfficiënt (Tabel 2);
- (2) point-kriging, gebruikmakend van de 154 punten met gegevens over zinkgehalten in de bodem;

(3) point-co-kriging, gebruikmakend van dezelfde 154 punten en de aanvullende 309 hoogtepunten.



Figuur 4 Monsterpunten in een deel van het onderzoeksgebied

De zinkconcentraties in de bodem werden geschat op de 45 testpunten en vergeleken met de werkelijke waarden. De mediaan en de interkwartiel range van de frequentieverdelingen van de aldus berekende absolute schattingsfouten staan weergegeven in Tabel 4 en werden gebruikt om de drie methoden met elkaar te vergelijken. Procedure (2) en (3) werden eveneens gebruikt voor de vervaardiging van isolijnenkaarten van zinkgehalten in de bodem van het studiegebied (Figuur 7).

Tabel 4 laat zien dat er aanmerkelijke verschillen bestaan tussen de resultaten van de verschillende schattingsmethoden. De beide kriging-interpolatoren leiden tot betere resultaten dan het gebruik van een lineaire regressie en, ondanks de lage correlatie-coëfficiënt van -0.48, de co-kriging-interpolatie heeft geleid tot geringere schattingsfouten dan de gewone kriging-interpolatie.

Figuur 8a en 8b geven blokdiagrammen van de schattingsvariantie in het zinkgehalte door interpolatie met resp. point-kriging en point-co-kriging. De point co-kriging variantie is over het gehele gebied aanmerkelijk lager dan voor point-kriging en, zoals mocht worden

verwacht uit het grotere aantal gebruikte gegevens, veel stabiel van aard. Dit verschil wordt verder uitgewerkt in Figuur 9, dat de isolijnen van het verschil tussen Figuur 8a en 8b weergeeft. In het noorden van het onderzoeksgebied is de co-kriging-schattingsvariantie

Tabel 4 Mediaan en interkwartielafstand (mg/kg) van absolute en gekwadrateerde voorspelfouten (n=45)

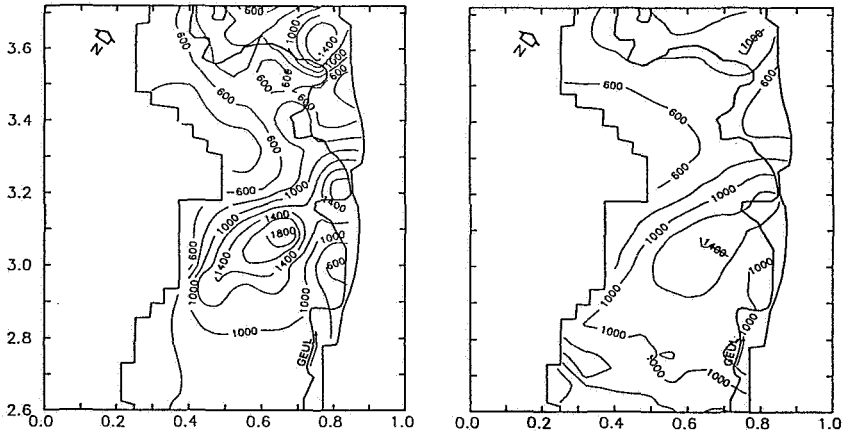
		mediaan	interkwartiel-afstand
lineaire regressie	(AF)	353	453
	(KF)	124,357	351,102
kriging	(AF)	311	400
	(KF)	97,014	243,853
co-kriging	(AF)	283	410
	(KF)	80,241	234,954

AF: absolute fout; KF: gekwadrateerde fout.

niet alleen veel stabiel van aard dan de kriging-schattingsvariantie, maar bedraagt bovendien 10 tot 20 % minder. Dit verschil loopt op tot ca. 30 % in het zuiden van het onderzoeksgebied, waar relatief weinig gegevens over het zinkgehalte in de bodem bekend waren. De lage co-kriging-schattingsvariantie voor zink is het gevolg van de beschikbaarheid van aanvullende hoogtegegevens.

5 DISCUSSIE

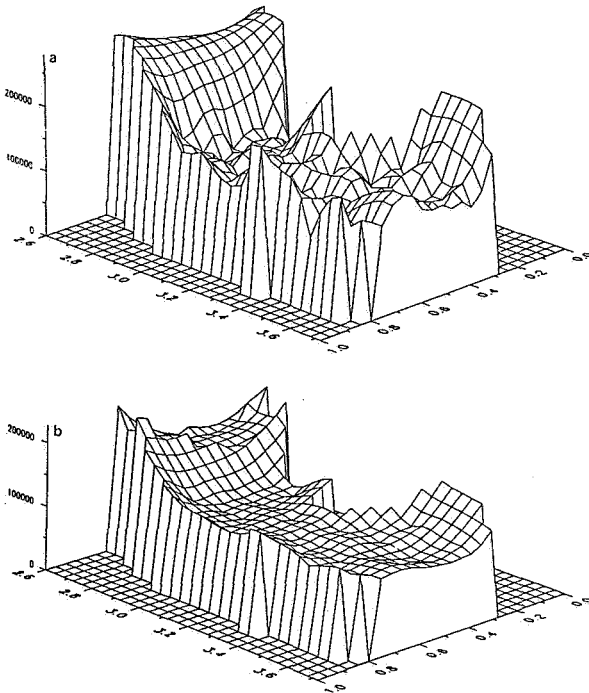
Een recente ontwikkeling bij het modelleren van de mogelijke gezondheidsrisico's als gevolg van bodemverontreiniging, is het gebruik van Monte Carlo simulatietechnieken (Leenaers et al., in prep.). Deze technieken maken het mogelijk de voortplanting van onzekerheden in de invoerparameters van een model te analyseren en de daaruit voortvloeiende spreiding in de modeluitkomsten te kwantificeren (Heuvelink et al., 1990). Om optimaal van deze techniek gebruik te kunnen maken is het dus noodzakelijk dat de spreiding van alle invoerparameters, zoals bijvoorbeeld de concentratie van een schadelijke component in de bodem, bekend is. De in het bovenstaande besproken geostatistische



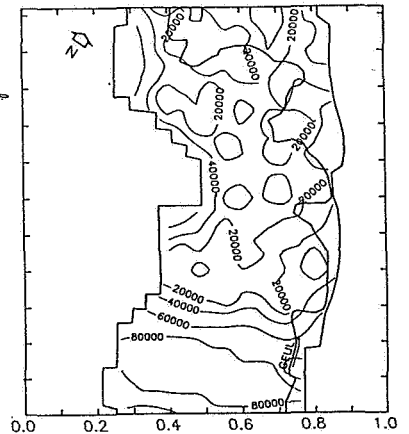
Figuur 7 Isolijnenkaarten van zinkgehalten in de bodem van het studiegebied, gemaakt met (a) point-kriging en (b) point-co-kriging

interpolatietechnieken kriging en co-kriging komen, in tegenstelling tot andere veel gebruikte ruimtelijke voorspellingstechnieken, aan deze eis tegemoet door naast een voorspelling ook de variantie van de voorspelling op te leveren.

Een reeds genoemde invoerparameter van risicobeoordelingsmodellen is de concentratie van een schadelijke component in de bodem, die, na vermenigvuldiging met de experimenteel vastgestelde bodemingestie, leidt tot een schatting van de inname van de schadelijke component door de bevolkingsgroep onder beschouwing. Voor de schatting van deze inname is het niet alleen gewenst dat de spreiding in de concentratie van de schadelijke component bekend is, maar ook dat deze spreiding zo klein mogelijk is. Een vermindering van de spreiding in de invoerparameters leidt immers tot een vermindering van de spreiding in de modeluitkomsten en dientengevolge tot een betere schatting van mogelijke risico's voor de volksgezondheid. Het zal duidelijk zijn dat de door co-kriging tot stand gebrachte vermindering van de schattingsvariantie, zelfs als die niet gepaard zou zijn gegaan met een vermindering van de absolute voorspelfout (zoals weergegeven in Tabel 4), in een dergelijk model kan leiden tot een betrouwbaarder risicoschatting.



Figuur 8 Blokdiagrammen van de variantie in het geschatte zinkgehalte door interpolatie met (a) point-kriging en (b) point-co-kriging



Figuur 9 Isolijnenkaart van het verschil tussen de kriging en co-kriging-varianties (Figuur 8a en 8b)

6 CONCLUSIES

De bovenstaande resultaten maken duidelijk dat het mogelijk is om met succes gebruik te maken van aanvullende hoogtegegevens bij het karteren van bodemverontreiniging in overstromingsgebieden. De gehanteerde co-kriging-procedure levert niet alleen lagere voorspelfouten, maar resulteert bovendien in een stabielere en lagere schattingsvariantie.

Hoogtegegevens zijn in grote hoeveelheden en goedkoop beschikbaar op de kaarten van de Topografische Dienst. De kwaliteitsverbetering van de verontreinigingskaarten kon dan ook zonder noemenswaardige kosten tot stand worden gebracht. Het gebruik van co-kriging was in dit geval zinvol omdat er vanuit geomorfologisch oogpunt aanleiding was om het verband tussen de hoogteligging en de mate van verontreiniging verder te onderzoeken. Bestaand inzicht in de aard van verspreidingsmechanismen zou ook in andere gevallen kunnen leiden tot een verbetering van de ruimtelijke voorspelling van de daaruit voortvloeiende patronen.

LITERATUUR

- COOPER, R.M. and ISTOK, J.D.; 1988. Geostatistics applied to groundwater contamination, I: Methodology. *Journal of Environmental Engineering*, 114, 2: 270-286.
- DEFIZE, P.R.; 1991. De fundamente van Kriging en Co-Kriging. CHO-TNO Rapporten en Nota's No. 24.
- HEUVELINK, G.B.M., BURROUGH, P.A. and LEENAERS, H.; 1990. Error propagation in spatial modelling with GIS. *Proceedings of the First European Conference on Geographical Information Systems*, 453-462.
- JOURNEL, A.G. and HUIJBREGTS, Ch.J.; 1978. *Mining Geostatistics*. Academic Press, New York.
- LEENAERS, H. and OKX, J.P.; 1989. The use of Digital Elevation Models for flood hazard mapping. *Earth Surface Processes and Landforms*, 14: 631-640.
- LEENAERS, H., RANG, M.C. and SCHOUTEN, C.J.; 1988. Variability of the metal content of flood deposits. *Environmental Geology and Water Science*, 11, 1: 95-106.
- LEENAERS, H., RANG, M.C. and RANG, D.M.C.; in prep. Omgaan met onzekerheid bij het modelleren van gezondheidsrisico's als gevolg van bodemverontreiniging. *Tijdschrift voor de Sociale Gezondheidszorg*.
- McBRATNEY, A.B. and WEBSTER, R.; 1986. Optimal interpolation and isarithmic mapping of soil properties: V. Co-regionalization and multi-sampling strategy. *Journal of Soil Science*, 34: 137-162.
- RANG, M.C., KLEIJN, C.E. and SCHOUTEN, C.J.; 1987. Mapping of soil pollution by application of classical geomorphological and pedological field techniques. In: Gardiner, V. (ed.); 1987. *International Geomorphology*, part I: 1029-1044. Wiley and Sons.
- VAN DE WESTERINGH, W.; 1980. Soil conditions, soil carbonates and former vegetation in the Geul valley from Gulpen to Meerssen. *Mededelingen Landbouw Hogeschool Wageningen*, 80, 8, 1-26.1.

KALMAN FILTERING BIJ RUIMTELIJKE INTERPOLATIE VAN GRONDWATER-STIJGHOOGTEN

Frans C. van Geer

SAMENVATTING

In beginsel kan het stijghoogte verloop in de ruimte en in de tijd worden beschreven met fysische wetten. Deze beschrijving zal afwijken van de werkelijk optredende stijghoogte door het gebrek aan systeemkennis en onvolkomenheden in de numerieke oplossingsmethoden. Los van deze beschrijving van het grondwaterproces wordt informatie ingewonnen over de actuele stijghoogte door middel van stijghoogte waarnemingen. Met Kalman filtering is het mogelijk de fysische beschrijving en de waarnemingen zodanig te combineren dat op elk meettijdstip optimaal gebruik wordt gemaakt van beide informatiebronnen (optimaal in de zin van minimale varianties). In deze presentatie ligt de nadruk op het ruimtelijke aspecten van Kalman filtering; ruimtelijke interpolatie. Behalve een optimale ruimtelijke interpolatie wordt ook de covariantiematrix van de interpolatie fout bepaald, die een maat is voor de onzekerheid van die interpolatie. Van het Kalman filter algoritme wordt geen afleiding gegeven, alleen de werking wordt besproken en geplaatst in een geohydrologische context. De methode wordt geïllustreerd met een praktijk voorbeeld.

1 INTRODUCTIE

Zolang er mensen met geohydrologisch onderzoek bezig zijn, stelt men zich de vraag hoe de grondwaterstijghoogte verloopt in de ruimte en in de tijd. Om een antwoord te geven op die vraag zijn theorieën ontwikkeld (bv. de wet van Darcy) en wordt de stijghoogte op vele punten waargenomen. In verband met het thema van de lezingendag "ruimtelijke variabiliteit", zal ik mij in hoofdzaak beperken tot de ruimtelijke aspecten. De vraag wordt daarmee gereduceerd tot het (re)construeren van het verloop van de stijghoogte in de ruimte, gegeven een aantal waarnemingen en met kennis van de theoretische grondslagen van grondwaterstroming.

In de geohydrologie wordt deze vraag traditioneel beantwoord met behulp van een conceptueel model, dat vervolgens, overeenkomstig de fysische wetten, vertaald wordt in een differentiaalvergelijking. Deze vergelijking wordt tegenwoordig meestal numeriek opgelost en het resultaat is een berekend stijghoogte verloop in de ruimte (en de tijd). De beschikbare stijghoogte waarnemingen binnen het modelgebied, worden daarbij uitsluitend gebruikt voor het ijken van de modelparameters. Na de ijking is de berekende stijghoogte alleen afhankelijk van de randvoorwaarden en de begincondities. Een aldus verkregen stijghoogteverloop is feitelijk geen interpolatie, maar eerder een simulatie.

Met de komst van de geostatistiek is het mogelijk geworden om ruimtelijk verdeelde variabelen te interpoleren op basis van de statistische structuur van waarnemingen. Het meest toegepaste in dit verband zijn Kriging en aanverwante interpolatietechnieken. Hoewel deze technieken zijn ontwikkeld voor tijdonafhankelijke variabelen, zoals bodemeigenschappen, kunnen ook (met enige voorzichtigheid) tijdsafhankelijke variabelen zoals stijghoogtevlakken worden geïnterpoleerd. Een belangrijk voordeel van geostatistische interpolatie is dat het resulterende stijghoogtevlak geconditioneerd is op de waarnemingen. Daarom is er in dit geval wel sprake van een echte interpolatie.

Binnen de in deze bijdrage beschreven Kalman filtering is het mogelijk om een aantal voordelen van beide bovengenoemde benaderingen te combineren. Enerzijds wordt er gebruik gemaakt van systeemkennis in de vorm van een differentie vergelijking, terwijl anderzijds de interpolatie wordt geconditioneerd op de waargenomen stijghoogte. Bovendien wordt evenals bij de geostatistische benadering, uit de statistische structuur een variantie

bepaald als maat voor de betrouwbaarheid. Eerlijkheidshalve moet worden opgemerkt dat deze Kalman filtering nogal bewerkelijk is en vooralsnog zeer rekenintensief.

In deze bijdrage ligt het accent vooral op de gedachtevorming. Daarom wordt er eerst aandacht besteed aan de uitgangspunten voordat het Kalman filter rekenschema wordt gegeven. Het rekenschema zelf wordt slechts toegelicht en niet afgeleid, terwijl aan het onderwerp ijken meer aandacht wordt besteed. Vervolgens wordt de techniek geïllustreerd aan de hand van uitgewerkte voorbeelden. Ten slotte zullen voor en nadelen worden besproken en worden enige conclusies getrokken.

2 UITGANGSPUNTEN

Om de stijghoogte te interpoleren kan via twee wegen informatie over de werkelijke grondwaterstroming worden gebruikt:

- systeemkennis;
- stijghoogte waarnemingen.

2.1 Systeemkennis

Op basis van geologische en hydrologische informatie kan een zo groot mogelijk deel van de grondwaterbeweging worden verklaard. Vrijwel altijd gebeurt dit via een numeriek geohydrologisch model. De parameters worden met behulp van waarnemingen geïjkt, waarna "het model" klaar is en de gewenste interpolatie kan worden uitgevoerd. In de deterministische gedachtegang kan de werkelijkheid oneindig dicht worden benaderd door meer detail in het model op te nemen, dus door kleinere rooster afstanden en meer gedifferentieerde gegevens over parameterwaarden. Het verschil tussen de werkelijkheid en het model wordt als het ware gezien als een soort afbreekfout en niet als een conceptuele grootheid. Omdat aangenomen wordt dat die afbreekfout binnen de gewenste marges kan worden gebracht, wordt (na ijking) de informatie uit de verschillen tussen werkelijkheid en model niet verder benut.

Bij stochastische modelconcept wordt het verschil tussen de werkelijkheid en het model expliciet in rekening gebracht. Daarmee wordt inherent geaccepteerd dat met het model niet

de volledige werkelijkheid kan worden beschreven. Het verschil tussen modelbeschrijving en werkelijkheid wordt systeemruis genoemd.

$$\text{"werkelijkheid"} = \text{"model"} + \text{"systeemruis"}$$

De gedachtengang is nu dat de globale grondwaterstroming met een relatief weinig gedetailleerd model kan worden beschreven. Om toch een gedetailleerd beeld van de werkelijkheid te krijgen wordt gebruik gemaakt van de statistische structuur in de systeemruis en van de actuele waarnemingen.

We gaan uit van een niet-stationair, numeriek, deterministisch stijghoogte model (bv. MODFLOW, McDonnald en Harbaugh, 1987). Zo'n model relateert de stijghoogte op een aantal roosterpunten op het tijdstip k aan de stijghoogte op het voorgaande tijdstip ($k-1$) en aan een drijvende kracht gedurende de tijdstap. Die drijvende kracht, ook input genoemd, wordt gevormd door randvoorwaarden, voeding van het grondwater, onttrekkingen enz. Doorgaans wordt voor de oplossing een impliciet eindig(e) differentie schema, of een eindig(e) elementen schema gebruikt. Voor de duidelijkheid zullen de formules hier alleen in explicite vorm worden gegeven. Uit de systeemkennis volgt de eerste basisvergelijking van het Kalman filter; de systeemvergelijking:

$$h_k = A h_{k-1} + B u_k + w_k \quad (1)$$

hierin is h_k een vector bestaande uit de stijghoogte op alle roosterpunten op het tijdstip k

u_k een vector bestaande uit de input van het systeem

w_k de systeemruisvector die voor elk roosterpunt het verschil tussen het model en de werkelijkheid aangeeft

A en B coëfficiënten matrices die volgen uit de model beschrijving.

De eerste twee termen in het rechter lid van (1) representeren niets anders dan het numerieke stijghoogte model.

2.2 Stijghoogte waarnemingen

De stijghoogte kan op een aantal punten in de ruimte en in de tijd worden waargenomen. De informatie uit deze waarnemingen zijn niet afhankelijk van systeemkennis of van een

modelbeschrijving. De relatie tussen de waarnemingen en de werkelijke stijghoogte is alleen afhankelijk van de locaties van de meetpunten. Bovendien zijn waarnemingen behept met meetfouten. Voor de numerieke verwerking van de meetinformatie wordt de stijghoogte op gelijke wijze gediscretiseerd als in het numerieke model. De relatie tussen de waarnemingen en de gediscretiseerde stijghoogte is de tweede basisvergelijking van het Kalman filter; de meetvergelijking:

$$y_k = C h_k + v_k \quad (2)$$

hierin is: y_k de meetvector, waarvan de elementen de waarnemingen zijn op tijdstip k
 v_k een vector, waarvan de elementen de waarnemingsfouten zijn op tijdstip k
 C de meetmatrix, die uitsluitend afhangt van de locaties van de meetpunten.

Indien de meetlocaties per meettijdstip verschillen (bv. verschillende meetfrequenties of missende waarnemingen) verschilt ook de meetmatrix per meettijdstip. Echter, aangezien het accent hier ligt op de ruimtelijke aspecten, wordt aangenomen dat alle meetlocaties op elk meettijdstip worden waargenomen, zodat de meetmatrix onafhankelijk is van de tijd.

3 REKENSHEMA

Kalman filtering, zoals dat bekend is binnen de systeem theorie, is zeer algemeen opgezet en op zeer veel processen toepasbaar. Voor de interpolatie van de stijghoogte kan worden volstaan met een eenvoudige lineaire versie, die in de hierna volgende sectie zonder bewijs wordt gegeven. Het filter algoritme is o.a. afgeleid in Van Geer (1987). Voor een uitgebreidere behandeling van het Kalman filter wordt verwezen naar de literatuur (bv. Sage & Melsa, 1976).

Het doel van de berekening is een conditionele interpolatie. Dat wil zeggen we willen de "beste" schatting van de stijghoogte gegeven een aantal waarnemingen. In dit geval wordt met "beste" bedoeld, de schatting met een minimale variantie van de interpolatie fout. Bovendien willen we ook, als maat voor hoe goed de interpolatie is, de variantie van de interpolatie fout weten. Voor de leesbaarheid worden eerst een paar definities gegeven.

* \hat{h}_k de beste schatting van de stijghoogte op tijdstip k gegeven waarnemingen tot en met tijdstip k.

Kortheidshalve zal deze grootheid de optimale schatting worden genoemd.

* h_k^* de beste schatting van de stijghoogte op tijdstip k gegeven waarnemingen tot en met tijdstip k-1.

Deze grootheid wordt predictie genoemd.

* P_k de covariantie matrix van de fout in de optimale schatting

* M_k de covariantie matrix van de fout in de predictie

* Q de covariantie matrix van de systeemruis

* R de covariantie matrix van de meetfouten

* K_k de Kalman gain

* I de eenheidsmatrix

Er kan worden bewezen dat voor een proces waarvoor de vergelijkingen (1) en (2) gelden en waarbij bovendien de systeemruis en de meetfout multi-dimensionale witte ruis processen zijn, de optimale schatting en de bijbehorende covariantie matrix kunnen worden berekend met het algoritme in tabel 1.

Tabel 1. Kalman filter algoritme.

$$h_k^* = A \hat{h}_{k-1} + B u_k \quad (3)$$

$$M_k = A P_{k-1} A^T + Q \quad (4)$$

$$K_k = M_k C^T \{ C M_k C^T + R \}^{-1} \quad (5)$$

$$\hat{h}_k = h_k^* + K_k \{ y_k - C h_k^* \} \quad (6)$$

$$P_k = \{ I - K_k C \} M_k \quad (7)$$

3.1 Predictie correctie methode

Het Kalman filter algoritme, zoals gegeven in tabel 1, kan worden beschouwd als een predictor-corrector methode. Stel dat op het tijdstip k-1 de optimale schatting \hat{h}_{k-1} en de

corresponderende covariantie matrix P_{k-1} bekend zijn. Dan kan met vergelijking (3) een predictie van de stijghoogte op het volgende tijdstip worden berekend. De covariantie matrix van deze predictie (de maat voor de onzekerheid in de predictie) wordt gegeven met (4). Uit deze laatste vergelijking blijkt dat de onzekerheid in de predictie wordt veroorzaakt door de onzekerheid in de optimale schatting op tijdstip $k-1$ en de covariantie van de systeemruis (Q).

Wanneer op het tijdstip k de waarnemingen y_k beschikbaar komen wordt de predictie gecorrigeerd volgens vergelijking (6). De correctie is evenredig met het verschil tussen de waarnemingen en de predictie daarvan. Dit verschil wordt de innovatie n_k genoemd:

$$n_k = y_k - C h_k^* \quad (8)$$

De wegingsmatrix, of Kalman gain K_k , in vergelijking (6) is afhankelijk van de verhouding tussen de onzekerheid in de predictie en de meetfouten (resp. M_k en R). De Kalman gain wordt berekend met vergelijking (5). Ten slotte wordt met (7) de covariantie matrix van de optimale schatting berekend. De vergelijkingen (6) en (7) leveren de begincondities voor de volgende tijdstap.

De optimale schatting \hat{h}_k kan gezien worden als interpolatie van de stijghoogte waarbij de procesbeschrijving en de waarnemingen zodanig gecombineerd zijn, dat de covariantie matrix van de interpolatie fout minimaal is.

3.2 IJken

3.2.1 Structuur in de matrices

Om het algoritme uit tabel 1. te kunnen toepassen dienen behalve de begincondities \hat{h}_{k-1} en P_{k-1} ook de matrices A , B , C , R en Q bekend te zijn.

De meetmatrix C is alleen afhankelijk van de locaties van de meetpunten. Ook de covariantie matrix van de meetfouten R kan eenvoudig worden afgeleid omdat kan worden aangenomen dat de meetfouten onderling onafhankelijk zijn. De matrix R reduceert daarmee tot een diagonaal matrix met op de diagonaal de variantie van de meetfout. Deze variantie

wordt bepaald door de nauwkeurigheid van het meetinstrument en door de afleesfouten van de waarnemer. In Nederland, waar de stijghoogte zich doorgaans dicht onder het maaiveld bevindt wordt meestal een meetfoutvariantie van 1 cm^2 aangehouden. Overigens zijn de meetfouten van ondergeschikte betekenis vergeleken bij de overige bronnen van onzekerheid.

De matrices A, B, en Q zijn niet zo eenvoudig in te vullen en zullen moeten worden geijkt aan de waarnemingen. De dimensie van deze matrices zijn: A $n \times n$, B $n \times m$, en Q $n \times n$, waarbij n het aantal roosterpunten is en m het aantal inputvariabelen. Het zal duidelijk zijn dat de elementen van deze matrices niet met het beschikbare waarnemingsmateriaal onafhankelijk van elkaar kunnen worden geijkt. Het aantal onafhankelijke parameters dient derhalve te worden gereduceerd, door middel van het aannemen van een structuur in de matrices.

De structuur in de matrices A en B wordt hier bekend verondersteld. Deze matrices zijn afhankelijk van de geohydrologische parameters en het gekozen numerieke schema. In de hierna te behandelen toepassing zijn de matrices A en B identiek aan de coëfficiënten matrices in het programma MODFLOW (zij het dat in MODFLOW een implicite vorm wordt gebruikt).

Een covariantiematrix van de systeemruis is een minder gangbaar begrip in de geohydrologie. De systeemruis variantie op één bepaalde locatie is een maat voor de mate waarin het model de werkelijkheid beschrijft. Hoe groter deze variantie hoe slechter het model. De varianties vormen de diagonaal van de systeemruis covariantie matrix. De eerste stap die we nu doen is het modelgebied indelen in subgebieden, waar binnen de variantie van de systeemruis constant wordt verondersteld (analoog aan het definiëren van bv. subgebieden met constante kD waarde). Het indelen in subgebieden is per toepassing verschillend en geschiedt op basis van geohydrologisch inzicht en ervaring. Indien bijvoorbeeld de onzekerheid voornamelijk wordt veroorzaakt door onzekerheid over de aanvulling door neerslag, speelt de toestand van de onverzadigde zone, die niet in het model voorkomt, een belangrijke rol. Het ligt dan voor de hand de subgebieden met gelijke variantie te kiezen overeenkomstig de dikte van de onverzadigde zone. In de meeste gevallen kan met een klein aantal verschillende varianties het hele gebied worden gekarakteriseerd. De tweede stap bij het aanbrengen van een structuur in de systeemruis

covariantie matrix is het bepalen van de nevendiagonalen. Een element van een nevendiagonaal geeft de samenhang aan tussen de modelfouten op twee verschillende ruimtelijke locaties. Deze samenhang zal in het algemeen afnemen naar mate de afstand tussen de locaties groter wordt. Uit de geostatistiek zijn een aantal mogelijkheden bekend om zo'n samenhang met enkele parameters te beschrijven. Tot dusver hebben wij gekozen voor een exponentieel verband, dat met één parameter (een correlatie coëfficiënt tussen twee elementen van de systeemruis op een eenheidsafstand) is te beschrijven.

Het element van de systeemruis covariantie matrix is bepaald door de uitdrukking:

$$q_{i,j} = \rho^d * \sigma_i * \sigma_j \quad (9)$$

hierin is $q_{i,j}$ een element van de systeemruis covariantie matrix corresponderend met de locaties i en j

ρ de correlatie coëfficiënt op de eenheidsafstand

d de afstand tussen locaties i en j

σ_i de standaard afwijking van de systeemruis in punt i .

3.2.2 IJkings criteria

Nu het aantal onafhankelijke elementen in de matrices A , B en Q is gereduceerd door het aanbrengen van een structuur, kunnen ze worden geijkt met de waarnemingen. Daartoe worden twee criteria gebruikt ten aanzien van de innovatie vector. De elementen van deze vector zijn de verschillen tussen de waargenomen stijghoogte en de predictie van die waarneming. Indien het rekenschema T tijdstappen wordt doorgerekend beschikken we per meetlocatie dus over een innovatie tijdreeks die T tijdstappen lang is.

In de eerste plaats eisen we dat de predictie geen systematische afwijkingen vertoont. Dat houdt in dat voor elke meetlocatie de verwachtingswaarde van de innovatie gelijk is aan nul. Aangezien bewezen kan worden dat de innovaties onafhankelijk in de tijd zijn, kan de verwachtingswaarde eenvoudig worden geschat door het bepalen van het rekenkundige gemiddelde. Uit vergelijking (8) blijkt dat voor een gegeven set waarnemingen de innovatie alleen afhankelijk is van de predictie. Door middel van trial and error kunnen de parameterwaarden zodanig worden aangepast dat het gemiddelde van de innovatie niet

significant verschilt van nul. In de praktijk is gebleken dat de gemiddelde waarde van de innovaties in belangrijke mate afhangen van de geohydrologische parameters in de matrices A en B. Deze procedure van ijken wordt ook in "gewone" geohydrologische modellering veelvuldig toegepast, zij het dat dan niet de innovaties worden gebruikt maar het verschil tussen de modelsimulatie en de waarnemingen.

In de tweede plaats eisen we dat de covariantie matrix van de innovaties zoals die wordt "voorspeld" door het filter algoritme, overeenkomt met de berekende covariantie matrix uit waargenomen innovaties. De covariantie matrix van de innovaties geeft een maat voor de spreiding in, en de ruimtelijke samenhang tussen de innovaties. Uit het algoritme kan de covariantie matrix (Z) van de innovaties worden afgeleid:

$$Z = C M_k C^T + R \quad (10)$$

Omdat de invloed van de beginconditie P_0 na verloop van tijd is uitgedempt en de matrices C en R bekend zijn, is deze covariantie matrix alleen afhankelijk van de matrices A, B en Q. Met name de systeemruis covariantie matrix Q is van grote invloed op de matrix Z. Alle covariantie matrices, ook van de innovaties, kunnen zonder gebruik te maken van actuele waarnemingen worden bepaald. Door nu de te ijken parameters (standaardafwijkingen en correlatie coëfficiënt van de systeemruis en geohydrologische parameters) aan te passen, wordt de matrix Z in overeenstemming gebracht met de experimentele covariantie matrix die uit de innovatie series kan worden bepaald.

4 TOEPASSING

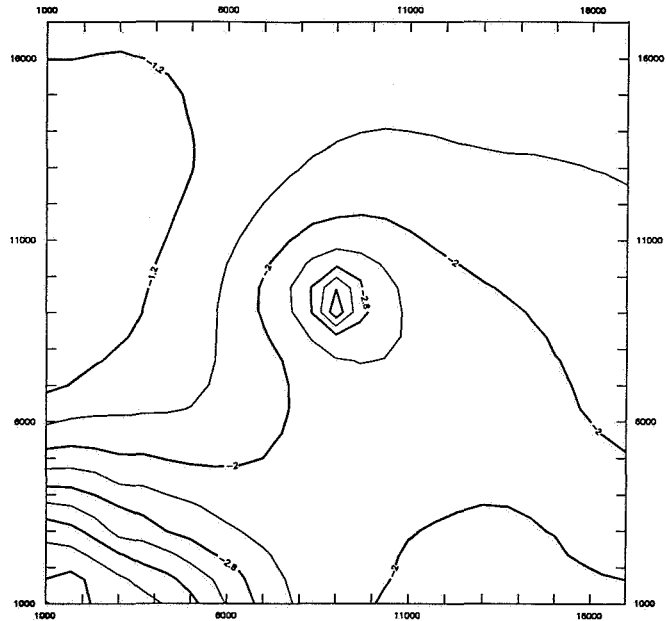
Ter illustratie van de interpolatie met behulp van Kalman filtering wordt in dit hoofdstuk een toepassing gepresenteerd van een stijghoogte interpolatie rond de waterwinplaats Spannenburg (Friesland). Een uitgebreide beschrijving van deze toepassing is gegeven in Van Geer en Te Stroet (1989). Hier wordt alleen de relevante informatie voor de ruimtelijke interpolatie zeer kort beschreven.

Het gebied rond de winplaats is geschematiseerd tot één goed doorlatend pakket (kD-waarden van 4000 tot 8000 m²/d) met een kleine bergingscoëfficiënt (S-waarde 10⁻⁴)

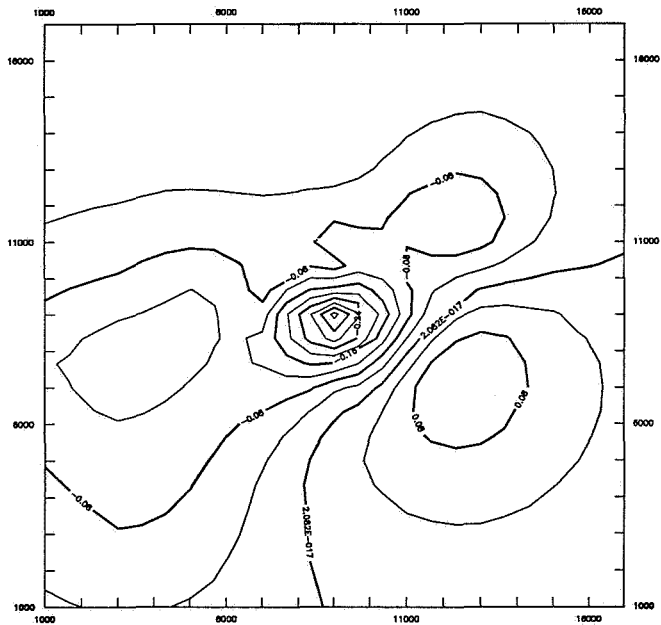
dat wordt afgedekt door een nogal heterogeen topsysteem met c-waarden van 15 tot 300 dagen. In het centrum van het modelgebied wordt onttrokken en de 18 waarnemingspunten liggen in min of meer loodrecht op elkaar staande raaien. Het gebied is gediscrètiseerd met rooster afstanden van 2000 m aan de randen tot 500 m bij de onttrekking. In totaal zijn 225 roosterblokken gebruikt. De tijdstap is overeenkomstig de meetfrequentie gesteld op 15.2 dagen. De berekeningen zijn uitgevoerd over de periode 1976 tot 1983. De hierna gegeven figuren hebben alle betrekking op één tijdstip.

In figuur 1 is de interpolatie van de stijghoogte gegeven zoals die met het Kalman filter algoritme is bepaald (de optimale schatting). Duidelijk is het dal in het midden van het modelgebied. Dit dal komt overeen met het centrum van de winning. De lage stijghoogte in de hoek links onder wordt veroorzaakt door de Noordoostpolder.

Voor de illustratie van het Kalman filter is figuur 2 echter meer informatief. In deze figuur is het verschil tussen de predictie en de optimale schatting gegeven op één tijdstip. Met andere woorden figuur 2 geeft de correctie van de predictie als gevolg van de actueel waargenomen stijghoogte op 18 meetlocaties. Opvallend is de grote correctie ter plaatse van de winning. Dit wordt in hoofdzaak veroorzaakt doordat de winningscijfers zijn gebruikt in de vorm van maandtotalen, waardoor de onzekerheid in het hydrologische model groot is. In de nabijheid van de winning is de onzekerheid in de winningscijfers overheersend. Daardoor kunnen de verschillen tussen de waarnemingen en de predictie groot zijn en daarmee ook de correcties (zie verg. (6)). In de rest van het gebied zijn de correcties veel kleiner. Opgemerkt dient te worden dat de stijghoogte ook wordt gecorrigeerd op niet bemeten locaties. Door de hoge ruimtelijke samenhang in dit gebied zijn de afzonderlijke meetpunten niet in correctie waarneembaar.

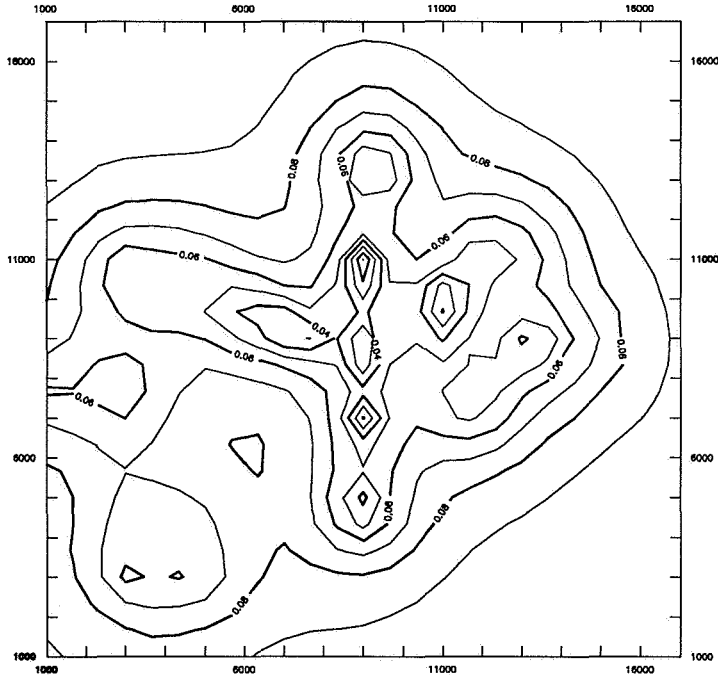


Figuur 1. Interpolatie van de stijghoogte (in m NAP).



Figuur 2. Correctie van de predictie met actuele waarnemingen (in m).

Figuur 3 ten slotte, geeft de standaardafwijking van de interpolatie fout. Deze standaardafwijking is een maat voor de betrouwbaarheid van de interpolatie. Uiteraard is in deze figuur de invloed van de afzonderlijke meetpunten wel herkenbaar. Dit in tegenstelling tot de figuren 1 en 2.



Figuur 3. Standaardafwijking van de interpolatiefout (in m).

5 DISCUSSIE

In de hoofdstukken hiervoor is beschreven hoe op basis van een numeriek stijghoogte model en Kalman filtering een ruimtelijke interpolatie kan worden gemaakt. Belangrijke voordelen van deze methode zijn dat gebruik wordt gemaakt van zowel de systeemkennis (in de vorm van een geohydrologisch model) als van de actuele stijghoogte waarnemingen. In deze discussie wil ik voornamelijk ingaan op verschillen tussen en overeenkomsten met een geostatistische benadering. In beide benaderingen is de gedachtengang dat de werkelijkheid is te verdelen in een bekend, deterministisch deel en een stochastisch deel. Deze twee delen worden hierna afzonderlijk besproken.

5.1 Deterministisch deel

Het deterministische deel in het Kalman filter is het geohydrologische model terwijl dit deel in de geostatistiek vaak wordt aangeduid met trendvlak of drift. De Kalman filter benadering is gebaseerd op een onderliggende partiële differentiaal vergelijking, waardoor een bekend geohydrologisch model zoals MODFLOW op natuurlijke wijze in het rekenschema past. De geostatistiek daarentegen, is in beginsel ontwikkeld voor tijdsafhankelijke variabelen en de schatting van het trendvlak is meestal gebaseerd op het fitten van een functie op de waarnemingen (bv. een polynoom). Bij Kalman filtering is het daardoor mogelijk meer detail in het trendvlak op te nemen en rekening te houden met traagheidseffecten. Daartegenover staat dat een geostatistische benadering aanzienlijk veel minder reken- en persoonstijd vraagt.

5.2 Stochastisch deel

Het stochastische deel wordt bij de geostatistische benadering geheel vastgelegd met een variogram, of een gegeneraliseerde covariantie functie. Daarbij gaat men er in principe van uit dat de samenhang tussen twee punten in de ruimte alleen afhangt van de onderlinge afstand terwijl de variantie over het hele gebied constant is. Bij de Kalman filter benadering is een volledige covariantie matrix gedefiniëerd en niet alleen de coëfficiënten van een variogram. Om die matrix te kunnen ijken moet het aantal te ijken parameters worden gereduceerd. Daartoe hebben we een structuur ingevoerd die is gegeven in vergelijking (9). Daarbij is de ruimtelijke samenhang met een exponentiële functie beschreven. Dit komt overeen met een genormeerd exponentieel variogram. De variantie is echter plaatsafhankelijk, waardoor de grootte van het stochastische deel in de ruimte kan variëren. Deze extra vrijheid in de covariantie structuur is mogelijk omdat we voor de ijking bij het Kalman filter beschikken over tijdreeks informatie op elke meetlocatie.

5.3 Slotopmerkingen

Resumerend kan worden gesteld dat de interpolatie met Kalman filtering grote mogelijkheden biedt, met name waar het gaat om het in rekening brengen van gedetailleerde informatie over de geohydrologische situatie. Daartegenover staat dat het Kalman filter zeer bewerkelijk is.

Binnen het kader van de lezingendag "Ruimtelijke statistiek in bodem en water" is alleen aandacht besteed aan de ruimtelijke aspecten van Kalman filtering. Het dient echter te worden vermeld dat deze techniek zowel de ruimtelijke als de tijdsaspecten in beschouwing neemt.

LITERATUUR

- GEER F.C. VAN; 1987. Applications of Kalman filtering in the analysis and design of groundwater monitoring networks. TH-Delft.
- GEER F.C. VAN en TE STROET, C.B.M.; 1989. Computercode C.I.S.K.A. toegepast op testcase Spannenburg. DGV-TNO. OS 90-05-A, Delft.
- McDONALD, M.G. and HARBAUGH, A.W.; 1988. A modular three-dimensional finite difference groundwater flow model. U.S. Geological Survey, report 83-875.
- SAGE, A.P. and MELSA, J.L.; 1976. Estimation Theory with Applications to Forecasting and Control. McGraw-Hill.

COVARIANTIESTRUCTUUR VAN KAARTEN

J.V. Witter, A.C. van Eijnsbergen en M.A.J. van Montfort.

SAMENVATTING

Kaarten verkregen door het toepassen van een interpolatie- dan wel karteringsprocedure, worden vaak gebruikt als invoer in een simulatiemodel. Dergelijke kaarten dienen om de ruimtelijke verdeling van bijvoorbeeld helling of infiltratiecapaciteit in een erosiemodel, of om de neerslag in een neerslag-afvoer model te beschrijven. In het laatste geval is het gebruikelijk om reeksen van dergelijke kaarten te gebruiken, één voor iedere tijdstap van de simulatie. Men beseft vaak onvoldoende dat dergelijke kaarten in hun ruimtelijke en temporele covariantiestructuur verschillen van de toevalsvelden die zij zouden moeten representeren. Dit beïnvloedt de simulatieresultaten. Dit artikel geeft een overzicht van het probleem, geeft procedures ter schatting van de mate waarin de covariantiestructuur van een bepaalde realisatie afwijkt van de werkelijke covariantiestructuur, en geeft enkele suggesties om het probleem te verhelpen. Het effect van een foutieve ruimtelijke en temporele covariantiestructuur op de uitkomst van simulatiemodellen zal worden geïllustreerd aan de hand van een neerslag-afvoer model.

1 INLEIDING

Geschatte realisaties van toevalsvelden worden vaak gebruikt als invoer in een simulatiemodel van bijvoorbeeld erosie of afvoer. Dergelijke geschatte realisaties zijn óf in de vorm van een tabel met (x,y) coördinaten en bijbehorende geïnterpoleerde waarden, verkregen door het toepassen van een interpolatieprocedure, óf in de vorm van een kaart verkregen door het toepassen van een karteringsprocedure. Praktisch gezien bestaat er geen verschil tussen beide procedures: kaarten worden immers getekend op grond van óf een soort van afstand-gewogen interpolatie (in het geval van hoogtelijnen kaarten), óf een met Thiessen-polygonen vergelijkbare methode (in het geval van thematische kaarten). In beide gevallen wordt de ruimtelijke covariantiestructuur beïnvloed. Terwille van een bondig taalgebruik worden hier resultaten van beide procedures met "kaart" aangeduid. Als schattingen van opeenvolgende realisaties worden gebruikt om de ruimtelijke verdeling van een variabele van het simulatiemodel te simuleren, wordt bovendien de temporele covariantiestructuur beïnvloed.

Door de introductie van Geografische Informatie Systemen (GIS) is de beschikbaarheid van dergelijke kaarten voor gebruik in simulatiemodellen toegenomen. Alhoewel de toegankelijkheid van deze kaarten vaak wordt gezien als een "interface" probleem tussen GIS en simulatiemodellen, is ons inziens de van de werkelijkheid afwijkende covariantiestructuur van deze kaarten een ernstiger "interface" probleem.

Zoals door Journel en Huijbregts (1978, p. 493) wordt benadrukt, is bij interpolatie het doel om op iedere locatie binnen het gebied van studie een waarde te schatten die zo dicht mogelijk de werkelijke waarde van de beoogde variabele benadert. Om een zo goed mogelijke schatting te verkrijgen, stelt men in het geval van kriging de eisen van zuiverheid en minimalisatie van de verwachte gekwadrateerde schattingsfout. Er is echter geen reden om aan te nemen dat deze schattingen ook de werkelijke ruimtelijke variatie vertonen. Bij gebruik van deze geschatte realisaties in simulatiemodellen zal juist de ruimtelijke variabiliteit belangrijk zijn, en Journel en Huijbregts (1978) presenteren een techniek die op iedere locatie binnen het studiegebied een "schatting" (een schatting in de gebruikelijke zin van het woord, namelijk gebaseerd op een steekproeffunctie, maar aangevuld met de resultaten van een simulatie) oplevert, die niet het "best" is in termen van minimale verwachte kwadratische fout, maar wel een ruimtelijke structuur heeft die overeenkomt met de werkelijkheid. Journel en Huijbregts noemen deze techniek conditionele simulatie. Deze "schatting" op grond van de conditionele simulatie is tevens zuiver.

De ruimtelijke covariantiestructuur wordt door het toepassen van interpolatie-, of karteringsmethoden in het algemeen glad gestreken. De interpolatie vertoont dus meer ruimtelijke samenhang dan de werkelijkheid. Over de effecten in het tijdsdomein bestaat nog weinig literatuur. Interpolatie voor opeenvolgende tijdstippen leidt tot twee tegenstrijdige effecten. Enerzijds zijn de realisaties voor de verschillende tijdstippen gladder dan de werkelijkheid, wat waarschijnlijk leidt tot verhoging van de temporele correlatie. Anderzijds wordt bij de interpolatie voor de opeenvolgende tijdstippen de temporele autocorrelatie niet expliciet gemodelleerd. Dit leidt zeker tot een verlaging van de temporele correlatie tussen opeenvolgende realisaties. Welk van de effecten overheerst is moeilijk aan te geven.

In dit artikel ligt de nadruk op de effecten van interpolatie op de ruimtelijke covariantiestructuur ("dispersie variantie"). Deze zijn beschreven in Journel en Huijbregts (1978). Recentelijk zijn kriging procedures ontwikkeld waarbij de temporele correlatie expliciet gemodelleerd wordt (Rouhani en Wackernagel, 1990). Alhoewel dit een verbetering is ten opzichte van traditionele procedures waarbij interpolaties voor opeenvolgende tijdstippen onafhankelijk van elkaar geschieden, blijft de covariantiestructuur van de interpolaties verschillen van die van de werkelijke velden.

Dit artikel begint met een overzicht van de effecten van interpolatie op de dispersie variantie, gebaseerd op Journel en Huijbregts (1978). Daarna behandelen we de techniek van conditionele simulatie, bedoeld om de dispersievariantie van de geïnterpoleerde velden te doen overeenstemmen met die van de werkelijke velden. Deze behandeling is eveneens gebaseerd op Journel en Huijbregts (1978). Vervolgens komen de effecten van interpolatie op de temporele covariantiestructuur aan de orde, gevolgd door een toepassing in de neerslag-afvoer modellering met zowel ruimtelijke als temporele aspecten.

2 DISPERSIE VARIANTIE

We gaan hier uit van een denkbeeldig onderzoek waarvoor de gemiddelde neerslaghoeveelheden over een zeker tijdsinterval (bijvoorbeeld een dag) nodig zijn voor een aantal deelgebieden v_i ($i = 1, \dots, n$) binnen een groot gebied V . Noteren we de respectievelijke gebiedsgemiddelden als $z(v_i)$ ($i = 1, \dots, n$) en $z(V)$, dan zullen deze in het algemeen worden geschat door gebruik te maken van puntwaarnemingen van de dagneerslag binnen het gebied, te noteren als $z(u_j)$ ($j = 1, \dots, m$). Er zijn dus drie schaalniveaus: de punten u_j , de deelgebieden v_i , alle binnen het gebied V .

Journal en Huijbregts (1978, p. 460 en volgende) geven het volgende voorbeeld. Het gebied V is rechthoekig en bestaat uit 50×10 deelgebieden; voor elk van deze 500 vierkante deelgebieden v_i ($i = 1, \dots, 500$) is het ertsgehalte bekend op 121 plaatsen u_j ($j = 1, \dots, 121$), gerangschikt volgens een regelmatig raster. Op grond van de 121 waarnemingen per deelgebied kan voor elk van deze deelgebieden een "werkelijk" gebiedsgemiddelde $z(v_i)$ worden bepaald als het rekenkundig gemiddelde van de waarnemingen binnen v_i .

De ruimtelijke variatie van een toevalsveld $Z(u)$ kan men karakteriseren door de zogenaamde dispersie variantie. Merk op dat het toevalsveld wordt aangeduid met Z en de realisatie ervan met z . Indien het gebied V volledig gepartitioneerd is in N even grote gebieden v_i ($i = 1, \dots, N$), dan is de dispersie variantie tussen de v 's binnen V :

$$D^2(v/V) = E \left\{ \sum_{i=1}^N [Z(v_i) - Z(V)]^2 / N \right\}, \quad (1)$$

waarin

$Z(v_i)$: gemiddelde van de toevalsfunctie over v_i

$Z(V)$: gemiddelde van de toevalsfunctie over V ; tevens het gemiddelde van de $Z(v_i)$.

De dispersie variantie neemt toe met de grootte van het gebied V , en neemt af met de grootte van de deelgebieden v . De dispersie variantie zal ook afnemen naarmate de ruimtelijke samenhang tussen nabijgelegen waarnemingen toeneemt.

Indien van een aantal min of meer uniform over het gebied V verspreide deelgebiedjes ($i = 1, \dots, n$; $n \leq N$) de gemiddelden $z(v_i)$ bekend zijn (bijvoorbeeld op grond van een interpolatie-procedure), en bovendien n niet te klein is, is de dispersie variantie te schatten als

$$\frac{1}{n} \sum_i (z(v_i) - \bar{z})^2, \quad (2)$$

met

$$\bar{z} = \frac{1}{n} \sum_i z(v_i).$$

Doorgaans wordt de ruimtelijke samenhang van een toevalsveld $Z(u)$, uitgedrukt door de semi-variantie $\gamma(h)$, gedefiniëerd als $\frac{1}{2} \text{Var}[Z(u) - Z(u+h)]$. Hierin is h de onderlinge

afstandsvector, en aangenomen wordt dat de semi-variantie niet van u afhangt. Tevens wordt hier aangenomen dat de semi-variantie isotroop is m.a.w. dat de lengte van h slechts relevant is. Bij afwezigheid van trend geldt dan

$$\gamma(h) = \frac{1}{2} E[(Z(u) - Z(u+h))^2]. \quad (3)$$

De semi-variantie kan worden geschat door de steekproef semi-variantie

$$g(h) = \sum_i [z(u_i) - z(u_i + h)]^2 / (2n(h)), \quad (4)$$

waarin

$n(h)$: aantal waarnemingsparen op afstandsvector h .

Aan de steekproef semi-variantie kan een semi-variantiemodel worden aangepast, bijvoorbeeld een lineair model

$$\gamma(h) = C\delta(h) + ah, \quad (5)$$

of een exponentieel model

$$\gamma(h) = C\delta(h) + a(1 - \exp(-bh)), \quad (6)$$

waarin C , a en b parameters zijn, en $\delta(h)$ de functie is waarvoor geldt:

$$\delta(h) = 1 \text{ als } h \neq 0 \quad \text{en} \quad \delta(h) = 0 \text{ als } h = 0.$$

De term $C\delta(h)$ is het nugget-effect in de semivariantie.

Bij het exponentiële model stabiliseert de semi-variantie zich voor grote onderlinge afstanden h naar een constante waarde ($C+a$), de "sill" genaamd. Deze is gelijk aan de puntvariantie $\text{Var } Z(u)$. Het lineaire model heeft geen sill.

Voor de dispersie variantie $D^2(v/V)$ geldt (Journel en Huijbregts, 1978; p.67):

$$D^2(v/V) = \bar{\gamma}(V, V) - \bar{\gamma}(v, v), \quad (7)$$

waarin

$\bar{\gamma}(V, V)$: gemiddelde waarde van de semi-variantie als beide punten het gebied V doorlopen

$\bar{\gamma}(v, v)$: analoog voor v .

Met behulp van (7) kan de werkelijke dispersie variantie worden geschat, indien het semi-variantie model bekend is.

Voor een bepaald gebied V bestaat de limiet $D^2(0/V)$ van $D^2(v/V)$ als de grootte van de v 's naar nul nadert. Indien het semi-variogram geen "sill" heeft, blijft $D^2(0/V)$ aangroeien als de grootte van V toeneemt. De puntvariantie bestaat dan niet; de "puntvariantie" $D^2(0/V)$ binnen een begrensde gebied V blijft echter een zinvol gegeven.

Een andere kijk op de dispersievariantie kan men als volgt verkrijgen:

$$D^2(v/V) = E \left(\frac{\sum_{i=1}^N (Z(V) - Z(v_i))^2}{N} \right) = \frac{1}{N} \sum_{i=1}^N (E(Z(V) - Z(v_i)))^2,$$

d.i. het gemiddelde van $E(Z(V) - Z(v_i))^2$; deze laatste uitdrukking is de variantie van de voorspellingsfout van $Z(V)$ door $Z(v_i)$.

Er blijkt te gelden (Journal en Huijbregts, 1978; p.54):

$$D^2(v/V) = \frac{1}{N} \sum_{i=1}^N (2\bar{\gamma}(v_i, V) - \bar{\gamma}(v_i, v_i) - \bar{\gamma}(V, V)). \quad (8)$$

Middelen van het rechterlid resulteert inderdaad in

$$D^2(v/V) = 2\bar{\gamma}(V, V) - \bar{\gamma}(v, v) - \bar{\gamma}(V, V) = \bar{\gamma}(V, V) - \bar{\gamma}(v, v). \quad (9)$$

3 INTERPOLATIE EN DISPERSIE VARIANTIE

In hydrologisch onderzoek worden veelal gebiedsgemiddelden $z(v_i)$ ($i = 1, \dots, n$) geschat met behulp van interpolatiemethoden, op grond van "punt" waarnemingen $z(u_j)$ ($j = 1, \dots, m$) binnen een groter gebied V . In het geval van interpolatie naar punten geldt daarbij: $v_i \rightarrow 0$ ($i = 1, \dots, n$).

We beschouwen hier twee interpolatiemethoden, namelijk kriging en de (Thiessen-) polygon methode, welke laatste methode veelal wordt toegepast bij het vervaardigen van thematische kaarten. De dispersie variantie van de geïnterpoleerde velden wordt aangeduid als respectievelijk $DK^2(v/V)$ en $DP^2(v/V)$. In geval van een volledige partitie van het gebied V in N gebieden geldt voor $DK^2(v/V)$ (Journal en Huijbregts, 1978, p. 451)

$$DK^2(v/V) = D^2(v/V) - \bar{\sigma}_{Kv}^2 + \sigma_{Kv}^2 \quad (10)$$

waarin

$\bar{\sigma}_{Kv}^2$: gemiddelde kriging variantie (gemiddelde verwachte kwadratische fout bij het kringen) voor de eenheden v_1, \dots, v_N

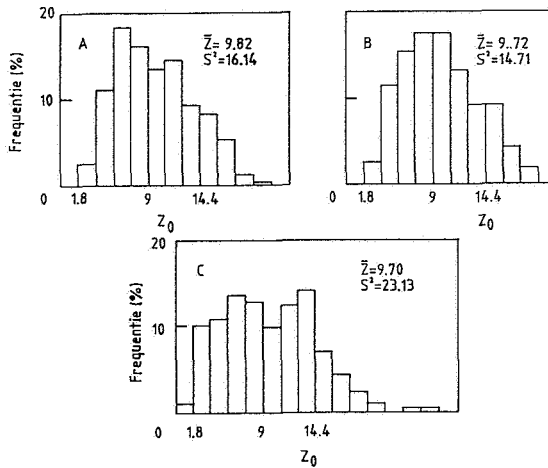
σ_{Kv}^2 : kriging variantie voor V.

Aangezien $\bar{\sigma}_{Kv}^2 - \sigma_{Kv}^2$ altijd positief is, geldt dat $DK^2(v/V)$ kleiner is dan $D^2(v/V)$. De grootte van het "smoothing effect" wordt bepaald door $\bar{\sigma}_{Kv}^2$, omdat σ_{Kv}^2 veel kleiner is. Vergelijking 10 geldt bij een "globale" kriging procedure (de volledige data configuratie wordt bij elke schatting gebruikt) en is een gevolg van de eigenschap dat een geschat increment ongecorrleerd is met de schattingsfout.

Voor kriging schattingen gebaseerd op lokale omgevingen geldt de bovenstaande gelijkheid niet. Journel en Huijbregts geven ook voor deze situatie een formule. Geheel overeenkomstig de verwachting is de reductie in dispersie variantie bij kriging uitgaande van een lokale omgeving geringer. Er is zelfs de mogelijkheid van overschatten van de dispersie variantie. Journel en Huijbregts (1978, p. 452) merken in dit verband echter op dat het "smoothing effect" $DK^2(v/V) \leq D^2(v/V)$ in de hun bekende toepassingen altijd is opgetreden. Stel dat de neerslagen over een deelgebied v binnen een groter gebied V zijn voorspeld met behulp van kriging van puntneerslagen. Op grond van de kriging schattingen $\hat{z}(v_i)$ kan de kansverdeling van gebiedsneerslagen over v worden geschat. De variantie behorende bij deze kansverdeling schat $DK^2(v/V)$. De geschatte kansverdeling zal echter op grond van Vergelijking 10 een dunnere "rechter staart" hebben dan de werkelijke kansverdeling. Aangezien veel hydrologische uitspraken juist gebaseerd zijn op analyse van extreme neerslagen dan wel overschrijdingen van een drempelwaarde, is dit ongewenst.

Journel en Huijbregts (p. 460 ev) illustreren de reductie in dispersie variantie aan de hand van het hiervoor vermelde voorbeeld van ertsgehalten. Na schatting van het semi-variogram werden zowel met behulp van kriging, gebaseerd op lokale omgevingen, als met de polygoon methode, gebiedsgemiddelden $z(v_i)$ geschat voor de binnenste 384 deelgebieden. Voor de gemiddelde kriging variantie gold $\bar{\sigma}_{Kv}^2 = 1.35$.

Figuur 1 geeft de histogrammen van de "werkelijke" gebiedsgemiddelden en van de geschatte gebiedsgemiddelden met behulp van kriging of de polygoonmethode. Uit vergelijking van de histogrammen A en B blijkt de variantie van de met kriging geschatte gebiedsgemiddelden geringer dan die van de "werkelijke" gebiedsgemiddelden, en inderdaad wordt vergelijking 10 bevestigd, immers $DK^2(v/V) = 14.71 \approx 16.14 - 1.35$.



Figuur 1. Histogrammen van de "werkelijke" gebiedsgemiddelde ertsgehalten (A) en van de geschatte gebiedsgemiddelden op grond van de kriging methode (B) en de polygoon methode (C).

Van de polygoon schatting van $z(v_i)$, waarbij $z(v_i)$ wordt gelijk gesteld aan de waarde in het centrale punt binnen v_i , blijkt in dit voorbeeld de geschatte variantie groter te zijn dan die van de "werkelijke" gebiedsgemiddelden. Volgens Journel en Huijbregts (p. 452) is dit in het algemeen het geval. De variantie van gemiddelde neerslagen over deelgebieden v_i binnen een gebied V zal, indien voorspeld met behulp van de polygoon methode, dus in het algemeen onrealistisch hoog zijn, en onrealistisch laag indien voorspeld met de kriging methode.

Voor een goed begrip dient bij de bovenstaande conclusie echter de volgende kanttekening te worden geplaatst. Bij voorspelling met behulp van kriging is er geen feitelijke ondergrens aan de omvang van het gebied v_i waarvoor de voorspelling kan plaatsvinden. Bij de polygoon methode is dat wel het geval: het aantal polygoonen is namelijk gelijk aan het aantal waarnemingspunten (bij de Thiessen polygoon methode) of gelijk aan het aantal onderscheiden landschappelijke eenheden (bij een bodemkundige kartering). Binnen deze polygoonen is elke variatie evenwel afwezig.

4 CONDITIONELE SIMULATIE

Indien men velden wil simuleren van ruimtelijke variabelen, waarbij de meetwaarden in de meetpunten onveranderd worden overgenomen en de dispersie variantie niet wordt gladgestreken, kan gebruik worden gemaakt van de techniek van conditionele simulatie (Journel en Huijbregts, 1978; p. 493 en volgende). Hierbij wordt, uitgaande van waarnemingen $z(u_j)$, $j = 1, \dots, m$, in een gebied V een realisatie $\hat{z}_{cs}(u)$ van het toevalsveld geschat die

- conditioneel is op de waarnemingen, dus

$$\hat{z}_{cs}(u_j) = z(u_j), j = 1, \dots, m;$$
- zuiver is; dus $E(\hat{z}_{cs}(u) - Z(u)) = 0$
- een semi-variogram heeft dat overeenkomt met dat van $z(u)$.

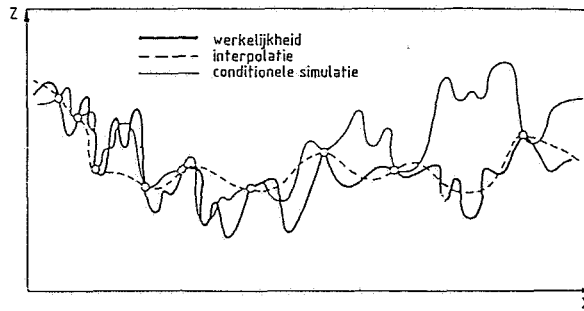
De procedure verloopt als volgt:

- (1) Maak een kaart, dat wil zeggen, schat een realisatie $\hat{z}(u)$ in de punten van een fijnmazig raster, met behulp van kriging en op grond van de waarnemingen en het (eventueel geschatte) semi-variogram.
- (2) Simuleer bij datzelfde semi-variogram, bijvoorbeeld met de "turning bands" methode (Appendix I) onconditionele waarden $z_s(u)$ in de waarnemings- en rasterpunten.
- (3) Interpoleer de simulatie $z_s(u)$ naar de rasterpunten met behulp van kriging, op grond van slechts de gesimuleerde waarden in de waarnemingspunten. Dit resulteert in $\hat{z}_s(u)$. In de waarnemingspunten geldt: $z_s(u) = \hat{z}_s(u)$, omdat kriging een exacte interpolator levert.
- (4) De conditionele simulatie \hat{z}_{cs} wordt nu voor elke rasterpunt gevonden als:

$$\hat{z}_{cs}(u) = \hat{z}(u) + [z_s(u) - \hat{z}_s(u)]. \quad (11)$$

- (5) Door de stappen (2) tot en met (4) te herhalen, kunnen steeds nieuwe conditionele simulaties worden gevonden.

Figuur 2 geeft een voorbeeld van de conditionele simulatie \hat{z}_{cs} en de kriging interpolatie \hat{z} . Men ziet dat de variantie van de conditionele simulatie veel groter is dan die van de geïnterpoleerde waarden. Er geldt dat de dispersie variantie van de conditionele simulatie, $\text{var}(\hat{z}_{cs})$, gelijk is aan de dispersie variantie van $Z(u)$. Dit geldt weer omdat een geschat increment ongecorrleerd is met de schattingsfout. Alhoewel hier conditionele simulatie is toegelicht aan de hand van interpolatie naar puntlocaties, is het bovenstaande onafhankelijk van de grootte van het gebied waarnaar wordt geïnterpoleerd. De plaatscoördinaat u mag dus overal worden gelezen als de aanduiding van een deelgebied v .



Figuur 2. Werkelijke waarden van een ruimtelijke variabele langs een raai, alsmede kriging schattingen en een conditionele simulatie (naar: Journel en Huijbregts, 1978).

5 TEMPORELE COVARIANTIESTRUCTUUR

Behalve de ruimtelijke covariantiestructuur is ook de temporele van belang, indien bijvoorbeeld opeenvolgende neerslagvelden worden geschat voor neerslag-afvoer modellering. In feite wordt dan de ruimtelijke-temporele variabele $Z(v,t)$ beschouwd, waarbij met $Z(v,t)$ in dit voorbeeld de gemiddelde neerslaghoeveelheid over gebied v en over tijdsinterval t wordt bedoeld.

Voor de eerste orde autocorrelatiecoëfficiënt, $\rho^*(1)$, van kriging schattingen van puntneerslagen van een zekere duur, bijv. dagneerslagen, geldt (zie Appendix II):

$$1 + \rho^*(1) = (1 + \rho(1)) \{ \overline{\rho_2(h)} / \overline{\rho_1(h)} \}, \quad (12)$$

waarin

$\rho(1)$: eerste orde autocorrelatiecoëfficiënt voor puntneerslagen van die duur

$\overline{\rho_1(h)}$: gemiddelde over V van de ruimtelijke correlatiecoëfficiënt van puntneerslagen van die duur

$\overline{\rho_2(h)}$: gemiddelde over V van de ruimtelijke correlatiecoëfficiënt van puntneerslagen van dubbele duur (bijv. 2-daagse neerslagen).

Voor neerslagen over een zeker deelgebied v geldt, analoog aan Vergelijking 12:

$$1 + \rho_v^*(1) = (1 + \rho_v(1)) \{ \overline{\rho_2(h)} / \overline{\rho_1(h)} \}, \quad (13)$$

waarin

$\rho_v^*(1)$: eerste orde autocorrelatiecoëfficiënt voor kriging schattingen van de gebiedsneerslag van zekere duur

$\rho_v(1)$: eerste orde autocorrelatiecoëfficiënt voor gebiedsneerslagen van die duur.

Vergelijkingen 12 en 13 maken het mogelijk om de gevolgen van interpoleren op het autocorrelatieverloop na te gaan. Aangezien geldt $\rho_2(h) > \rho_1(h)$ impliceert Vergelijking 13 een verhoging van de eerste orde autocorrelatiecoëfficiënt van de kriging schattingen, dus een gladstrijkend effect.

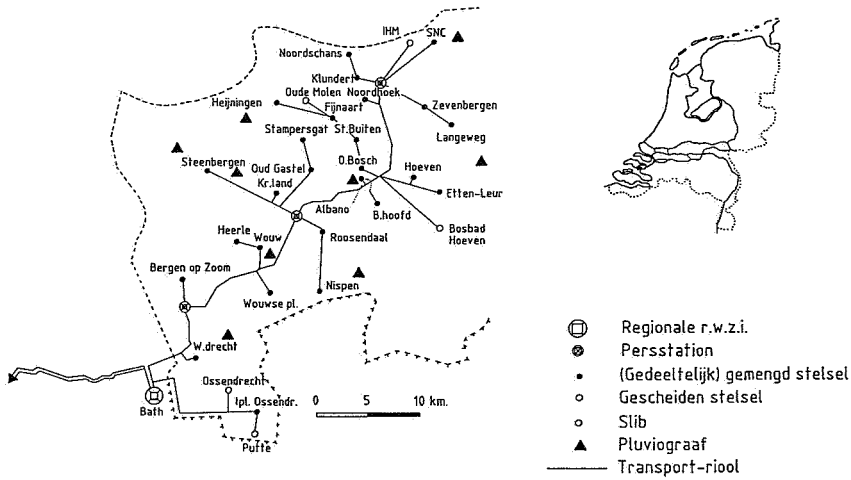
In feite zijn de gemiddelde ruimtelijke correlatiecoëfficiënten in de Vergelijkingen 12 en 13 gewogen gemiddelden. Aangezien bij kriging van gebiedsneerslagen op grond van puntneerslagen de gewichten λ_i ($i = 1, \dots, n$) iets zullen verschillen van die bij punt-kriging, dienen de gewogen gemiddelden in Vergelijking 12 en 13 op iets andere wijze te worden bepaald. De fout die gemaakt wordt door in beide vergelijkingen het rekenkundig gemiddelde te nemen, zal echter gering zijn.

6 VOORBEELD

In STORA (1990) is de waterbalans doorgerekend van het aanvoerstel van de regionale waterzuiveringsinstallatie (rwzi) Bath. Dit aanvoerstelsel bestaat uit een persleiding met 26 daarop aangesloten gemengde rioleringsstelsels, 3 gescheiden rioleringsstelsels, en slibaanvoer vanuit een tweetal plaatsen (zie Figuur 3 en Tabel 1).

Tabel 1. Op de rwzi Bath aangesloten rioleringsstelsels (uit: STORA, 1990).

Gemengde stelsels:		
Shell Nederland (SNC)	Etten-Leur	
Noordschans	Hoeven	Roosendaal
Klundert	Bosschenhoofd	Nispen
Zevenbergen	Albano	Heerle
Langeweg	Stampersgat	Wouw
Noordhoek	Oud Gastel	Wouwse Plantage
Fijnaart	Steenbergen	Bergen op Zoom
Heijningen	Kruisland	Woensdrecht
Standaardbuiten	Oudenbosch	Legerplaats Ossendrecht
Gescheiden stelsels:		
Industrieterrein Moerdijk	Bosbad Hoeven	Oude Molen
Slibaanvoer:		
Ossendrecht	Putte	



Figuur 3. Overzicht aanvoerstelsel regionale rwzi Bath, situatie 1983-1985 (uit: STORA, 1990).

Aangezien de berging in de persleiding verwaarloosbaar is, bestaat de waterbalans van het totale aanvoerstelsel in feite uit de optelsom van de waterbalansen van de aangesloten kernen. De waterbalansen zijn opgesteld voor berekeningsintervallen van een half uur, en kunnen als volgt worden geschreven:

$$Q(i, t) = c_i v_i \hat{z}(v_i, t - t') - Q_0(i, t) + \Delta B(i, t) + Q_d(i, t) + Q_b(i, t), \quad (14)$$

waarin

- $Q(i, t)$: vanuit kern i in de persleiding geïnjecteerd debiet gedurende tijdsinterval t ; deze debieten zijn geregistreerd [$L^3 T^{-1}$]
- c_i : afvoercoëfficiënt voor kern i ; deze is bepaald door middel van calibratie [-]
- v_i : verhard oppervlak van kern i ; deze is bekend [L^2]
- t' : vertragingstijd tussen het vallen en tot afvoer komen van de neerslag, eveneens door calibratie bepaald [-]
- $\hat{z}(v_i, t-t')$: kriging schatting van de gebiedsneerslag over v_i voor het tijdsinterval $t-t'$ [LT^{-1}]
- $Q_0(i, t)$: overstortend debiet [$L^3 T^{-1}$]
- $\Delta B(i, t)$: verandering van de berging in het rioleringsstelsel, negatief bij afname van de beschikbare berging [$L^3 T^{-1}$]

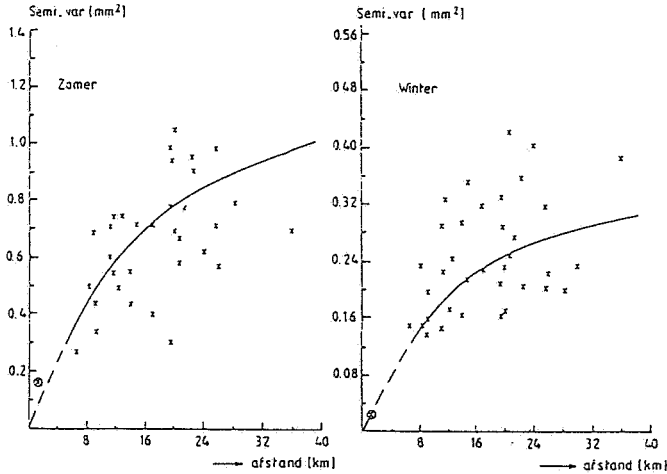
- $Q_d(i,t)$: droogweerafvoer; deze is bekend [L^3T^{-1}]
 $Q_b(i,t)$: aanvoer vanuit "bovenstroomse" kernen; deze is bekend
 op grond van de registraties van die kernen [L^3T^{-1}].

In feite wordt in Vergelijking 14 dus voor elke kern voor elk interval van een half uur de som $(-Q_0(i,t) + \Delta B(i,t))$ berekend. Aangezien de totale berging van elk rioleringsstelsel bekend is, is $\Delta B(i,t)$ naar boven begrensd. Indien de berging groter zou moeten worden dan de totale beschikbare berging, treedt een overstorting op.

De neerslagterm $\hat{z}(v_i, t-t')$ in Vergelijking 14 is voor de 26 aangesloten kernen met een gemengd rioolstelsel bepaald met behulp van kriging. De hierbij gebruikte gegevens zijn afkomstig van 9 pluviografen in het gebied (Figuur 3). De oppervlakte van het totale gebied V waarbinnen de kernen gelegen zijn, bedraagt circa 450 km^2 . Voor de meeste kernen geldt $v_i \ll V$, en is voor de schatting \hat{z} gebruik gemaakt van punt-kriging, aangezien onder deze voorwaarden in dit voorbeeld de verschillen tussen punt- en blok-kriging verwaarloosbaar zijn.

De waterbalans van het aanvoerstelsel is doorgerekend voor een 37-tal zware neerslagen. Daarbij is onderscheid gemaakt tussen het zomer- (mei t/m september) en het winterseizoen (rest van het jaar). In Figuur 4 zijn de geschatte semi-varianties en aangepaste semi-variogrammen voor deze neerslagen afgebeeld. Aanpassing van de semi-variogrammen geschiedde met een gewogen kleinste kwadratenmethode (weging door \sqrt{n}) in verband met het sterk uiteenlopend aantal waarnemingsparen waarop de geschatte semi-varianties berusten. Indien de semi-varianties worden geschat met behulp van een enkelvoudige realisatie benadering, verdient het aanbeveling om bij het aanpassen van semi-variogrammodellen geen gebruik te maken van geschatte semi-varianties voor onderlinge afstanden groter dan de helft van de karakteristieke lengte van het gebied. In het geval van West-Brabant bedraagt deze karakteristieke lengte ongeveer 40 km. Aangezien hier gebruik is gemaakt van de meervoudige realisatie benadering, is het semi-variogrammodel aangepast aan alle geschatte semi-varianties. De semi-varianties in Figuur 4 voor een onderlinge afstand van 1.44 km zijn berekend met behulp van gegevens van drie registrerende regenmeters in het hydrologisch proefgebied van Hupsel. Aan de geschatte semi-varianties in Figuur 4 zijn exponentiële semi-variogrammodellen aangepast zonder nugget-effect met als resultaat

zomer: $\gamma(h) = 1.095(1-\exp(-0.065 h))$
 winter: $\gamma(h) = 0.325(1-\exp(-0.073 h))$



Figuur 4. Geschatte semi-varianties en aangepaste exponentiële semi-variogrammen.

In dit voorbeeld zullen de effecten van interpolatie toegespitst worden op de gevolgen die de interpolatie heeft op de dispersie variantie van de neerslagen van de 26 kernen. Omdat de 26 kernen min of meer uniform verspreid liggen zal deze dispersie variantie ongeveer gelijk zijn aan $D^2(v/V)$. In Tabel 2 worden de geschatte dispersie varianties van de kriging schattingen (\hat{DK}^2) en de polygoon schattingen (\hat{DP}^2) vergeleken met de theoretische waarden van de dispersie variantie ($D^2(v/V)$) en de steekproefvariantie van de gemeten puntneerslagen, s^2 , voor het hierboven gedefiniëerde zomer- en winterseizoen. $D^2(v/V)$ is berekend met behulp van Vergelijking 9. Daarbij is uitgegaan van $v = 0 \text{ km}^2$ en $V = 450 \text{ km}^2$, waarbij het gebied geschematiseerd is tot een rechthoek met lengte-breedte verhouding van 2. In dat geval (Matérn, 1960) is de gemiddelde onderlinge afstand tussen punten in het gebied $\sqrt{450} * 0.5691 = 12 \text{ km}$. $D^2(v/V)$ wordt met een geringe overschatting gesteld op $\gamma(12 \text{ km})$. DK^2 en DP^2 zijn geschat als:

$$\frac{1}{N_1} \sum_{i=1}^{N_1} \left[\frac{1}{N_2(i)} \sum_{j=1}^{N_2(i)} \left\{ \frac{1}{25} \sum_{k=1}^{26} (\hat{z}(v_k) - \hat{z}(V))^2 \right\} \right], \quad (15)$$

waarin

- N_1 : aantal neerslaggebeurtenissen in zomer, resp. winter
 $N_2(i)$: aantal halfuurlijkse intervallen in neerslaggebeurtenis i
 $\hat{z}(v_k)$: kriging of polygoon schatting van de neerslag over deelgebied v_k over interval j van neerslaggebeurtenis i
 $\hat{z}(V)$: $\sum_{k=1}^{26} \hat{z}(v_k)/26$.

Tabel 2. Theoretische dispersie variantie $D^2(v/V)$ en geschatte waarden van $DK^2(v/V)$ en $DP^2(v/V)$ [mm^2].

Dispersie variantie	Zomer	Winter
$\hat{D}K^2$	0.23	0.08
$\hat{D}P^2$	1.02	0.58
$D^2(v/V)$	0.59	0.19
s^2	1.01	0.54

De dispersie variantie van de kriging schattingen blijkt aanzienlijk lager, en die van de polygoonschattingen aanzienlijk hoger te zijn dan de theoretische waarde van de dispersie variantie. Uit Tabel 2 blijkt dat zelfs geldt: $\hat{D}P^2 > s^2$. Om de invloed aan te geven van een toename van het aantal gebieden v_i waarvoor de neerslag moet worden geschat op $\hat{D}P^2$, is $\hat{D}P^2$ herberekend, maar nu met een (arbitrair) 25 maal zo groot aantal gebieden v_i per polygoon. Dit leidt niet tot een noemswaardige reductie in $\hat{D}P^2$. Inderdaad kan worden afgeleid (Appendix III) dat $\hat{D}P^2$ bij benadering gelijk is aan s^2 .

Schattingen van de autocorrelatiecoëfficiënten van de kriging schattingen van de neerslagen over de op de persleiding aangesloten kernen, $\rho_v^*(k)$, staan vermeld in Tabel 3, voor orde $k = 1, 2$ en 3 . De tabel vermeldt tevens de schattingen van de autocorrelatiecoëfficiënten voor puntneerslagen, $\rho(k)$, voor $k = 1, 2$ en 3 . De eerste orde autocorrelatiecoëfficiënten van de werkelijke, maar niet waargenomen gebiedsneerslagen, $\rho_v(1)$, zijn geschat met behulp van Vergelijking 12. Het quotiënt van gebiedsgemiddelden van ruimtelijke correlatiecoëfficiënten in het rechterlid van Vergelijking 12 is geschat door gebruik te maken van regenwaarnemingen in het Hupselsche Beek gebied. De resultaten zijn eveneens vermeld in Tabel 3. Het autocorrelatieverloop van de werkelijke gebiedsneerslag wordt door de met behulp van kriging geschatte gebiedsneerslag redelijk benaderd. Uit Tabel 3 blijkt eveneens dat bij gebruik maken van polygoonschattingen van de neerslag, de neerslagschattingen een te geringe autocorrelatie hebben.

Tabel 3. Schattingen van de autocorrelatiecoëfficiënten van puntneerslagen, $\rho(k)$, van kriging schattingen van de neerslagen over de op de persleiding aangesloten kernen, $\rho_v^*(k)$, en van de eerste orde autocorrelatiecoëfficiënten, $\rho_v(1)$, van de werkelijke gebiedsneerslag.

Orde k	Zomer			Winter		
	$\rho(k)$	$\rho_v^*(k)$	$\rho_v(k)$	$\rho(k)$	$\rho_v^*(k)$	$\rho_v(k)$
1	0.35	0.45	0.48	0.50	0.61	0.57
2	0.15	0.20		0.32	0.30	
3	0.06	0.07		0.10	0.13	

Vervolgens is nagegaan hoe de dispersie variantie en de autocorrelatiestructuur van de kriging schattingen verbeterd kan worden door conditionele simulatie van neerslagvelden. Om de hoeveelheid rekenwerk te beperken zijn geen conditionele neerslagvelden gesimuleerd van alle 37 in Stora (1990) beschouwde neerslagvelden, maar slechts van één: die van 6 februari 1984. Deze heeft een duur van 27 uur. Gemiddeld over West-Brabant valt er 18.5 mm neerslag; over Zevenbergen valt in totaal 27.1 mm neerslag. Er zijn 10 conditionele simulaties uitgevoerd, waarbij gebruik is gemaakt van de turning bands methode. Het aantal van 10 conditionele simulaties is te gering om algemeen geldende conclusies te trekken. De gemiddelde dispersie variantie van de 10 conditionele simulaties is een factor 2.1 hoger dan die van de kriging schattingen. Dit komt goed overeen met de reductie voor winterse neerslagen van \hat{DK}^2 ten opzichte van $D^2(v/V)$ in Tabel 2.

Voor deze verbetering moet echter een prijs worden betaald in de vorm van een verslechtering van het autocorrelatieverloop. De gemiddelde eerste orde autocorrelatiecoëfficiënt van de 10 conditionele simulaties bedraagt 0.31 tegen 0.50 voor de kriging schattingen.

De invloed van conditionele simulaties op de resultaten van het hierboven beschreven simulatiemodel van het aanvoerstelsel van de rwzi Bath, is nagegaan voor één van de op de persleiding aangesloten kernen, Zevenbergen. Het gemaal Zevenbergen verpompt ook de afvoer vanuit Langeweg (Figuur 3). Het gemiddeld aantal overstortingen per 10 conditionele simulaties bedraagt 2.2 (2 voor de kriging schattingen), terwijl de gemiddelde overgestorte hoeveelheid 6932 m³ bedraagt (4627 m³ voor de kriging schattingen). Deze verschillen kunnen ten dele worden verklaard doordat de gemiddelde neerslaghoeveelheid voor de 10 conditionele simulaties 30.2 mm is, tegen 27.1 mm voor de kriging schatting. Dit is enerzijds

een gevolg van het feit dat conditionele simulatie kan resulteren in negatieve neerslagen (die vervolgens zijn gecorrigeerd tot nul). Anderzijds is het aantal beschouwde conditionele simulaties (10) nogal laag. Echter, ook in STORA (1990) werd een forse toename van overgestorte hoeveelheden (met een factor 1.44) gevonden bij gebruik maken van een simulatietechniek. Deze simulatietechniek was echter niet conditioneel op de waarnemingen, en nam de ruimtelijke structuur van de neerslag niet in beschouwing.

7 CONCLUSIES

Als kaarten die zijn verkregen door het toepassen van een interpolatie- dan wel karteringsprocedure worden gebruikt als invoer in een simulatiemodel, verdient de covariantiestructuur van de kaarten (geschatte realisaties van toevalsvelden) bijzondere aandacht. Algemeen kan worden gesteld dat de polygoonmethode tot onbevredigende resultaten leidt: zowel de ruimtelijke als de temporele variatie worden overschat. Dit vindt zijn weerslag in een te hoge dispersie variantie en een te lage autocorrelatiecoëfficiënt.

Interpolatie middels kriging leidt tot onderschatten van de ruimtelijke variatie; de temporele variatie wordt echter bevredigend weergegeven. Met behulp van conditionele simulatie kan de ruimtelijke variatie waarheidsgetrouw worden weergegeven. De prijs die hiervoor moet worden betaald is echter een verslechtering van de weergave van het autocorrelatieverloop.

De methode van conditionele simulatie dient dan ook uitgebreid te worden in de zin dat ook de temporele covariantiestructuur bewaard blijft. Ook de mogelijkheid dat het simulatieresultaat negatief is, dient te worden uitgesloten. Rekentechnisch is conditionele simulatie met behulp van de turning bands methode een praktisch alternatief voor andere simulatiemethoden van toevalsvelden.

De toename van zowel de ruimtelijke als temporele variatie van de resultaten van conditionele simulatie ten opzichte van de kriging schattingen, vindt zijn weerslag in de resultaten van simulatiemodellen. In het hierboven gegeven voorbeeld van een neerslag-afvoermodel van een stedelijk gebied neemt de hoeveelheid overgestort water aanzienlijk toe.

LITERATUUR

- ABRAMOWITZ, M., and I.A. STEGUN, 1965. Handbook of mathematical functions. Dover, New York.
- BRAS, R.L., and I. RODRIGUEZ-ITURBE, 1984. Random functions and hydrology. Addison-Wesley, Reading.
- BUISHAND, T.A., 1977. De variantie van de gebiedsneerslag als functie van puntneerslagen en hun onderlinge samenhang. Mededelingen Landbouwhogeschool 77-10, Wageningen.
- JOURNEL, A.G. and HUIJBREGTS, Ch.J., 1978. Mining geostatistics. Academic Press, Londen.
- MATERN, B., 1960. Spatial Variation. Stochastic models and their application to some problems in forest surveys and other sampling investigations. Meddelanden fran Statens Skogsforskningsinstitut Band 49-Nr. 5.
- MATHERON, G., 1973. The intrinsic random functions and their applications. Advances of Applied Probability 5: 439-468.
- ROUHANI, S. and WACKERNAGEL, H., 1990. Multivariate geostatistical approach to space-time data analysis. Water Resources Research 26: 585-591.
- STORA, 1990. Eindrapport Stora-onderzoek 3.1.0: Dimensionering regionale rioolwatertransportstelsels. Stora, Den Haag.

APPENDIX I TURNING BANDS METHODE

Bij het simuleren van de ongeconditioneerde toevalsvelden $z_s(u)$ naar een fijnmazig raster in 2 of 3 dimensies stuit men bij het toepassen van gebruikelijke methodes, zoals bijvoorbeeld de Choleski methode, al gauw op de excessieve aanspraken die deze methoden maken op computergeheugen en computertijd. Een alternatief is dan de zogenaamde "turning bands" methode (Matheron, 1973). Hier wordt de beschrijving van de methode in Bras en Rodriguez-Iturbe (1984) gevolgd.

Verondersteld wordt dat het toevalsveld gemiddelde nul heeft en Normaal verdeeld is. De simulatie van een toevalsveld in 2 of 3 dimensies wordt nu opgebouwd als de som van simulaties z_i langs de lijnen l_i ($i = 1, \dots, L$), waarvan de hoek θ_i met de horizontale x-as uniform verdeeld is over de eenheidscirkel, respectievelijk - bol (zie voor het 2-dimensionale geval Figuur I-1, overgenomen uit Bras en Rodriguez-Iturbe, 1984). Voor een punt u_0 geldt:

$$z_s(u_0) = \frac{1}{\sqrt{L}} \sum_{i=1}^L z_i(\xi_{0i}). \quad (I-1)$$

De benodigde semi-variantie structuur langs een lijn l_i kan worden gerelateerd aan de benodigde (eventueel geschatte) semi-variantie in 2 of 3 dimensies. In het geval van een exponentieel semi-variogram in 2 dimensies, $\gamma_2(h)$, volgens:

$$\gamma_2(h) = a(1 - \exp(-bh)),$$

geldt voor de semi-variantie langs een lijn, $\gamma_1(h)$:

$$\gamma_1(h) = a \left[\frac{\pi}{2} bh \{I_0(bh) - L_0(bh)\} \right],$$

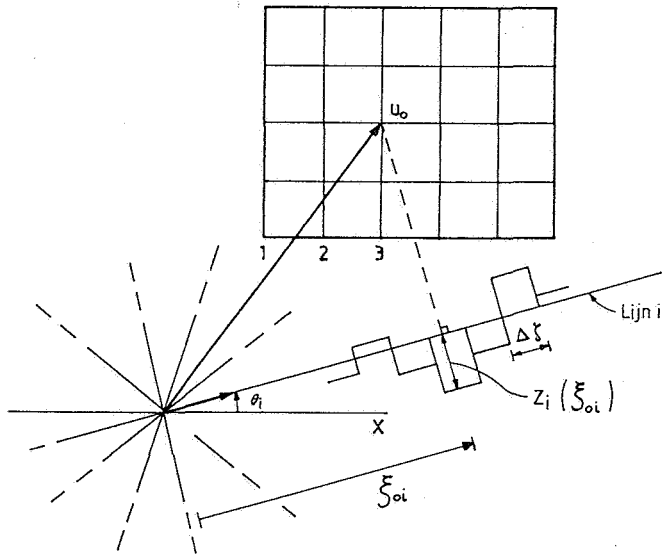
waarin

I_0 : Bessel functie van orde nul

L_0 : gemodificeerde Struve functie van orde nul.

Benaderingen voor beide standaardfuncties $I_0(\cdot)$ en $L_0(\cdot)$ zijn te vinden in Abramowitz en Stegun (1965).

Simulatie van $z_s(u)$ met de "turning bands" methode reproduceert de univariate verdeling van de toevalsfunctie en de eerste twee momenten van de toevalsfunctie: het gemiddelde (door bij de op de hierboven aangegeven wijze verkregen simulatie het gemiddelde op te tellen), en de semi-variantie. Alhoewel hiertoe voor L in Vergelijking I-1 moet gelden, $L \rightarrow \infty$, blijkt $L = 15$ te voldoen in praktische toepassing, zowel voor simulaties in twee als in drie dimensies (Journel en Huijbregts, 1978).



Figuur I-1. Schematische voorstelling van het verkrijgen van het ongeconditioneerde toevalsveld $z_s(u)$ op grond van simulaties langs lijnen l_i ($i = 1, \dots, L$) met behulp van de "turning bands" methode.

APPENDIX II AFLEIDING VAN VERGELIJKING 12

Beschouw een ruimtelijk-temporele variabele $Z(u,t)$, en de kriging schatting in een punt met coördinatenvector u_0 , $\hat{Z}(u_0, t) = \sum_j \lambda_j Z(u_j, t)$. Voor de variantie van de som van krigingschattingen in twee opvolgende tijdsintervallen, t en $t+1$, geldt enerzijds:

$$\begin{aligned}
 \text{Var}[\hat{Z}(u_0, t) + \hat{Z}(u_0, t + 1)] &= 2\text{Var}[\hat{Z}(u_0, t)] + 2\text{Cov}[\hat{Z}(u_0, t), \hat{Z}(u_0, t + 1)] \\
 &= 2\text{Var}[\hat{Z}(u_0, t)] + 2\rho^*(1)\text{Var}[\hat{Z}(u_0, t)] \\
 &= 2(1 + \rho^*(1))\text{Var}[\sum_i \lambda_i Z(u_i, t)] \\
 &= 2(1 + \rho^*(1))\sum_i \sum_j \lambda_i \lambda_j \text{Cov}[Z(u_i, t), Z(u_j, t)] \\
 &= 2(1 + \rho^*(1))\sum_i \sum_j \lambda_i \lambda_j \rho_1(u_i, u_j) \text{Var}Z(u, t) \\
 &= 2(1 + \rho^*(1))\overline{\rho_1(h)} \text{Var}[Z(u, t)], \tag{II - 1}
 \end{aligned}$$

waarin

- $\rho^*(1)$: eerste orde autocorrelatiecoëfficiënt voor kriging schattingen van puntneerslagen van zekere duur
- $\rho_1(u_i, u_j)$: ruimtelijke correlatiecoëfficiënt van neerslagen van zekere duur op locaties u_i en u_j
- $\overline{\rho_1(h)}$: gewogen gemiddelde over V van de $\rho_1(u_i, u_j)$'s.

Anderzijds geldt:

$$\begin{aligned}
 \text{Var}[\hat{Z}(u_0, t) + \hat{Z}(u_0, t + 1)] &= \text{Var}[\sum_i \lambda_i Z(u_i, t) + \sum_i \lambda_i Z(u_i, t + 1)] \\
 &= \text{Var}[\sum_i \lambda_i (Z(u_i, t) + Z(u_i, t + 1))] \\
 &= \sum_i \sum_j \lambda_i \lambda_j \text{Cov}\{[Z(u_i, t) + Z(u_i, t + 1)], [Z(u_j, t) + Z(u_j, t + 1)]\} \\
 &= \overline{\rho_2(h)} \text{Var}[Z(u, t) + Z(u, t + 1)] \\
 &= 2\overline{\rho_2(h)}(1 + \rho(1))\text{Var}[Z(u, t)], \tag{II-2}
 \end{aligned}$$

waarin

$\rho(1)$: eerste orde autocorrelatiecoëfficiënt voor puntneerslagen van de beschouwde duur

$\overline{\rho_2(h)}$: gewogen gemiddelde over V van de ruimtelijke correlatiecoëfficiënt van puntneerslagen van dubbele duur.

Uit (II-1) en (II-2) volgt:

$$1 + \rho^*(1) = (1 + \rho(1)) \{ \overline{\rho_2(h)} / \overline{\rho_1(h)} \}. \tag{II-3}$$

Bij de afleiding van Vergelijking II-3 voor de autocorrelatiecoëfficiënt van kriging schattingen van de puntneerslag is de afleiding gevolgd in Buishand (1977) voor de autocorrelatiecoëfficiënt van gebiedsneerslagen.

APPENDIX III VERBAND TUSSEN DP^2 EN s^2

Beschouw de puntwaarnemingen $z(u_j)$ ($j = 1, \dots, m$) binnen een gebied V aan een ruimtelijke variabele. De m waarnemingspunten leggen m polygoon vast. Laat m_j het aantal punten, resp. gebieden zijn in polygoon j ($j = 1, \dots, m$) rondom waarnemingspunt u_j , waarvoor de neerslag geschat moet worden met de polygoon methode. De steekproefvariantie van de waarnemingen kan worden geschat als:

$$s^2 = \frac{1}{m} \sum_j [z(u_j) - \bar{z}]^2, \quad (III - 1)$$

waarin

$$\bar{z} = \frac{1}{m} \sum_j z(u_j).$$

De dispersie variantie van de polygoon schattingen $DP^2(0/V) = DP^2(v/V)$ kan worden geschat, analoog aan Vergelijking III-1, door

$$\sum_j m_j [z(u_j) - \bar{z}]^2 / \sum m_j, \quad (III - 2)$$

Indien de m_j onderling niet te sterk verschillen, zijn de uitkomsten van beide vergelijkingen bij benadering gelijk.

DE ROL VAN DE GEOSTATISTIEK IN EEN INTELLIGENT GIS

P.A. Burrough

SAMENVATTING

Het toenemende belang voor het modelleren en simuleren van processen in bodem en water (gewas groei, erosie, grondwaterbeweging, enz) heeft geleid tot een willekeurige koppeling tussen GIS-databestanden en allerlei soorten modellen. Omdat de huidige Geografische Informatiesystemen niet in staat zijn de ruimtelijke variatie van niet-exacte gegevens op een stochastische wijze te kunnen beschrijven is er weinig inzicht in de kwaliteit of de betrouwbaarheid van modeluitkomsten. De ruimtelijke statistiek kan hier behulpzaam zijn door a) het beschrijven van fouten, b) het gebruik van interpolatie om het ruimtelijk oplossend vermogen te kunnen verbeteren, c) om meetnetten te kunnen optimaliseren, en d) om het voortplanten van fouten door modellen te kunnen analyseren. Als de ervaring over het goed gebruik van GIS, modellen en de ruimtelijke statistiek wordt geformaliseerd en in een aan het GIS-gebonden kennissysteem ter beschikking gesteld, zouden de gebruikers beter weten hoe ze resultaten met van te voren bepaalde kwaliteitseisen het beste zou kunnen bereiken.

1 INLEIDING

De andere bijdragen in deze series hebben enkele belangrijke aspecten van de ruimtelijke statistiek naar voren gebracht. Uit deze en andere bronnen (zie bijvoorbeeld Webster en Oliver 1990) is de *essentie* van de ruimtelijke statistiek weergegeven als a) het beschrijven van ruimtelijke- en tijdgebonden variatie in geval van geleidelijke veranderingen b) het doen van optimale voorspellingen met minimale voorspelfout uit door steekproef verkregen gegevens. De benadering van de ruimtelijke statistiek is dat het ruimtelijke verloop van de waarde van een of ander attribuut van de bodem of van het water continu is niet volledig te beschrijven. Het ruimtelijke verloop kan niet alleen door exacte structuur-eenheden of deterministische modellen beschreven worden, maar moet ook rekening houden met rondom fluctuaties. Deze rondom fluctuaties kunnen ook onderverdeeld worden in *ruimtelijke gecorreleerde variaties* en *niet ruimtelijke gecorreleerde variaties* ofwel "ruis" of "nugget". De ruimtelijk gecorreleerde variaties worden door middel van ruimtelijke covariantiefuncties (zoals variogrammen of intrinsic rondom functies) vast gelegd; de ruis geeft de grenzen aan voor nadere precisering.

De ruimtelijke statistiek is daarom voor de volgende redenen belangrijk:

1. Het geeft inzicht in de ruimtelijke structuur van niet-zichtbare attributen van het water en de bodem.
2. Het houdt rekening met niet-exact aspecten van ruimtelijke verschillen in het landschap.
3. Geleidelijke veranderingen in attribuutwaarden c.q. 'zachte' grenzen kunnen worden gemodelleerd.
4. Het geeft schattingen van voorspelfouten die in foutenanalyse en voor de optimalisatie van onderzoekprocedures kunnen worden gebruikt.

Voordat het mogelijk is te zien hoe deze aspecten van de ruimtelijke statistiek in Geografische Informatiesystemen (GIS) ingezet kunnen worden is het noodzakelijk eerst een kort overzicht van de huidige GIS te geven.

2 GEOGRAFISCHE INFORMATIESYSTEMEN

Geografische Informatie Systemen (GIS) zijn computer gereedschappen voor het bewaren, opvragen, verwerken, analyseren en weergeven van ruimtelijke gegevens (Burrough 1986). De basisgegevens zijn meestal afkomstig van topografische kaarten, thematische kaarten, teledetectie systemen (o.a. remote sensing), veldwaarnemingen en andere bronnen, zoals censusrapporten, postcodegegevens en kadastrale gegevens. Deze basisgegevens worden in een GIS opgeslagen als eenvoudige punt-, lijn- of vlakobjecten. Bijvoorbeeld, een peilbuislokatie, een beek en eenheid op een bodemkaart worden respectievelijk als punt-, lijn- en vlakobject beschreven.

Naast de ruimtelijke ligging van het object worden eveneens de kenmerken van het object (de zgn. attributen) en de relaties met andere objecten (de zgn. topologie) opgeslagen. De attributen worden vaak apart in een Relational Data Base Management System (RDBMS) opgeslagen. De vorm van een object kan op twee manieren worden beschreven: a) door middel van vectoren die de exacte vorm van het object beschrijven, of b) door middel van een raster dat de vorm van een object met behulp van gridcellen benadert. Beide systemen kunnen verschillende soorten gegevens (ook wel 'features' genoemd) in aparte lagen ('overlays') bewaren.

Vectorsystemen zijn in staat tot een grafische weergave met een kwaliteit vergelijkbaar met die van handgetekende kaarten. Vectorsystemen zijn verder noodzakelijk voor het beschrijven van de topologie van bijvoorbeeld riviersystemen en wegnetten. In vectorsystemen zijn de belangrijkste analysemethoden: Booleanoperaties (Structured Query Language, SQL), het maken van bufferzones rondom bestaande objecten (bijvoorbeeld een weg), het combineren van kaarten met verschillende thema's (bijvoorbeeld bodemgebruik en bodemtypen) door middel van overlays, en het analyseren van routes door een netwerk. In rastersystemen wordt de vorm van een object impliciet beschreven door een verzameling gridcellen. Hoe fijner het net, hoe nauwkeuriger de vorm, geleidelijke veranderingen gemodelleerd als discrete veranderingen. Beide systemen - vector en raster - gebruiken een model van de werkelijkheid die bestaat uit vastgelegde, exacte objecten en hun bijbehorende exacte attribuutwaarden.

Rastersystemen zijn ook geschikt voor het modelleren en het verwerken van geleidelijk veranderende kenmerken, zoals de door remote sensing opgenomen reflectie waarden of de

hoogte van het land (digitale terrein modellen). Een van de grootste voordelen van rastersystemen is de gelijkvormigheid van de basis elementen (vierkanten) in de kaartlagen. Deze eigenschap van rasterkaarten maakt het mogelijk per locatie (per grid cel) nieuwe attributwaarden te berekenen uit de waarden van andere attributen op dezelfde locatie. Nieuwe attributwaarden kunnen ook uit aangrenzende cellen afgeleid worden, zoals bijvoorbeeld de hellingshoek en richting van de helling (expositie) van het terrein. Deze kunnen daarna gebruikt worden bij het opstellen van afstromingsmodellen en voor het kwantitatieve analyse van afwateringsgebieden (Van Deursen, in voorbereiding). De wiskundige verwerking van rasterkaarten in GIS vindt plaats met behulp van map algebra (Burrough 1986).

Niet alle ruimtelijke fenomenen kunnen worden beschreven met behulp van exact te definiëren objecten zoals punten, lijnen, polygonen of glad verlopende vlakken. Veel aspecten van de natuur, zoals bijvoorbeeld de vorm van het landschap, de gehalten van opgeloste stoffen in grond- en oppervlaktewateren en de concentraties van chemicaliën in de lucht, kunnen alleen worden gemeten op puntlocaties. Om een vlakvullend beeld te krijgen van het ruimtelijke verloop van deze attributen, moet de puntinformatie door middel van *interpolatie* in vlakinformatie worden omgezet. Interpolatie is het voorspellen van een attributwaarde in het gebied tussen de datapunten. Voor het maken van digitale terreinmodellen worden gegevens van stereoluchtfoto's of satellietbeelden gebruikt. In principe zijn er weinig problemen voor het inwinnen van voldoende informatie op elk gewenst oplossend vermogen - de interpolaties worden op basis van veel gegevens uitgevoerd. In tegenstelling, voor het in kaart brengen van de ruimtelijke verspreiding van de fysische en chemische eigenschappen van bodem, water en lucht kan men alleen met steekproeven werken. Deze werkzaamheden brengen meestal grote kosten met zich mee die vaak een belangrijk beperkende factor is. De interpolaties van deze soorten attributen worden daarom op relatief weinig informatie gebaseerd.

De resultaten van interpolatie kunnen op verschillende manieren in een GIS worden opgeslagen. Als er veel datapunten zijn en het vlak is goed gedefinieerd, zoals bij digitale terreinmodellen, kunnen in vectorsystemen *Triangulated Irregular Networks* (TIN's) worden gebruikt. Vectorsystemen kunnen continuvlakken ook door middel van contouren of isolijnen afbeelden. In rastersystemen worden de geïnterpoleerde attributwaarden per cel uitgerekend, wat de wiskundige koppeling van deze geïnterpoleerde gegevens met andere

gegevens zoals die vanuit de remote sensing enz. vergemakkelijkt.

Als er niet zo veel datapunten zijn dan gebruiken de meeste GI-systemen heel eenvoudige 'inverse distance weighting' methoden of 'spline' functies om attribuutwaarden vanuit puntlocaties tot vlakken te interpoleren. De keuze van interpolatietechnieken is vaak willekeurig en CAD/CAM- overwegingen spelen daarin een belangrijke rol. Zulke technieken zijn snel uit te voeren, gemakkelijk te begrijpen en worden gesteund door het geloof dat continuverschillen in attribuutwaarden altijd glad en vloeiend moeten verlopen. De geschiktheid van beide methoden voor een bepaalde toepassing kunnen echter niet worden onderzocht, en bovendien verschaffen ze geen inzicht in de nauwkeurigheid of mate van onzekerheid of de invloed van ruis en meetfouten op de resultaten.

3 MODELLEREN MET GIS

Voor vele toepassingen van GIS in bodem-, water en milieu-onderzoek is men niet alleen geïnteresseerd in de waarde van een attribuut op een bepaalde lokatie, maar wil het gegevensbestand ook gebruiken als invoer voor eenvoudige of complexe modellen van bepaalde processen, zoals het vaststellen van de bodemgeschiktheid voor één of ander gewas, of het berekenen van potentiële gewasopbrengsten, het schatten van de hoeveelheid aan erosie of oppervlakte-afstroming van een afwateringsgebied of het onderzoeken van grondwaterbeweging en waterkwaliteit. De bewerkingen voor deze modellen kunnen zowel logische operaties (SQL) als wiskundige methoden omvatten (Burrough 1989a). Modellen gebaseerd op logische algebra kunnen worden gebruikt voor het afleiden van geschiktheidsklassen uit meerdere kaartlagen (bijv: ALS de grondsoort klei is EN de grondwatertrap heeft waarde 1, DAN is het gebied ongeschikt voor tarwe). Eenvoudig wiskundige modellen gebruikt men bijvoorbeeld om de waarde van een duur attribuut af te leiden uit een ander goedkoop attribuut. Een voorbeeld van een dergelijk model is het schatten van de hoeveelheid bodemvocht bij het verwelkingspunt (pF 4.2) uit gegevens over de bulkdichtheid en het vochtgehalte bij veldcapaciteit (pF 2.0) (Heuvelink et al. 1989).

De manieren waarmee gegevens uit een GIS worden gekoppeld aan modellen hangt sterk samen met het aard van de GIS-datamodellen. In vectorsystemen worden modelberekeningen voor iedere ruimtelijk eenheid of polygoon apart uitgevoerd. Als met behulp van het model:

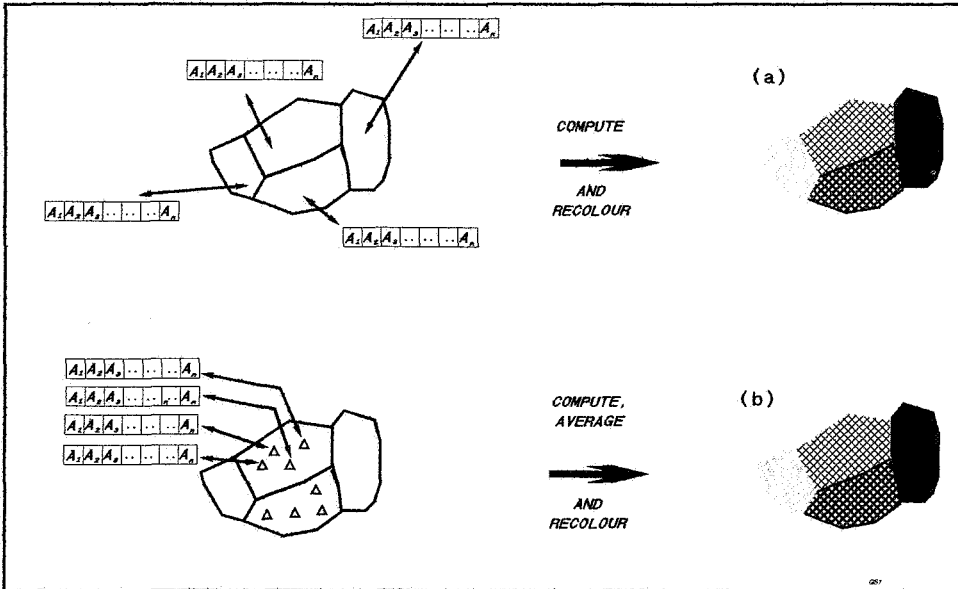
$$U = f(A_1, A_2, \dots, A_n)$$

een nieuwe attribuutwaarde U wordt berekend uit de attributen A_n dan krijgt de betreffende eenheid een nieuw attribuut met bijbehorende nieuwe waarde. Het resultaat wordt zichtbaar gemaakt door middel van een kleur of signatuur (Figuur 1a).

De nauwkeurigheid van het modelresultaat kan verbeterd worden als binnen het polygoon enkele datapunten met dezelfde invoergegevens aanwezig zijn. Door middel van een point-in-polygon zoekmethode kunnen voor iedere eenheid de bijbehorende punten worden gevonden. Het model wordt vervolgens op ieder individueel punt toegepast en de puntgebonden resultaten worden gebruikt voor het berekenen van een gebiedsgemiddelde (Figuur 1b). Hoe meer punten, hoe beter het resultaat. De *ruimtelijke oplossing* van het resultaat blijft echter aan de maat van de ruimtelijke eenheid gebonden.

Dezelfde operaties kunnen ook op de cellen van een raster-GIS uitgevoerd worden, waarbij de omvang van de rastercel een maat is voor het ruimtelijk oplossend vermogen. Als de celwaarden door het verrasteren van een polygoonkaart tot stand zijn gekomen levert deze methode precies hetzelfde resultaat op als wanneer dezelfde berekening wordt uitgevoerd per polygoon. Als het model op elke cel apart is berekend, levert deze methode veel meer rekenwerk op dan met de vector polygoon met geen beter resultaat.

Maar, indien elke cel in het raster-GIS onafhankelijke informatie zou bevatten, en als de omvang van de cel kleiner is dan de polygoon in een vector-GIS, dan kan het runnen van een model in een raster-GIS resultaten met een hogere ruimtelijke oplossend vermogen leveren dan in een vector-GIS. Het is goed denkbaar dat in plaats van het uitrekenen van een *gemiddelde* zoals in Figuur 1b, men de gegevens vanuit de puntwaarnemingen tot een raster zou kunnen *interpoleren* en daarmee het oplossende vermogen van het modelresultaat verhogen. Deze procedure zou misschien aanzienlijke verbeteringen met zich kunnen meebrengen, hetgeen in het onderstaande nader zal worden toegelicht.

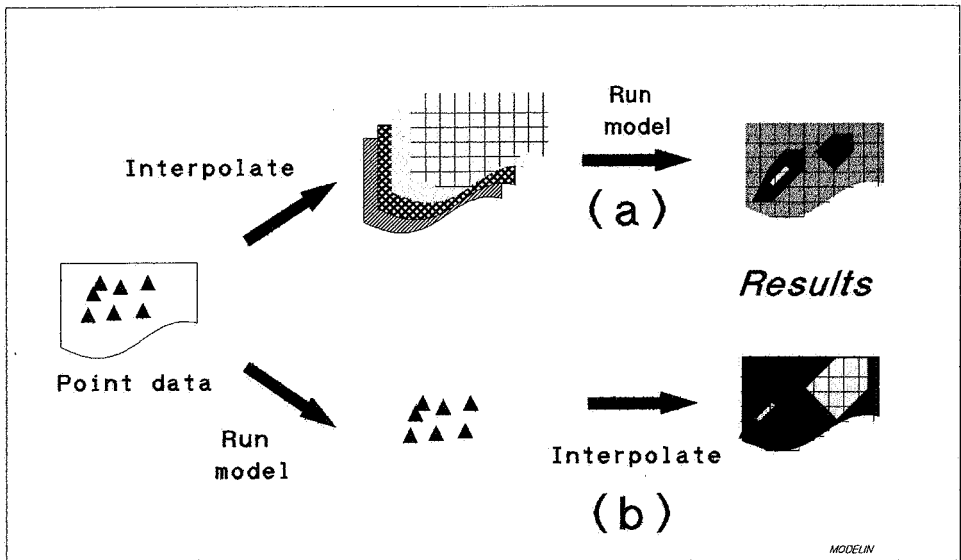


Figuur 1 Het uitvoeren van modellen in een standaard vector GIS.

Bij het runnen van een kwantitatief model in samenhang met uit steekproef verkregen gegevens en een raster GIS zijn er twee procedures mogelijk.

- a) Men interpoleert elk oorspronkelijk attribuut van de puntwaarnemingen tot een grid, en daarna wordt het model op elke gridlocatie uitgevoerd. De resultaten worden daarna als rasterbeeld of als contourlijnen getekend, (Figuur 2a) of
- b) Het model wordt op ieder datapunt uitgevoerd en de resultaten worden daarna tot een raster geïnterpoleerd en afgebeeld (Figuur 2b).

De voordelen van beide ten opzichte van het berekenen van het model op een set gegevens per polygoon is een verhoging van het ruimtelijk oplossend vermogen. Dit kan van groot belang zijn indien er sprake is van een belangrijk laterale variatie binnen de oorspronkelijke polygonen, of het gaat om een dynamisch proces waar de tijdstippen en maten van eindige elementen in evenwicht moeten zijn. De voordelen van procedure a) zijn dat laterale interacties in het model meegenomen kunnen worden en gegevens afkomstig van andere raster databestanden (b.v. satellietbeelden) ook direct als modelinvoer kunnen worden gebruikt.



Figuur 2 Het uitvoeren van modellen met behulp van interpolatie van punten naar raster-vlakken.

De nadelen van deze procedure zijn de grote rekestijden die bij complexe modellen voorkomen. De voordelen van procedure b) zijn dat resultaten van zeer complexe modellen op een beperkte aantal punten berekend kunnen worden, wat rekestijd spaart, maar laterale interacties kunnen niet zo eenvoudig worden meegenomen.

4 DE KWALITEIT VAN MODELRESULTATEN

Modelresultaten worden slechts bruikbaar als ze aan geaccepteerde kwaliteitseisen voldoen. In de wereld van de GI-systemen is het begrip 'kwaliteit' moeilijk te omschrijven. Tot voor kort gebruikte men het woord alleen in verband met de kartografische kwaliteit van de uitvoer. Met het toenemend gebruik van GIS voor modelleringsdoeleinden heeft het woord echter een nieuwe betekenis gekregen. Men begint te beseffen dat de kwaliteit van modelresultaten sterk wisselend is en samenhangt met een complex van factoren. Maten

voor kwaliteit zijn noodzakelijk, zoals het hanteren van betrouwbaarheidsmarges, wil men het beleid met behulp van modelresultaten goed kunnen onderbouwen.

Uit het voorgaande is duidelijk geworden dat er een aantal tegenstrijdige aspecten kleven aan het opslaan en verwerken van gegevens met de huidige GI-systemen. In vectorsystemen kan de ligging van objecten heel nauwkeurig worden vastgelegd maar is weinig informatie beschikbaar over ruimtelijke variaties binnen de eenheid. Bovendien, veel "objecten" in bodem en water zijn niet als exact te beschouwen (Burrough 1989b). In rastersystemen is het moeilijk de exacte ligging van een object vast te stellen, maar het geleidelijk veranderende verloop van attribuutwaarden kan makkelijker worden beschreven. Veel moderne GIS hebben dit probleem onderkend en bieden een combinatie van raster en vector functionaliteit maar, ze hebben geen mogelijkheden om nog de fouten in de oorspronkelijke attribuutwaarden, nog de voortplanting van fouten in modellen te kunnen behandelen of te kunnen minimaliseren. De huidige GIS hebben ook geen mogelijkheden voor het formaliseren van kennis over hoe bepaalde analyseprocedures het beste kunnen worden uitgevoerd, wat betekent dat de "rule base" voor analyse buiten het systeem blijft.

Kennis over de kwaliteit van modelresultaten ontbreekt vaak omdat uit budgettaire overwegingen men geen waarneming meer wil maken dan minimaal noodzakelijk is. Dit is kortzichtig, want verschillende modellen kunnen resultaten van diverse kwaliteit opleveren, en het is niet duidelijk dat de meest gedetailleerde modellen altijd de beste resultaten voor een bepaalde hoeveelheid aan inspanning leveren. Er is een duidelijk behoefte inzicht te verkrijgen in hoe men, gebruikmakend van een GIS en de bijbehorende modellen, uit de beschikbare gegevensbestanden de beste resultaten kan afleiden bij een minimum aan uitgaven. De geostatistiek kan in dit kader een belangrijke rol spelen bij het creëren van een intelligent GIS.

5 WAT ONTBREEKT BIJ DE HUIDIGE GIS?

Uit het voorgaande kan geconcludeerd worden dat bij het huidige GIS ontbreken:

1. Een Datamodel dat rekening houdt met niet-exacte gegevens;
2. Methoden voor optimale interpolatie;
3. Methoden voor het optimaliseren van meetnetten;

4. Methoden voor het verbeteren van de precisie van de uitkomsten van modellen;
5. Methoden voor het doorrekenen van fouten;
6. Methoden voor het opslaan en beschikbaar stellen van kennis over hoe GIS-analyses het beste uitgevoerd kunnen worden.

De ruimtelijke statistiek kan hiermee behulpzaam zijn.

6 DE RUIMTELIJKE STATISTIEK ALS COMPONENT VAN EEN INTELLIGENT GIS

De ruimtelijke statistiek kan een aantal van de boven genoemde tekortkomingen verhelpen. Door gebruik van een stochastisch model in plaats van exacte datamodellen houdt de ruimtelijke statistiek rekening met niet-exacte gegevens, zowel in de verspreiding van een fenomeen als in de attribuutwaarden. Daarnaast door het bepalen van covariantiestructuren kan de ruimtelijke statistiek bruikbare informatie leveren ten behoeve van optimale interpolatie, meetnet optimalisatie en het doorrekenen van fouten in modellen.

6.1 Interpolatiemethoden

Er zijn een groot aantal geostatistische interpolatiemethoden, waaronder simple and ordinary kriging, univariaat en multivariaat kriging (co-kriging), universal kriging, disjunctie kriging en probability kriging, Kalman filtering, enz. die kunnen in of naast een GIS gebruikt worden. De voorspellingen kunnen betrekking hebben op punten (point kriging) of op raster cellen of blokken (block kriging), waarbij in enkele gevallen bovendien de mogelijkheid bestaat het onderzoeksgebied vooraf in te delen in een aantal min of meer homogene strata (Stein et al., 1989). Alle technieken leveren voorspellingen van de waarde van een attribuut en schattingen van de voorspelfout op niet bezochte plaatsen.

Geostatistische interpolatietechnieken kunnen zonder meer direct aan bestaande raster-GIS gekoppeld worden. Invoer voor de geostatistiek is afkomstig van de in het GIS aanwezige puntgegevens. Het berekenen van covariantiestructuren en het fitten van variogrammodellen kunnen makkelijk op personal computers gedaan worden (Englund and Sparks 1988, Yates 1990) en grootschalige interpolatie (kriging) kan op UNIX-werkstations of andere computers

uitgevoerd worden. De resultaten kunnen echter als isolijnen of als rasters met behulp van standaard GIS opties getekend worden.

6.2 Covariantiestructuren

Veel methoden in de ruimtelijke statistiek maken gebruik van analyse van de ruimtelijke- en tijdscovariantie functies van attributen. Dit kan belangrijk informatie opleveren zoals:

- a) de juiste gewichten voor een optimale voorspelling met een bekende schattingsvariantie (kriging enz.);
- b) informatie over de beste ruimtelijke- en tijdschalen voor het bemonsteren van de verschillende attributen - dit kan erg belangrijk zijn als deze gegevens later in modellen worden gebruikt (zie Burrough 1991);
- c) specificaties voor het optimaliseren van meetnetten waarbij de verkregen resultaten aan bepaalde foutenmarges moeten voldoen. (van Geer 1987, McBratney et al 1981).

ad a) Als de covariantiestructuur niet verschilt van die van witte ruis dan is het duidelijk dat met het gebruikte bemonsteringsstrategie interpolatie niet zinvol is. Geen andere interpolatietechniek geeft zulke resultaten.

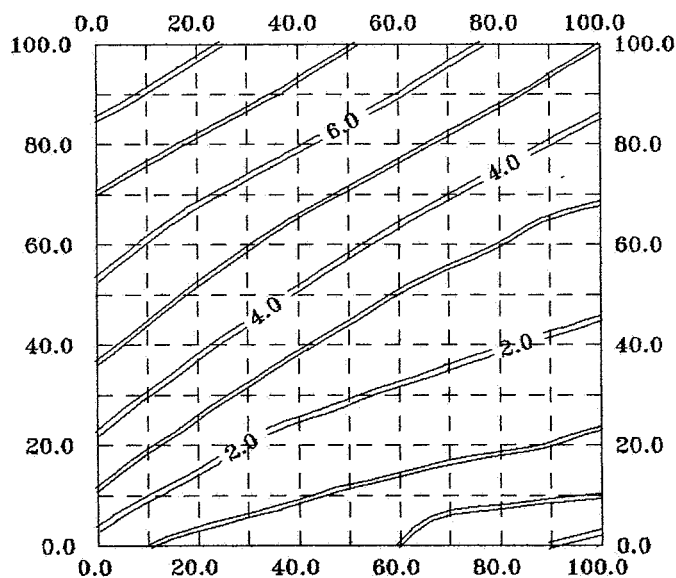
ad b) Het runnen van modellen met diverse invoergegevens kan onzinnige resultaten opleveren als de covariantiestructuren van de verschillende attributen sterk van elkaar verschillen. Alle gegevens voor een model kunnen beter op dezelfde "support" gemeten worden.

ad c) De "kriging error" ofwel de fout bij interpolatie is een functie van de interpolatietechniek (zie onder), het aantal datapunten, de configuratie van de datapunten en het covariantiestructuur (b.v. het variogram). Als het variogram bekend is, en het data configuratie ook is vastgesteld kan men de schattingsfout berekenen als functie van (i) het afstand tussen de datapunten op een regelmatig grid en (ii) de maat van het grid cel of blok (Figuur 3). Vanuit zulke analyses is het mogelijk direct af te kunnen lezen welke combinaties van cel grote en bemonsteringsafstand resultaten met een gegeven schattingsfout zou kunnen opleveren.

6.3 GIS, modellen en de voortplanting van fouten.

De methoden van de ruimtelijke statistiek kunnen modelresultaten verbeteren en tevens informatie over de kwaliteit van de resultaten leveren. Dit kan variëren van onafhankelijke toetsing door gebruikmaking van testgegevens tot jack-knifing van de schattingsfouten verkregen door methoden als kriging en Kalman filtering waar schattingsfouten worden geleverd door de techniek.

Het hebben van schattingsfouten per attribuut en per locatie kan het analyseren van de voortplanting van fouten in modellen aanzienlijk verhelpen, vooral als het model wordt uitgevoerd op gegevens in een raster GIS.



Figuur 3 Een voorbeeld van de afhankelijkheid van de schattingsfouten bij blokkriking afhankelijk van Blokgröte (x-as) en meetnetafstand (y-as).

Als een model van de generische vorm:

$$U = f(A_1, A_2, \dots, A_n)$$

op elke gridcel in een raster GIS wordt gebruikt, dan is de kwaliteit van het antwoord afhankelijk van drie aspecten:

- het model zelf: hoe goed beschrijft het model de werkelijkheid op de gekozen ruimtelijke schaal en/of tijdschaal?
- de nauwkeurigheid waarmee het model wordt gecalibreerd (instelling van het model op een gekozen gebied);
- de ruimtelijke/temporele variabiliteit van de inputgegevens.

Een intelligente GIS zou de gebruiker in staat moeten stellen procedures te kiezen om de fouten in de uitvoer zo klein mogelijk te houden. Als de resultaten veel nauwkeuriger zijn dan nodig, dan kunnen mogelijk besparingen worden bereikt bij de inrichting van een meetnet. Als de resultaten te grof zijn, dan zullen vragen moeten worden beantwoord ten aanzien van verbeteringen van de invoer en/of het model, en de bijbehorende kosten.

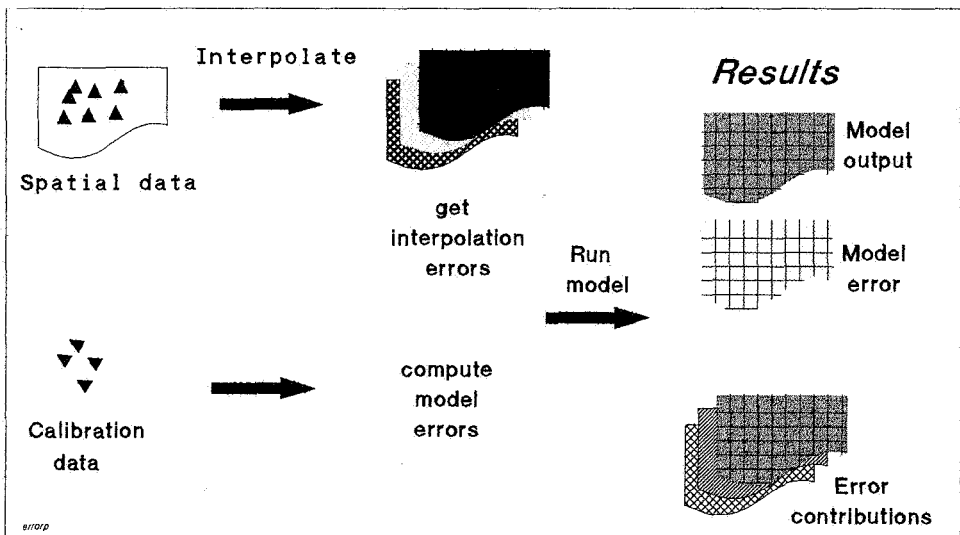
De gevoeligheid van modellen voor fouten in de invoer kan met behulp van Monte Carlo methoden worden onderzocht, maar vergen voor complexe modellen enorme rekentijden (Hazelhoff et al. 1990). Alternatieve methoden zijn het gebruik van een Taylorreeks benadering en Rosenblauth's methode. Heuvelink et al (1989) zijn erin geslaagd om voor *regressie-achtige modellen* van het type

$$U = \alpha_1 A_1, \alpha_2 A_2, \dots, \alpha_n A_n$$

die voor gehele grid overlays worden gebruikt, het modelfout δU en tevens de contributies van elk $\delta \alpha_i$ en δA_n afzonderlijk door te rekenen (Figuur 4). Dit resultaat levert inzicht in welke van de inputs en welke parameters de grootste onzekerheid oplevert. Als de α 's de grootste bijdrage leveren, dan is het raadzaam het model beter te calibreren; als ruimtelijke verschillen een of meer van de grootste bijdrage leveren, dan is het raadzaam te kijken of deze nog nauwkeuriger geïnterpoleerd zouden moeten of kunnen worden en eventueel andere wegen tot verbetering van de resultaten te onderzoeken. Interpolatie binnen exacte eenheden verhoogt het oplossend vermogen en tevens verlaagt het de absolute fouten in de eindige elementen (de Roo et al 1989).

Met behulp van de ruimtelijke statistiek kunnen de volgende vijf alternatieven verbeteringen in modelresultaten opleveren:

- gebruik een interpolatiemethode voor de schatting i.p.v. van een gemiddelde waarde per ruimtelijke eenheid, of kies een betere interpolatiemethode (bijvoorbeeld co-kriging in plaats van univariaatkriging);



Figuur 4 De voortplanting van fouten in regressiemodellen toegepast op GIS-databestanden.

- verzamel meer gegevens (optimalisering c.q. aanpassing van meetnet);
- verzamel andere gegevens (veel goedkope gegevens in plaats van weinig dure gegevens, bijv. relatieve hoogtes in plaats van gehalten aan zware metalen - zie Leenaers et al 1989);
- gebruik andere of betere modellen, die ook andere soorten gegevens nodig hebben;
- verbeter de modelcalibratie (vraagt meer detailwaarnemingen om model echt geschikt te maken voor het gebied).

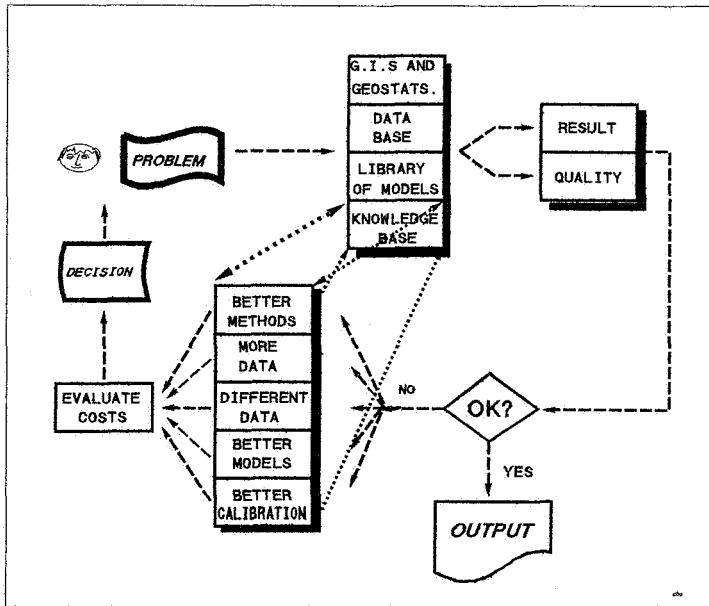
Elk van deze keuze brengt bijbehorende kosten met zich mee. Een kosten/baten-analyse zou de gebruiker in staat moeten stellen om een rationale beslissing te kunnen nemen (Figuur 5).

7 KENNISSYSTEMEN

Er is veel vraag om met behulp van een GIS ruimtelijke processen te kunnen modelleren. Het is duidelijk dat de methoden van de ruimtelijke statistiek belangrijke bijdragen aan het verbeteren van modelresultaten kunnen leveren, waaronder ook kennis over foutenmarges en de voortplanting van fouten. Wij zijn in een stadium aangekomen waar het noodzakelijk is de bestaande kennis en ervaring over de bijdrage van de ruimtelijke statistiek te bundelen en te formaliseren. Deze geformaliseerde kennis kan in een kennissysteem naast het GIS gezet worden om gebruikers te assisteren bij het kiezen van data-analyse-methoden en modellen om resultaten met een *van te voren bepaalde kwaliteit* te kunnen realiseren.

Een voorbeeld van zo'n regel zou kunnen zijn: "Als men het ruimtelijk patroon van bodemverontreiniging in een uiterwaard wil modelleren, dan is het raadzaam de verontreinigingsgegevens aan te vullen met hoogtegegevens en een interpolatie uit te voeren met co-kriging". Dit soort praktische kennis en informatie over het soort variogram dat in een bepaald geval mag worden verwacht zou moeten worden vastgelegd in een kennissysteem. De waarde en de 'intelligentie' van een GIS worden daarmee aanzienlijk verhoogd (Figuur 5).

Figuur 5 is een zeer eenvoudig schema voor een intelligent GIS. De gebruiker vraagt een modelresultaat met een bepaalde betrouwbaarheid. Op basis van bestaande inzichten en regels wordt een verwerkingsprocedure op gang gezet. De fouten in de uitvoer worden berekend als functie van de modelparameters en de modelinvoer. Het resultaat wordt vergeleken met de specificaties. Zijn de resultaten voldoende nauwkeurig, dan kan uitvoer worden gegenereerd. Zijn de resultaten niet nauwkeurig genoeg, dan moet de gebruiker een beslissing nemen. Hij concludeert dat zijn eisen te streng waren en accepteert de resultaten, of hij kiest voor een verbetering van het model of de invoer na afweging van de bijkomende kosten. De kennis en regels die door dit proces worden opgeleverd, kunnen later worden



Figuur 5 Een voorbeeld van een intelligente GIS met geostatistiek.

ingebracht in het kennissysteem, hetgeen leidt tot een algemene verbetering van het systeem.

Een intelligent GIS zou niet alleen moeten beschikken over technieken voor het optimaal schatten van attribuutwaarden op onbemonsterde plaatsen, maar ook informatie over hoe de methoden te werk gaan en over hun voor- en nadelen. Met betrekking tot de modellen moeten niet alleen de verwerkingsmodule, maar ook informatie over de schaal van toepassing, de soorten gegevens en hun vereiste nauwkeurigheid aanwezig zijn. Daarnaast dient er een reeks methoden beschikbaar te zijn voor het doorrekenen van fouten, voor meetnetoptimalisatie en voor het uitvoeren van kosten/baten-analyses. Dit geheel zou niet alleen een krachtig gereedschap kunnen zijn dat onze kennis zeer effectief benut, maar ook door het gebruik onze kennis kunnen vergroten. Als men weet dat er geen antwoord is op de vraag "wat is de kwaliteit van het resultaat?", dan worden zulke vragen niet gesteld -

gebruikers hebben immers meestal geen besef over de relaties tussen vraagstelling, verwachte kwaliteitseisen en praktische mogelijkheden. Er worden inzichten verkregen in de zwakste aspecten van modellen, hetgeen tot modelverbetering kan leiden. Dure monitoring programma's zouden goed doorgelicht kunnen worden om te kijken of de met veel kosten verzamelde gegevens wel allemaal nodig waren. Het grootste wetenschappelijke voordeel is misschien wel dat de hele procedure van gegevensverzameling tot en met modelverwerking wordt geformaliseerd. Dit kan leiden tot het opstellen van regels met algemene geldigheid.

Wij zijn in zekere zin niet meer zo ver weg van dit ideale GIS. Het is nu zonder meer mogelijk geostatistische methoden aan GIS-databestanden te koppelen en gegevens te kunnen interpoleren. Er zijn methoden voor het optimaliseren van meetnetten, zowel voor ruimtelijk- als voor tijdsgebonden gegevens, en er zijn ook methoden die onder bepaalde omstandigheden het doorrekenen van fouten aan kunnen. Wat nog ontbreekt zijn geformaliseerde kennis over het fitten van variogrammen, kennis over het kiezen van de juiste interpolatietechniek, en algemeen geldende methoden voor het schatten van de fouten. Het is aantrekkelijk te denken dat krigingfouten absoluut zijn. Ze zijn echter afhankelijk van veel factoren, zoals het experimentele variogram, het gewitte model, de kriging methode, het verspreidingspatroon van de datapunten, het aantal datapunten dat wordt gebruikt, de blokmaat en de support size. Over deze aspecten, en nog meer, is er gelukkig nog meer onderzoek nodig.

DANKWOORD

Mijn dank aan Dr H. Leenaers voor het leveren van kritisch commentaar en voor taalcorrecties.

LITERATUUR

- BURROUGH, P.A.; 1986. Principles of Geographical Information Systems for Land Resource Survey. Oxford University Press.
- BURROUGH, P.A.; 1989a. Matching spatial databases and quantitative models in land resource assessment. Soil Use and Management 5: 3-8.
- BURROUGH, P.A.; 1989b. Fuzzy mathematical methods for soil survey and land evaluation. J. Soil Sci. 40: 477-492, 1989.

- BURROUGH, P.A.; 1991. Sampling designs for Quantifying Map Unit Composition. In: L. Wilding and M. Mausbach (Eds). "Spatial variability and Map Units for Soil Surveys", International Soil Science Society Working Group of Soil and Moisture Variability in Time and Space /American Society of Agronomy, the Crop Science Society of America and the Soil Science Society of America. (in press).
- DE ROO, A.P.J.; HAZELHOFF, L. en BURROUGH, P.A., 1989. Soil Erosion Modelling using ANSWERS and Geographical Information Systems. *Earth Surface Processes and Landforms* 14: 517-532.
- ENGLUND, E. en SPARKS, A.; 1988. GEO-EAS User's Guide. Environmental Monitoring Systems Laboratory, Office of Research and Development, Environmental Protection Agency, Las Vegas, Nevada.
- HAZELHOFF, L., DE ROO, A.P.J. en HEUVELINK, G.B.M.; 1990. The use of Monte Carlo Simulations to Estimate the Effects of Spatial Variability of Infiltration on the Output of a Distributed Runoff and Erosion Model. *Proc. EGIS'90, First European Conference on Geographical Information Systems*, 442-452.
- HEUVELINK, G.B.M., BURROUGH, P.A. en STEIN, A.; 1989. Propagation of error in spatial modelling with GIS. *Int. J. Geographical Information Systems* 3: 303-322.
- LEENAERS, H., Okx,J.P E en Burrough, P.A.; 1989. Co-kriging: an accurate and inexpensive means of mapping floodplain soil pollution by using elevation data. In: *Geostatistics*, edited by M. Armstrong, Proceedings of the third Geostatistics Congress, Avignon, October 1988, Kluwer, 371-382.
- MCBRATNEY, A.B. en WEBSTER, R.; 1981. The design of optimal sampling schemes for local estimation and mapping of regionalized variables: 2 Program and examples. *Computers & Geosciences* 7: 335-65.
- STEIN, A., HOOGERWERF, M. en BOUMA, J.; 1988. Use of soil map delineations to improve (co)kriging of point data on moisture deficits. *Geoderma* 43 163-177.
- VAN GEER, F.C.; 1987. Applications of Kalman filtering in the analysis and design of groundwater monitoring networks. Institute of Applied Geoscience, TNO. Delft, Report PN 87-05, 130pp.
- WEBSTER, R. en OLIVER, M.A.; 1990. *Statistical Methods in Soil and Land Resource Survey*. Spatial Information Systems Series No. 1, Oxford University Press.
- YATES, S.; 1990. GEOPAK - Geostatistical programs. University of Santa Barbara, California?

COMMISSIE VOOR HYDROLOGISCH ONDERZOEK TNO

RAPPORTEN EN NOTA'S

- No. 1. Tweede rapport en aanbevelingen van de Contactgroep Archivering en Automatische Verwerking van hydrologische gegevens TNO.
Januari 1977.
- No. 2. Verslag en aanbevelingen van de ad hoc-Groep Grondwatermodellen en Computerprogrammatuur TNO.
Juli 1978.
- No. 3. De droogte in 1976.
Een samenvatting en overzicht van de over de droogte van 1976 verschenen literatuur - P.K.M. v.d. Heijde.
Augustus 1978.
- No. 4. Nederlandse activiteiten in internationaal hydrologisch verband.
Lezingserie, gehouden op 25 april 1978 te Delft, aangevuld met (schematische) overzichten van internationale organisaties en een overzicht van hun vertegenwoordigers in Nederland.
Augustus 1978.
- No. 5. Waterkwaliteit in grondwaterstromingsstelsels.
Verslag van de Workshop op 1 en 2 april 1980 te Wageningen - (red. J.C. Hooghart), aangevuld met discussiebijdragen en een inventarisatie van het onderzoek in Nederland.
Augustus 1980.
- No. 6. Derde rapport en aanbevelingen van de Contactgroep Archivering en Automatische verwerking van hydrologische gegevens TNO.
Februari 1981.
- No. 7. Overzicht van de wensen van hydrologen en waterbeheerders ten aanzien van het operationele regenwaarnemingsnet van het KNMI - J.C. Hooghart.
Oktober 1981.
- No. 8.*) Verklarende Hydrologische Woordenlijst van de Gespreksgroep Hydrologische Terminologie.
- 8a. I. Water in de onverzadigde zone
 - II. Water in de verzadigde zone
Januari 1982.
 - 8b. III. Atmosferisch water
Juni 1983.
 - 8c. IV. Oppervlaktewater
Maart 1985.

*) Verouderd: vervangen door Rapporten + Nota's no. 16.

- No. 9. Waterkwaliteit en waterkwantiteit in het IJsselmeergebied.
Verslag van de 2e CHO-studiebijeenkomst op 2 en 3 november 1981, De Eemhof, Zuidelijk Flevoland - (red. J.C. Hooghart), aangevuld met discussiebijdragen.
Februari 1982.
- No. 10. Rapport en aanbevelingen
van de Contactgroep Grondwatermodellen, CHO-TNO.
April 1982.
- No. 11. Inventarisatie Grondwaterkwaliteitsmodellen.
L.J.M. Boumans.
Oktober 1982.
- No. 12. Grondwaterkwaliteit in relatie met onderzoek en beleid.
Verslag van de 3e CHO-studiebijeenkomst op 15 maart 1983 te Wageningen - (red. J.C. Hooghart), aangevuld met discussiebijdragen.
Juni 1983.
- No. 12a. Voorlopig overzicht van inventarisaties waarin grondwater(kwaliteits)modellen voorkomen of hiermee in verband staan.
J.C. Hooghart.
Januari 1984.
- No. 13. Vergelijking van modellen voor het onverzadigd grondwatersysteem en de verdamping.
Verslag van de 4e CHO-studiebijeenkomst op 24 oktober 1984, georganiseerd in samenwerking met de Studiegroep Hupselse Beek - (red. J.C. Hooghart).
Maart 1985.
- No. 14. Meten, meetnetten en optimale meetnetontwerpen ten dienste van het waterbeheer.
Verslag van:
- Voorjaarsbijeenkomst van de KIVI Sectie Waterbeheer:
"Meten voor waterbeheer", mei 1984.
- Colloquium van de Studiegroep Statistiek in de hydrologie CHO-TNO:
"Meetontwerp en optimalisatie", november 1984.
(red. P. v.d. Kloet en J.C. Hooghart).
Januari 1986.
- No. 15. Het hydrologisch systeem in het grensgebied Luik-Maasbracht.
Le système hydrologique dans la région frontalière Liège- Maasbracht.
Verslag van de 5e CHO-studiebijeenkomst op 13 december 1985, georganiseerd in samenwerking met de Nationale IHP-comité's van België en Nederland en de Contactgroep Hydrologie van het Nationaal Fonds voor Wetenschappelijk Onderzoek uit België.
(red. J.C. Hooghart).
April 1986.

- No. 16. Verklarende Hydrologische Woordenlijst van de Gespreksgroep Hydrologische Terminologie, waarin opgenomen de hoofdstukken:
 I Algemene termen
 II Atmosferisch Water
 III Water in de onverzadigde zone
 IV Water in de verzadigde zone
 V Oppervlaktewater
 Oktober 1986, hernieuwde uitgave.
- No. 17.*) Duurzaamheid rioolleidingen; een literatuurstudie naar aantastingsmechanismen.
 R.B. Polder.
 Februari 1987.
 *) Uitverkocht.
- No. 18. Ruimtelijke variabiliteit van bodem en water.
 Verslag van de 6e CHO-studiebijeenkomst op 22 oktober 1986.
 (red. J.C. Hooghart).
 Februari 1987.
- No. 19. Van Penman naar Makkink; een nieuwe berekeningswijze voor de klimatologische verdampingsgetallen.
 Eindrapport van de KNMI-Projectgroep en de CHO-Begeleidingsgroep Verdampingsberekening.
 (red. J.C. Hooghart en W.N. Lablans).
 December 1988.
- No. 20. Tijdreeksen in bodem en water.
 Inleidingen van de lezingendag op 25 januari 1989 van de NRLO-Werkgroep Ruimtelijke variabiliteit in bodem en water en de Studiegroep Statistiek in de Hydrologie van de CHO-TNO.
 December 1988.
- No. 21. Neerslagmeting en -voorspelling; toepassing van modern technieken, zoals radar- en satellietwaarnemingen.
 Verslag van de 7e CHO-studiebijeenkomst, georganiseerd in samenwerking met SAMWAT, op 16 november 1988.
 (red. J.C. Hooghart).
 Februari 1989.
- No. 22. Integraal Waterbeheer in het Goois/Utrechts stuwwallen- en plessengebied.
 Verslag van de op 7 april 1989 in Bussum gehouden themadag, georganiseerd door het Zuiveringschap Amstel en Gooiland en de Provincie Utrecht, in samenwerking met de CHO-TNO.
 (red. L. van Liere, R.M.M. Roijackers en P.J.T. Verstraelen).
 Augustus 1989.

- No. 23. Bodemwaterkwaliteit in wisselwerking met biologische, chemische en hydrologische processen.
Verslag van de 8e CHO-studiebijeenkomst op 8 mei 1990.
(red. J.C. Hooghart)
September 1990.
- No. 24. Ruimtelijke statistiek van bodem en water.
Inleidingen van de lezingendag op 24 januari 1991 van de NRLO-werkgroep Ruimtelijke variabiliteit van bodem en water en de Studiegroep Statistiek in de Hydrologie van de CHO-TNO.
(red. J.C. Hooghart)
Januari 1991.

Voor bestellingen en informatie: CHO-TNO
Postbus 6067
2600 JA DELFT

