

BASSON E M

HANTERING VAN TYDREEKSE MET VERLORE WAARNEMINGS  
MET BEHULP VAN DIE TOESTANDRUIMTE BENADERING EN  
DIE KALMANFILTER

Mwiskundige Statistiek

UP

1991

HANTERING VAN TYDREEKSE MET VERLORE WAARNEMINGS MET  
BEHULP VAN DIE TOESTANDRUIMTE BENADERING EN DIE  
KALMANFILTER

deur  
ELIZABETH M BASSON

Voorgele ter vervulling van 'n deel van die vereistes vir die graad  
Magister in Wiskundige Statistiek in die Fakulteit Natuurwetenskappe  
Universiteit van Pretoria  
Pretoria.

Julie 1991

## Voorwoord

Hiermee wil ek graag dr J.S.Galpin bedank vir haar ondersteuning gedurende die navorsing wat tot hierdie verhandeling gelei het, asook professor S.H.C. du Toit as studieleier.

## INHOUDSOPGAWE

Voorwoord

Hoofstuk 1 – Inleiding	1.1
Hoofstuk 2 – Toestandruimte modelle	2.1
2.1 Inleiding	2.1
2.2 Toestandruimte modelle	2.2
2.3 Die Kalmanfilter	2.5
2.4 Die voorspellings–fout–ontbinding van die aanneemlikheidsfunksie	2.6
2.5 Die toestandruimte voorstelling van 'n outoregressiewe– bewegendegemiddelde proses	2.9
2.6 Opsomming	2.15
Hoofstuk 3 – Die Kalmanfilter vergelykings	3.1
3.1 Inleiding	3.1
3.2 Basiese resultate	3.2
3.3 Die Kalmanfiltervergelykings	3.5
3.4 Algemene vorm van die Kalmanvergelykings	3.12
3.5 Aanvangswaardes vir die Kalmanfilter	3.13
3.6 Illustrasie van die verskillende tegnieke aan die hand van 'n MA(1)–proses	3.21
3.7 Samevatting	3.26

<b>Hoofstuk 4 – Hantering van tydreeks met verlore data met behulp</b>	
<b>van die toestandruimte benadering en die Kalmanfilter</b>	<b>4.1</b>
4.1 Inleiding	4.1
4.2 Verlore data en die Kalmanfilter	4.2
4.3 Die algoritme van Gardner, Harvey en Phillips(1980)	4.3
4.4 Samevatting	4.18
<b>Hoofstuk 5 – 'n Kort oorsig oor ander metodes wat verlore data</b>	
<b>hanteer</b>	<b>5.1</b>
5.1 Inleiding	5.1
5.2 'n Kort oorsig oor verskeie tegnieke om verlore en ongelyk	
" gespasieerde data te hanteer	5.2
5.2.1 Metodes wat die tydreeks eers verwerk	5.2
5.2.2 Metodes wat verlore data beraam en invul	5.4
5.3 Vergelyking van metodes	5.9
5.4 Opsomming	5.20
5.5 Bylae	5.21
<b>Hoofstuk 6 – Gevolgtrekkings en voorstelle vir verdere studie</b>	<b>6.1</b>
<b>Samevatting</b>	<b>7.1</b>
<b>Summary</b>	<b>8.1</b>
<b>Literatuurverwysings</b>	<b>9.1</b>

# HOOFSTUK 1

## INLEIDING

Vir navorsers, in byvoorbeeld omgewingsake, is dit belangrik om data wat oor 'n periode ingewin is, te analiseer. Die doel van so 'n analise kan byvoorbeeld wees om 'n tendens te bepaal en die tendens kan gebruik word om toekomstige gedrag te voorspel.

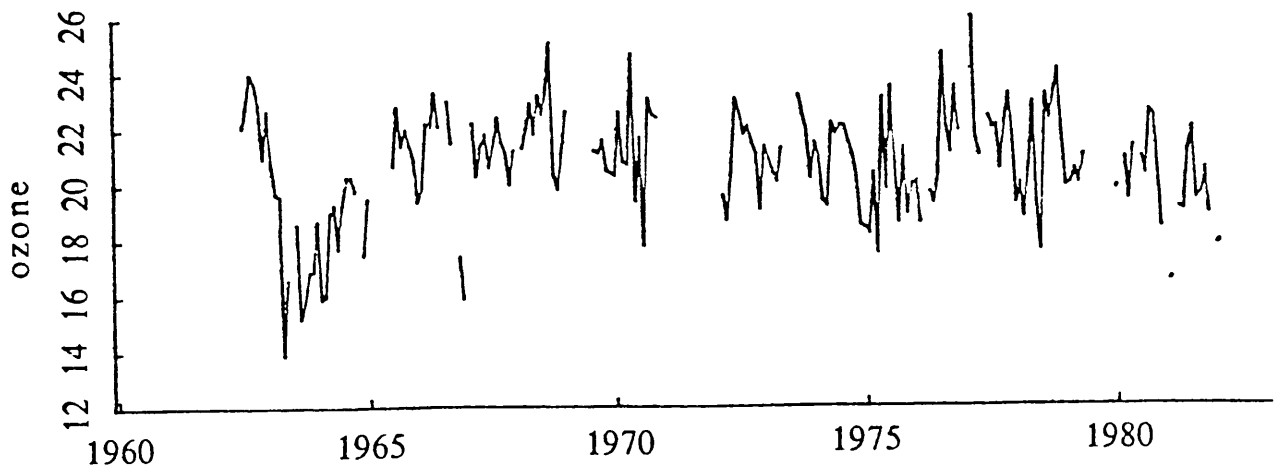
Aangesien die beskikbare data oor tyd gekorreleerd is, is die aangewese metode 'n tydreeksanalise. Die standaard tegnieke, en die meeste rekenaarpakette, vereis dat die waarnemings volledig en op opeenvolgende, gelykgespasiëerde tydsintervalle, beskikbaar is.

Die waarnemings vir 'n tydreeks word gewoonlik deur 'n persoon met 'n meetapparaat in naturomstandighede gemaak. Die faktore mens, instrument en natuur dra dan ook by tot 'n groot variasie in die hoeveelheid en kwaliteit van die langtermyn datastelle

Figuur 1 is 'n grafiek van stratosferiese osoonda(Wincek en Reinsel(1986)) en is 'n voorbeeld van 'n datastel waarvan die omvang van die vermiste data tot 20% beloop.

Figuur 1

### STRATOSFERIESE OSOONDATA OOR DIE TYDPERK 1963–1981



Die doel van hierdie studie is om metodes te bestudeer wat onvolledige datastelle hanteer, met spesiale verwysing na die metode waar van die Kalmanfilter gebruik gemaak word.

Om van die Kalmanfilter gebruik te maak, moet 'n outoregressief-bewegende-gemiddelde-model (ARMA-model) eers in 'n toestandruimte voorstelling geskryf word. Aangesien op verskeie plekke in die literatuur slegs die finale vorm van die toestandruimte voorstelling van 'n ARMA-model gegee word, word in Hoofstuk 2 'n weergawe van die afleiding van die proses gegee. In hierdie hoofstuk word die voorspellings-fout-ontbinding van die aanneemlikheidsfunksie ook uiteengesit as hulpresultaat wat sy toepassing in Hoofstuk 4 vind.

Die Kalmanfilter vergelykings word in Hoofstuk 3 bespreek en die teorie word in Hoofstuk 4 toegepas om tydreeks met verlore data te hanteer.

In Hoofstuk 5 word 'n kort oorsig oor ander metodes wat verlore data hanteer gegee en vier van hierdie metodes word gebruik om hulle te vergelyk teen die Kalmanfilter metode.

## HOOFSTUK 2

### TOESTANDRUIMTE MODELLE ( *state space models* )

#### 2.1 INLEIDING

In die praktyk is daar 'n groot aantal prosesse waarby die toestand van die proses op tydstip  $t$  'n funksie is van sy toestand op tydstip  $t-1$ . Daar het dan ook 'n studieveld ontwikkel om die gedrag van die proses op 'n tydstip te kan voorspel gebaseer op die inligting van sy gedrag op die vorige tydstip, en word daar gepraat van toestandruimte modelle.

Die tegnologie is aanvanklik hoofsaaklik deur ingenieurs gebruik. In die literatuur oor toestandruimte modelle is 'n tipiese sinsnede dan ook: "State space models were originally developed by control engineers". In 'n algemene toepassing word daar gekyk na 'n stel van toestandsveranderlikes wat oor tyd verander. Die toestandsveranderlikes kan byvoorbeeld 'n sein wees wat die posisie van 'n missiel aandui. Die "toestand" van die missiel is dan sy posisie op 'n sekere tydstip  $t$ , wat weer 'n funksie is van sy posisie op tydstip  $(t-1)$ . Die posisie word gemodelleer deur 'n eerste-orde-differensievergelyking. Daarenteen word sy posisie werklik waargeneem deur 'n lesing, gemeet byvoorbeeld deur 'n satelliet. Die lesing is natuurlik vervorm deur seinsteurings en sogenaamde witruis ( *white noise* ). In Afdeling 2.2 word die toestandruimte model wiskundig geformuleer in terme van die oorgangs- (in bogenoemde voorbeeld die differensievergelyking) en metingsvergelykings (in voorbeeld die satellietlesing) uiteengesit.

Die toestandruimte modelle gee aanleiding tot 'n stel vergelykings wat 'n beraming van 'n bepaalde "toestand" maak wat op 'n spesifieke stadium waargeneem word. Die probleem om die beramings, onder gegewe stel aannames, rekursief te doen, was oorspronklik opgelos deur Kalman wat dan ook aanleiding gegee het tot die bekende Kalmanfilters. In Afdeling 2.3 word die Kalmanfilters verder gedefinieer en in Hoofstuk 3 meer breedvoerig uiteengesit. Deur die gebruik van die Kalmanfilter word 'n reeks foutterme geproduseer. Die foutterme word in die voorspellings-fout-ontbinding van die aanneemlikheidsfunksie geïnkorporeer. In Afdeling 2.4 word die voorspellings-fout-ontbinding van die aanneemlikheidsfunksie uiteengesit en in Hoofstuk 3 verder daarna verwys.

Alhoewel die oorsprong van die toestandruimte modelle by die ingenieurswese te vinde is, het die toepassingsmoontlikhede wyer gekring na die ekonomie (Harvey en Pierce (1984), Harrison en Stevens (1976), Shumway en Stoffer (1982)), die medisyne en die grondwetenskappe (Shumway (1985)). Ook in die Statistiek het die toestandruimte benadering van 'n tydreeksmodel baie toepassingsmoontlikhede, onder andere om verlore data te kan hanteer. Aangesien op verskeie plekke in die literatuur slegs die finale vorm van die toestandruimte voorstelling van 'n outoregressiewe-bewegendegemiddelde tydreeks gegee word, word in Afdeling 2.5 'n weergawe van die afleiding van die proses gegee. Ten slotte (Afdeling 2.6) word die resultate van hierdie hoofstuk kortliks saamgevat.

## 2.2 TOESTANDRUIMTE MODELLE

In 'n toestandsruimtemodel word die aandag toegespits op 'n  $m \times 1$  vektor van toestandsverandelikes,  $\alpha_t$ . Die toestandsvektor is nie direk waarneembaar nie, maar sy ontwikkeling oor tyd word beheer deur 'n goedgedefinieerde proses. Die proses word

omskrywe deur 'n aanvangswaarde,  $\alpha_0$  en 'n oorgangsvergelyking (*transition equation*) wat geskryf word as 'n eerste orde Markovproses

$$\alpha_t = T_t \alpha_{t-1} + R_t \eta_t \quad t=1, \dots, T \quad 2.2.1$$

met  $T_t$  en  $R_t$  vaste matrikse van orde  $m \times m$  en  $m \times g$  respektiewelik en  $\eta_t$  'n  $g \times 1$  vektor van versteurings met gemiddeld nul en kovariansiematriks  $Q_t$ .

Die  $N$  veranderlikes wat in werklikheid waargeneem word, word omskryf deur 'n  $N \times 1$  vektor,  $y_t = (y_{t1}, \dots, y_{tN})$ , wat verband hou met die toestandsveranderlikes. Die verband word weergegee deur die metingsvergelyking (*measurement equation*) wat geskryf kan word as

$$y_t = Z_t \alpha_t + S_t \xi_t \quad t=1, \dots, T, \quad 2.2.2$$

waar  $Z_t$  en  $S_t$  vaste matrikse van orde  $N \times m$  en  $N \times n$  respektiewelik is. Die  $n \times 1$  vektor van versteurings,  $\xi_t$ , het gemiddeld nul en kovariansiematriks,  $H_t$ .

Roode (1987) gee die volgende benaming vir die konstante matrikse:

$T_t$ : Toestandoorgangsmatriks

$R_t$ : Aandryfmatriks

$Z_t$ : Uitvoermatriks

$S_t$ : Oorseinmatriks

Vergelykings 2.1.1 en 2.2.2 staan bekend as die **toestandruimte vorm** van 'n lineere dinamiese model (*linear dynamic model*).

Daar word aangeneem dat die versteurings in beide die metings—en oorgangsvergelyking reeks ongekorreleerd is. Verder is hulle ongekorreleerd met mekaar oor alle tydperiodes asook met die aanvangstoestandsvektor,  $\alpha_0$ . Die aannames kan as volg opgesom word:

$$\begin{bmatrix} \xi_t \\ \eta_t \end{bmatrix} \sim \text{WR} \left[ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} H_t & 0 \\ 0 & Q_t \end{bmatrix} \right], \quad t = 1, \dots, T$$

waarby

$$E \begin{bmatrix} \alpha_0 \\ \eta_t \end{bmatrix} = 0 \quad E \begin{bmatrix} \alpha_0 \\ \xi_t \end{bmatrix} = 0 \quad t=1, \dots, T$$

en WR staan vir 'witruis'.

In die meeste gevalle is slegs een waarneming per tydsinterval beskikbaar en reduceer 2.2.2 na 'n enkele metingsvergelyking

$$y_t = z_t' \alpha_t + \zeta_t \quad t = 1, \dots, T \quad 2.2.3$$

waar  $\zeta_t \sim \text{WR}(0, \sigma^2 h_t)$

$T_t$ ,  $R_t$  en  $Q_t$  van die oorgangsvergelykings is ook meestal tyd—onafhanklik en derhalwe vereenvoudig 2.2.1 na

$$\alpha_t = T \alpha_{t-1} + R \eta_t \quad t = 1, \dots, T \quad 2.2.4$$

waar  $\eta_t \sim \text{WR}(0, \sigma^2 Q)$ .

### Opmerking

Aangesien  $\sigma^2$  nie in die opdaterings—en voorspellingsvergelykings voorkom nie, kan dit, sonder verlies aan algemeenheid, in die uitdrukkings vir  $\text{var}(\xi_t)$  en  $\text{var}(\eta_t)$  weggelaat word. Die filter is dus wel afhanklik van die waardes van  $T$ ,  $R$ ,  $Q$  en  $h_t$ , maar  $\sigma^2$  hoef nie bekend te wees nie. Aangesien  $\sigma^2$  in baie modelle voorkom word dit vervolgens wel ingesluit.

Die Kalmanfiltervergelykings word afgelei vir die model 2.2.3 en 2.2.4 en kan veralgemeen word vir die model 2.2.1 en 2.2.2.

## 2.3 DIE KALMANFILTER

Aangesien die ingenieursliteratuur, waar die toestandsruimtemodelle en Kalmanfilters vrylik gebruik word, (sien bv. Kalman(1970)) moeilik lees, word daar hoofsaaklik van Harvey(1981) se beskrywing gebruik gemaak.

Die Kalmanfilter word gedefinieer as 'n stel vergelykings wat 'n beramer in staat stel om opgedateer te word die oomblik wanneer 'n nuwe waarneming beskikbaar word. Die proses word in twee stappe uitgevoer. In die eerste plek word daar, met behulp van voorspellingsvergelykings (*prediction equations*), 'n optimale voorspelling van die volgende waarneming, op grond van al die inligting huidiglik beskikbaar, gemaak. Die nuwe waarneming word dan geïnkorporeer in die beramer van die toestandsvektor met behulp van opdateringsvergelykings (*updating equations*).

Die filtervergelykings word rekursief toegepas soos wat nuwe waarnemings beskikbaar gestel word. As 'n neweproduk word 'n reeks voorspellingsfoute geproduseer wat weer ingespan word om MA—beramers te vind met behulp van die voorspellings—fout—ontbinding van die aanneemlikheidsfunksie.

## 2.4 DIE VOORSPELLINGS–FOUT–ONTBINDING VAN DIE AANNEEMLIKHEIDSFUNKSIE

Beskou 'n steekproef  $y_1, \dots, y_T$  getrek uit 'n meerveranderlike normaalverdeling met gemiddeld  $\mu$  en variansie–kovariansiematriks  $\sigma^2 V$ , dit is  $\mathbf{y} \sim N(\mu, \sigma^2 V)$ .

Die aanneemlikheidsfunksie van 'n steekproef  $y_1, \dots, y_T$  getrek uit 'n populasie met waarskynlikheidsverdeling  $N(\mu, \sigma^2 V)$ , is die waarskynlikheidsfunksie van die gesamentlike verdeling van  $T$  onderling onafhanklike stogastiese veranderlikes elk met 'n normaalverdeling. Dus volg dat

$$\ln L(\mathbf{y}) = -(T/2)\ln 2\pi - (T/2)\ln \sigma^2 - (1/2)\ln |V| - (1/2)\sigma^{-2}(\mathbf{y}-\mu)'V^{-1}(\mathbf{y}-\mu) \quad 2.4.1$$

as volg geskryf kan word

$$L(\mathbf{y}) = L(y_1, \dots, y_{T-1}) \cdot \ell(y_T/y_{T-1}, \dots, y_1) \quad 2.4.2$$

met  $\ell$  die voorwaardelike verdeling van die laaste waarneming, gegee al die vorige waarnemings.

Gevolgtik is:

$$\ln L(\mathbf{y}) = \ln L(y_1, \dots, y_{T-1}) + \ln \ell(y_T/y_{T-1}, \dots, y_1) \quad 2.4.3$$

Laat

$\hat{y}_{T/T-1}$  die beramer van  $y_T$  gegee  $y_{T-1}, \dots, y_1$  voorstel.

Die voorspellingsfout:

$y_T - \hat{y}_{T/T-1}$  kan ontbind word in twee dele:

$$y_T - \hat{y}_{T/T-1} = [y_T - E(y_T/y_{T-1}, \dots, y_1)] + [E(y_T/y_{T-1}, \dots, y_1) - \hat{y}_{T/T-1}]$$

waar  $E(y_T/y_{T-1}, \dots, y_1)$  die voorwaardelike verwagtingswaarde van  $y_T$  gegee  $y_{T-1}, \dots, y_1$  is.

Die **gemiddelde-kwadraatberamer(GKB)** (*mean square estimator*) van  $y_T$  onder voorwaarde van  $y_{T-1}, \dots, y_1$  is :

$$\begin{aligned} E[(y_T - \hat{y}_{T/T-1})^2] &= E[y_T - E(y_T/y_{T-1}, \dots, y_1)]^2 \\ &+ E[E(y_T/y_{T-1}, \dots, y_1) - \hat{y}_{T/T-1}]^2 \\ &= \text{var}(y_T/y_{T-1}, \dots, y_1) \\ &+ E[\hat{y}_{T/T-1} - E(y_T/y_{T-1}, \dots, y_1)]^2 \end{aligned}$$

Gevolgtik is die **minimum-gemiddelde-kwadraatberamer(MGKB)** van  $y_T$  gegee  $y_{T-1}, \dots, y_1$ :

$$\tilde{y}_{T/T-1} = E(y_T/y_{T-1}, \dots, y_1).$$

Die voorspellingsfoutvariëansie geassosieer met  $\tilde{y}_{T/T-1}$  word gegee deur

$$\begin{aligned} E(y_T - \tilde{y}_{T/T-1})^2 &= E[(y_T - E(y_T/y_{T-1}, \dots, y_1))]^2 \\ &= \text{var}(y_T/y_{T-1}, \dots, y_1) \\ &= \sigma^2_{f_T} \end{aligned}$$

Aangesien die voorwaardelike verdelings van 'n gesamentlike normaalverdeling weer normaal verdeel is, (Anderson 1966), volg dat

$$\ln \mathcal{L}(y_T/y_{T-1}, \dots, y_1) = -(1/2)\ln 2\pi - (1/2)\ln \sigma^2 - (1/2)\ln f_T - (1/2)\sigma^{-2}(y_T - \bar{y}_{T/T-1})^2 f_T,$$

Herhaaldelike toepassing van 2.4.2 lewer

$$\ln L(y) = \sum_{t=2}^T \ln \mathcal{L}(y_t/y_{t-1}, \dots, y_1) + \ln \mathcal{L}(y_1) \quad 2.4.4$$

met  $\mathcal{L}(y_1)$  die onvoorwaardelike verdeling van  $y_1$ . Sou  $\mu_1$  egter beskou word as die minimum-gemiddelde-kwadraatberamer van  $y_1$ , gegee geen vorige waarnemings, kan die term  $y_1 - \mu_1$  beskou word as die voorspellingsfout geassosieer met  $y_1$  en mag ons die variansie van  $y_1$  aandui as  $\sigma^2 f_1$ .

Gevolgtik word 2.4.4 die gesamentlike verdeling van T onafhanklike voorspellingsfoute,

$$v_t = y_t - \bar{y}_{t/t-1}, \quad t=1, \dots, T,$$

met gemiddeld nul, variansie  $\sigma^2 f_t$  en  $\bar{y}_{1/0} = \mu_1$ .

Derhalwe kan 2.4.4 geskryf word as:

$$\ln L(y) = -(T/2)\ln 2\pi - (T/2)\ln \sigma^2 - (1/2) \sum_{t=1}^T \ln f_t - (1/2)\sigma^{-2} \sum_{t=1}^T v_t^2 / f_t \quad 2.4.5$$

wat bekend staan as die voorspellings-fout-ontbinding van die aanneemlikheidsfunksie.

## 2.5 DIE TOESTANDRUIMTE VOORSTELLING VAN 'N OUTOREGRESSIEWE—BEWEGENDEGEMIDDELDE PROSES

Op verskeie plekke in die literatuur word slegs die finale vorm van die toestandruimte voorstelling van 'n outoregressiewe—bewegendegemiddelde proses (*autoregressive moving average* (ARMA) —model) gegee. Sien byvoorbeeld Harvey (1981) hoofstuk 4, Brockwell en Davis (1986) hoofstuk 12 en Roode (1987). Vir beter insae word daar gekyk na die afleiding soos weergegee deur Jones (1980).

Toestandruimte modelle van toevalsprosesse is gebaseer op die sogenaamde Markov—eienskap, wat impliseer dat die toekoms van 'n proses, gegee sy huidige toestand, onafhanklik is van sy verlede. Abraham en Ledolter (1983) sê dan ook: *In such a system the STATE of the process summarizes all the information from the past that is necessary to predict the future*

In die toestandruimte (ook genoem die Markoviaanse—) voorstelling word die toestandsvektor  $\alpha_t$  dus geïnterpreteer as die vektor bestaande uit al die inligting wat nodig en voldoende is om, op tydstip  $t$ , die reeks arbitrêr ver in die toekoms te voorspel.

### Voorbeeld 2.5.1

Beskou 'n outoregressiewe model van orde twee ('n AR(2)—model):

$$\begin{aligned} \hat{y}_{t+2} = y_{t+2/t} &= E_t[y_{t+2}] \\ &= E[y_{t+2}/y_s, s \leq t] \end{aligned}$$

$$\begin{aligned}
&= E_t[\phi_1 y_{t+1} + \phi_2 y_t + a_{t+1}] \\
&= \phi_1 \hat{y}_{t+1} + \phi_2 y_t \\
&= \phi_1 y_{t+1/t} + \phi_2 y_t
\end{aligned}
\tag{2.5.1}$$

Vir verdere verklaring van notasie en die minimum-gemiddelde-kwadraatfoutvoorspelling word die leser verwys na Box en Jenkins(1976), Hoofstuk 5.

Vir die AR(2)-model, om die reeks arbitrêr ver in die toekoms te voorspel is die paar  $y_t$  en  $\hat{y}_{t+1}$  nodig en sou  $\alpha_t$  dus geskryf kon word as:

$$\alpha_t = (y_t, y_{t+1/t})
\tag{2.5.2}$$

Soortgelyk kan die *toestand* van die algemene ARMA(p,q)-model, met gemiddeld nul:

$$y_t = \sum_{k=1}^p \phi_k y_{t-k} + \eta_t + \sum_{k=1}^q \theta_k \eta_{t-k} \quad (\text{Box en Jenkins-notasie})$$

geskryf word as:

$$\alpha_t = [y_{t/t}, y_{t+1/t}, \dots, y_{t+m-1/t}]'
\tag{2.5.3}$$

met

$$y_{t/t} = y_t$$

en

$$m = \text{maksimum}(p, q+1)$$

Die een-stap-voorspelling is:

$$y_{t+1/t} = \sum_{k=1}^p \phi_k y_{t+1-k} + \sum_{k=1}^q \theta_k \eta_{t+1-k}$$

Die j-stap-voorspelling gebruik die vorige voorspellings in die outoregressiewe deel:

$$\begin{aligned} y_{t+j/t} &= \sum_{k=1}^{j-1} \phi_k y_{t+j-k/t} + \sum_{k=j}^p \phi_k y_{t+j-k} \\ &\quad + \sum_{k=j}^q \theta_k \eta_{t+j-k} \end{aligned} \quad 2.5.4$$

Deur bv.  $p=2$ ,  $q=0$ ,  $j=2$  in 2.5.4 te stel word 2.5.1 gekry.

Soortgelyk is:

$$\begin{aligned} y_{t+j/t+1} &= \sum_{k=1}^{j-2} \phi_k y_{t+j-k/t+1} + \sum_{k=j-1}^p \phi_k y_{t+j-k} \\ &\quad + \sum_{k=j-1}^q \theta_k \eta_{t+j-k} \end{aligned} \quad 2.5.5$$

Uit 2.5.4 en 2.5.5 volg dat

$$\begin{aligned} y_{t+j/t+1} - y_{t+j/t} &= \sum_{k=1}^{j-1} \phi_k (y_{t+j-k/t+1} - y_{t+j-k/t}) \\ &\quad + \theta_{j-1} \eta_{t+1} \end{aligned} \quad 2.5.6$$

Beskou  $j=2$ :

$$\begin{aligned} y_{t+2/t+1} - y_{t+2/t} &= \phi_1(y_{t+1/t+1} - y_{t+1/t}) + \theta_1\eta_{t+1} \\ &= (\phi_1 + \theta_1)\eta_{t+1} \end{aligned} \quad 2.5.7$$

Beskou  $j=3$ :

$$\begin{aligned} y_{t+3/t+1} - y_{t+3/t} &= \phi_1(y_{t+2/t+1} - y_{t+2/t}) + \phi_2(y_{t+1/t+1} - y_{t+1/t}) \\ &\quad + \theta_2\eta_{t+1} \end{aligned}$$

M.b.v. 2.5.7 volg soortgelyk

$$\begin{aligned} &= \phi_1(\phi_1 + \theta_1)\eta_{t+1} + \phi_2\eta_{t+1} + \theta_2\eta_{t+1} \\ &= (\phi_1^2 + \phi_1\theta_1 + \phi_2 + \theta_2)\eta_{t+1} \end{aligned} \quad 2.5.8$$

Vanuit 2.5.7 en 2.5.8 is dit duidelik dat 2.5.6 'n rekursiewe vergelyking is wat slegs 'n funksie is van die ewekansige skokterm  $\eta$  op tydstep  $t+1$ .

Gevolglik kan 2.5.6 geskryf word as

$$y_{t+j/t+1} - y_{t+j/t} = \xi_j\eta_{t+1}$$

of

$$y_{t+j/t+1} = y_{t+j/t} + \xi_j\eta_{t+1} \quad 2.5.9$$

waar die  $g$ 's gegenerer word deur die rekursiewe betrekking

$$\begin{aligned} \xi_1 &= 1 \\ \xi_j &= \theta_{j-1} + \sum_{k=1}^{j-1} \phi_k \xi_{j-k} \end{aligned}$$

waar

$$\theta_j = 0 \quad \text{vir } j > q$$

Stel byvoorbeeld  $j=3$  om weer 2.5.8 te verkry

Die finale element in die toestandsvektor is:

$$y_{t+m/t+1} = y_{t+m/t} + g_m \eta_{t+1} \quad 2.5.10$$

Aangesien:

$$y_{t+m} = \sum_{k=1}^p \phi_k y_{t+m-k} + \sum_{k=1}^q \theta_k \eta_{t+m-k}$$

en

$$E_t[\eta_{t+k}] = 0 \quad k=1,2,\dots$$

kan 2.5.10 geskryf word as:

$$y_{t+m/t+1} = \sum_{k=1}^p \phi_k y_{t+m-k/t} + g_m \eta_{t+1} \quad 2.5.11$$

Die matriksformulering van 2.5.9 en 2.5.11 naamlik

$$\begin{bmatrix} y_{t+1} \\ y_{t+2/t+1} \\ \vdots \\ y_{t+m/t+1} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \dots & \dots & 0 \\ 0 & 0 & 1 & \dots & \dots & 0 \\ & & & \cdot & \cdot & \\ \phi_m & \cdot & \dots & \phi_2 & \phi_1 & \end{bmatrix} \begin{bmatrix} y_t \\ y_{t+1/t} \\ \vdots \\ y_{t+m-1/t} \end{bmatrix} + \begin{bmatrix} 1 \\ g_2 \\ \vdots \\ g_m \end{bmatrix} \eta_{t+1} \quad 2.5.12$$

waar  $\phi_i = 0$  vir  $i > p$

lewer die sogenaamde oorgangvergelings.

Bogenoemde kan ook geskryf word as:

$$\alpha_{t+1} = T\alpha_t + R\eta_{t+1} \quad \text{vergelijk met 2.2.1}$$

Vergelyking 2.2.12 staan bekend as die toestandsruimte vorm van 'n ARMA(p,q)–proses.

Die tweede vergelyking in die toestandsruimte voorstelling naamlik die waarnemingsvergelyking word gegee deur:

$$y_t = z\alpha_t + \xi_t$$

met

$$z = (1, 0, \dots, 0) \text{ en } \xi_t \text{ die foutterm.}$$

Die waarnemingsvergelyking onttrek dus bloot die eerste element van die toestandsvektor.

### Voorbeeld 2.5.2

Beskou weer die AR(2)–model:  $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + a_t$

Die vergelyking word uit die oorgangvergelyking herwin deur herhaalde substitusie:

$$\alpha_t = \begin{bmatrix} y_t \\ y_{t+1/t} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ \phi_2 & \phi_1 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t/t-1} \end{bmatrix} + \begin{bmatrix} 1 \\ \phi_1 \end{bmatrix} a_t$$

deur uitvermenigvuldiging vind ons:

$$y_t = y_{t/t-1} + a_t \tag{2.5.13}$$

wat korrek is aangesien die waargenome waarde gelyk is aan die som van die beraamde waarde en die foutterm.

Verder volg dat

$$\begin{aligned} y_{t+1/t} &= \phi_2 y_{t-1} + \phi_1 y_{t/t-1} + \phi_1 a_t \\ &= \phi_2 y_{t-1} + \phi_1 (y_{t/t-1} + a_t) \end{aligned}$$

$$= \phi_2 y_{t-1} + \phi_1 y_t \quad 2.5.14$$

Vervang 2.5.14 in 2.5.13, dan volg

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + a_t$$

### Voorbeeld 2.5.3

Op soortgelyke wyse kan bevestig word dat 'n MA(1)-model:  $y_t = \eta_t + \theta \eta_{t-1}$  geskryf kan word as:

$$\alpha_{t+1} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \alpha_t + \begin{bmatrix} 1 \\ \theta \end{bmatrix} \eta_{t+1}$$

## 2.6 OPSOMMING

In hierdie hoofstuk is begrippe gedefinieer en hulpresultate afgelei:

1. Die begrippe **toestandruimte modelle** en die **Kalmanfilter** is gedefinieer. In die volgende hoofstuk word die Kalmanfilter-vergelykings vervolgens afgelei.
2. Daar is aangetoon hoe 'n outoregressiewe-bewegendegemiddelde proses in 'n toestandruimte vorm omskryf kan word. Ter illustrasie is 'n paar voorbeelde behandel.
3. As hulpresultaat tot Hoofstuk 3 is die voorspellings-fout-ontbinding van die aanneemlikheidsfunksie uiteengesit.

## HOOFSTUK 3

### DIE KALMANFILTER VERGELYKINGS

#### 3.1 INLEIDING

Die toestandruimte modelle gee aanleiding tot 'n stel vergelykings wat 'n beraming van 'n spesifieke "toestand" maak wat op 'n gegewe stadium waargeneem is. Die probleem om die beramings, onder bepaalde aannames, rekursief te verkry, was oorspronklik opgelos deur Kalman (1960) en Kalman en Bucy (1961) en staan bekend as die sogenaamde Kalmanfilter.

Die Kalmanfilter word gedefinieer as 'n stel vergelykings wat 'n beramer in staat stel om opgedateer te word die oomblik wanneer 'n nuwe waarneming beskikbaar word. Die proses word in twee stappe uitgevoer. In die eerste plek word daar, met behulp van voorspellingsvergelykings, 'n optimale voorspelling van die volgende waarneming, op grond van al die inligting huidiglik beskikbaar gemaak. Die nuwe waarneming word dan geïnkorporeer in die beramer van die toestandsvektor met behulp van opdateringsvergelykings. In Afdeling 3.3 word hierdie vergelykings afgelei vir die spesiale geval van een waarneming per tydsinterval. Die resultate vir die algemene geval van  $N$  waarnemings per tydsinterval word in Afdeling 3.4 aangebied.

Die Kalmanfilter verskaf 'n optimale oplossing tot die probleem van vooruitskatting en opdatering. Sou die waarnemings normaal verdeel en die huidiglike beramer van die toestandsvektor die "beste" wees, sal die vooruitskatting- en die opdateringsberamers ook die beste wees.

Sou normaliteit nie aangeneem kan word nie, is soortgelyke resultate geldig, maar nou binne 'n klas van beramers en voorspellers wat lineër in die waarnemings is. Aangesien die toestandsvektor stogasties is, moet die interpretasie van die "beste" beramer met omsigtigheid hanteer word. Daar word eerder gekyk na die beramingsfout as na die beramer self. Dit lei tot die begrip van die **minimum gemiddelde kwadraatberamer** wat in Afdeling 3.2 die aandag geniet as voorloper tot die afleiding van die Kalmanvergelings.

In Afdeling 3.5 word aangetoon hoe om aanvangswaardes vir die onbekende parameters te bepaal terwyl die resultate prakties geïllustreer word met behulp van 'n MA(1)-proses soos uiteengesit in Afdeling 3.6.

Resultate verkry in hierdie hoofstuk word in Afdeling 3.7 saamgevat.

### 3.2 BASIESE RESULTATE (Harvey 1981)

Beskou die model

$$\mathbf{y} = \mathbf{Z} \boldsymbol{\alpha} + \boldsymbol{\xi}, \quad 3.2.1$$

waarby  $\mathbf{y}$  'n  $T \times 1$  vektor van waarnemings,  $\mathbf{Z}$  'n  $T \times m$  matriks van vaste waardes en  $\boldsymbol{\xi}$  'n vektor van versteurings met gemiddeld nul en kovariansie matriks  $\sigma^2 \boldsymbol{\Omega}$  is.

Vir  $\boldsymbol{\alpha}$  'n vaste vektor van onbekende parameters neem 3.2.1 die vorm van 'n algemene regressiemodel aan en is die gewone resultate geldig. Die **algemene kleinste kwadrate beramer**,  $\tilde{\mathbf{a}}$ , van  $\boldsymbol{\alpha}$  word gegee deur

$$\tilde{\mathbf{a}} = (\mathbf{Z}' \boldsymbol{\Omega}^{-1} \mathbf{Z})^{-1} \mathbf{Z}' \boldsymbol{\Omega}^{-1} \mathbf{y}$$

en is ook dieselfde as die **beste lineêre onsydige skatter** (b.l.o.s.) van  $\boldsymbol{\alpha}$ .

### 3.3

Die Gauss–Markovteorie beweer dat die "beste" daarop neerkom dat onder die klas van lineêr onsydige beramers die "beste" beramer die kleinste variansie besit.

Sou  $\alpha$  egter beskou word as stogasties in die sin dat dit ewekansig getrek is uit 'n prior verdeling voordat die waarnemings  $\mathbf{y}$  gegeneer is, word die statistiese eienskappe van  $\bar{\mathbf{a}}$  gegee in terme van die beramingsfout,  $\bar{\mathbf{a}} - \alpha$ . So word  $\bar{\mathbf{a}}$  nou 'n onvoorwaardelike onsydige beramer genoem indien die verwagtingswaarde van die beramingsfout nul is. Die eienskappe van  $\bar{\mathbf{a}} - \alpha$  is egter parallel aan die van die konvensionele algemene kleinste vierkante skatter en besit ook variansie/kovariansie matriks:  $\sigma^2(\mathbf{Z}'\Omega^{-1}\mathbf{Z})^{-1}$ .

'n Uitbreiding van die Gauss–Markovstelling is om aan te toon dat  $\bar{\mathbf{a}}$ , binne 'n klas van lineêr en onvoorwaardelik onsydige beramers, die **minimum gemiddelde kwadraadberamer** is:

Veronderstel, sonder verlies van algemeenheid, dat  $\Omega=I$  (die eenheidsmatriks).

Laat  $\hat{\mathbf{a}} = \mathbf{D}^* \mathbf{y}$  enige ander lineêr, onvoorwaardelik onsydige beramer van  $\alpha$  wees, dan is:

$$\begin{aligned} \hat{\mathbf{a}} - \alpha &= \mathbf{D}^* \mathbf{y} - \alpha \\ &= \mathbf{D}^* (\mathbf{Z}\alpha + \xi) - \alpha \\ &= [\mathbf{D}^* - (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}']\mathbf{Z}\alpha + [\mathbf{D}^* - (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' + (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}']\xi \\ &= \mathbf{D}\mathbf{Z}\alpha + [\mathbf{D} + (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}']\xi \end{aligned}$$

met

$$\mathbf{D} = \mathbf{D}^* - (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$$

Om onvoorwaardelik onsydig te wees, moet:

$$E[\hat{\mathbf{a}} - \alpha] = 0$$

dus moet:

$$\mathbf{D}\mathbf{Z} = 0$$

Onder hierdie voorwaarde is die kovariansiematriks van  $\hat{\mathbf{a}} - \boldsymbol{\alpha}$ , wat dan ook gelyk is aan die minimum gemiddelde kwadraatberamer van  $\boldsymbol{\alpha}$ , die volgende:

$$\begin{aligned} \text{var}(\hat{\mathbf{a}} - \boldsymbol{\alpha}) &= \text{E}[(\hat{\mathbf{a}} - \boldsymbol{\alpha})(\hat{\mathbf{a}} - \boldsymbol{\alpha})'] \\ &= [D + (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}']\text{E}(\boldsymbol{\xi}\boldsymbol{\xi}')[\mathbf{D}' + \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}] \end{aligned} \quad 3.2.2$$

Aangesien

$$\text{E}(\boldsymbol{\xi}\boldsymbol{\xi}') = \sigma^2\mathbf{I}$$

en

$$\mathbf{D}\mathbf{Z} = 0$$

vereenvoudig 3.2.2 na:

$$\text{var}(\hat{\mathbf{a}} - \boldsymbol{\alpha}) = \sigma^2\mathbf{D}\mathbf{D}' + \sigma^2(\mathbf{Z}'\mathbf{Z})^{-1}$$

Dit is hieruit duidelik dat die kovariansiematriks groter is as die van  $\tilde{\mathbf{a}} - \boldsymbol{\alpha}$ .

Sou die waarnemings normaal verdeel wees, is die optimale eienskappe van  $\tilde{\mathbf{a}}$  nie langer beperk tot die klas van lineêre beramers nie en word  $\tilde{\mathbf{a}}$  die **minimum gemiddelde kwadraatberamer** van  $\boldsymbol{\alpha}$ .

Veronderstel verder dat kennis oor  $\boldsymbol{\alpha}$  reeds beskikbaar is en vervat is in die vektor  $\mathbf{a}_0$ , sodanig dat  $\mathbf{a}_0 - \boldsymbol{\alpha}$  'n gemiddeld nul en kovariansiematriks  $\sigma^2\mathbf{P}_0$  ( $\mathbf{P}_0$  bekend), besit.

Skryf die aangevulde model as:

$$\begin{bmatrix} \mathbf{a}_0 \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{I} \\ \mathbf{Z} \end{bmatrix} \boldsymbol{\alpha} + \begin{bmatrix} \mathbf{a}_0 - \boldsymbol{\alpha} \\ \boldsymbol{\xi} \end{bmatrix}$$

Wat geskryf kan word as:

$$\mathbf{y}^\dagger = \mathbf{Z}^\dagger \boldsymbol{\alpha} + \boldsymbol{\xi}^\dagger$$

met  $\text{E}(\boldsymbol{\xi}^\dagger) = 0$  en  $\text{E}(\boldsymbol{\xi}^\dagger \boldsymbol{\xi}^{\dagger'}) = \sigma^2$

$$\begin{bmatrix} \mathbf{P}_0 & 0 \\ 0 & \boldsymbol{\Omega} \end{bmatrix}$$

Alhoewel  $\alpha$  nou sodanig gedefinieer is, dat dit die moontlikheid insluit dat sommige, of alle, elemente vas is, is die uitgebreide Gauss–Markovteorie steeds geldig en is die beramer

$$\bar{\mathbf{a}}^\dagger = (\mathbf{Z}^\dagger' \mathbf{V}^{-1} \mathbf{Z}^\dagger)^{-1} \mathbf{Z}^\dagger' \mathbf{V}^{-1} \mathbf{y}^\dagger \quad 3.2.3$$

die minimum gemiddelde lineêre kwadraatberamer van  $\alpha$  met die variansie kovariansie matriks van  $(\bar{\mathbf{a}}^\dagger - \alpha)$  gelyk aan  $\sigma^2 \mathbf{P}$ .

Wanneer die oorspronklike notasie vervang word in 3.2.3 word gekry

$$\begin{aligned} \bar{\mathbf{a}}^\dagger &= \mathbf{P}(\mathbf{P}_0^{-1} \mathbf{a}_0 + \mathbf{Z}' \Omega^{-1} \mathbf{y}), \\ \mathbf{P} &= (\mathbf{P}_0^{-1} + \mathbf{Z}' \Omega^{-1} \mathbf{Z})^{-1} \end{aligned}$$

### 3.3 DIE KALMANFILTERVERGELYKINGS.

In Hoofstuk 2 is die oorgangsvergelyking(2.2.1) en metingsvergelyking(2.2.2) beskou wat gerieflikheidsonthelwe hier herhaal word:

$$\alpha_t = \mathbf{T}_t \alpha_{t-1} + \mathbf{R}_t \eta_t \quad t=1, \dots, T \quad 3.3.1$$

en

$$\mathbf{y}_t = \mathbf{Z}_t \alpha_t + \mathbf{S}_t \xi_t \quad t=1, \dots, T \quad 3.3.2$$

waarby die  $N$ -veranderlike vektor  $\mathbf{y}_t = (y_{1t}, \dots, y_{Nt})'$ ,  $t = 1, \dots, N$  beskou is.

In baie gevalle is daar egter slegs een waarneming per tydsinterval beskikbaar.

Vergelyking 3.3.2 reduceer in sodanige gevalle tot 'n enkele metingsvergelyking:

$$y_t = \mathbf{z}_t' \alpha_t + \xi_t, \quad t = 1, \dots, T, \quad 3.3.3$$

waar  $\xi$  -  $WR(0, \sigma^2 h_t)$

Die oorgangsvergelyking 3.3.1 vereenvoudig na

$$\alpha_t = T\alpha_{t-1} + R\eta_t, t = 1, \dots, T \quad 3.3.4$$

met  $\eta_t \sim \text{WR}(0, \sigma^2 Q)$

Die teenwoordigheid van  $\sigma^2$  in  $\text{var}(\xi_t)$  en  $\text{var}(\eta_t)$  veroorsaak nie 'n verlies aan algemeenheid nie, aangesien dit nie in die Kalmanvergelykings verskyn nie.

Die Kalmanfiltervergelykings word vervolgens vir 3.3.3 en 3.3.4 afgelei en veralgemeen vir die sisteem 3.3.1 en 3.3.2.

Ten einde die nodige vergelykings af te lei, word die volgende aannames gemaak:

$\hat{\mathbf{a}}_t$  is die minimum-gemiddelde-lineêre-kwadraatberamer (MGLKB) van  $\alpha_t$  gebaseer op al die inligting tot en met die huidige waarneming,  $y_t$ .

$\hat{\mathbf{a}}_{t|t-1}$  is die minimum-gemiddelde-lineêre-kwadraatvoorspeller (MGLKV) van  $\alpha_t$  op tydstip  $t-1$ .

Soos reeds in Hoofstuk 2 gestel, word die Kalmanfilter-vergelykings verdeel in die sogenaamde voorspellings- en opdateringsvergelykings.(Harvey(1981)).

**Die voorspellings-vergelykings.**

Op tydstip  $t-1$  is al die beskikbare inligting vervat in  $\hat{\mathbf{a}}_{t-1}$ , die MGLKB, van  $\alpha_{t-1}$ , met kovariansiematriks  $\sigma^2 P_{t-1}$  en  $P_{t-1}$  bekend.

Die vergelyking  $\alpha_t = T\alpha_{t-1} + R\eta_t$  impliseer dat die MGLKB van  $\alpha_t$  op tydstip  $t-1$  gegee word deur (dit kan gesien word as 'n mens lees: "die verwagtingswaarde van  $\alpha_t$  op tydstip  $t-1$  (dit is  $a_{t/t-1}$ ) = T x die verwagtingswaarde van  $\alpha_{t-1}$  op tydstip  $t-1$ , (dit is  $a_{t-1}$ ) + R x die verwagtingswaarde van  $\eta_t$  op tydstip  $t-1$ , dit is nul")

$$a_{t/t-1} = T a_{t-1} \quad 3.3.5$$

met voorspellingsfout

$$\begin{aligned} a_{t/t-1} - \alpha_t &= T a_{t-1} - [T \alpha_{t-1} + R \eta_t] \\ &= T [a_{t-1} - \alpha_{t-1}] - R \eta_t \end{aligned}$$

Let op dat  $a_{t/t-1}$  onvoorwaardelik onsydig is, aangesien die verwagtingswaarde van die voorspellingsfout  $a_{t/t-1} - \alpha_t$  nul is.

Die kovariansiematriks van die voorspellingsfout word gegee deur:

$$\begin{aligned} &E \left[ \left[ a_{t/t-1} - \alpha_t \right] \left[ a_{t/t-1} - \alpha_t \right]' \right] \\ &= E \left[ \left[ T [a_{t-1} - \alpha_{t-1}] - R \eta_t \right] \left[ T [a_{t-1} - \alpha_{t-1}] - R \eta_t \right]' \right] \end{aligned}$$

aangesien die kruisproduk nul is reduceer dit na:

$$\begin{aligned} &= E \left[ T [a_{t-1} - \alpha_{t-1}] [a_{t-1} - \alpha_{t-1}]' T' \right] + E [R \eta_t \eta_t' R'] \\ &= \sigma^2 T P_{t-1} T' + \sigma^2 R Q R' \end{aligned}$$

$$= \sigma^2 [TP_{t-1}T' + RQR']$$

$$= \sigma^2 P_{t/t-1}$$

Dus  $\mathbf{a}_{t/t-1} - \boldsymbol{\alpha}_t$  het verwagtingswaarde nul en variansie-kovariansiematriks  $\sigma^2 P_{t/t-1}$  met

$$P_{t/t-1} = TP_{t-1}T' + RQR' \quad 3.3.6.$$

bekend.

Vanuit die waarnemingsvergelyking

$$y_t = \mathbf{z}'_t \boldsymbol{\alpha}_t + \xi_t$$

en gegee dat  $\mathbf{a}_{t/t-1}$  die MGLKV van  $\boldsymbol{\alpha}_t$  op tydstep  $t-1$  is, word die MGLKV van  $y_t$  op tydstep  $t-1$  gegee deur

$$\bar{y}_{t/t-1} = \mathbf{z}'_t \mathbf{a}_{t/t-1}$$

met die gepaardgaande voorspellingsfout:

$$v_t = y_t - \bar{y}_{t/t-1} \quad 3.3.7$$

$$= \mathbf{z}'_t [\boldsymbol{\alpha}_t - \mathbf{a}_{t/t-1}] + \xi_t$$

Aangesien die verwagtingswaarde van beide  $(\boldsymbol{\alpha}_t - \mathbf{a}_{t/t-1})$  en  $\xi_t$  nul is,

is  $E(v_t) = 0$  en

$$\begin{aligned} \text{Var}(v_t) &= E(v_t^2) \\ &= E[\mathbf{z}'_t (\boldsymbol{\alpha}_t - \mathbf{a}_{t/t-1}) + \xi_t]^2 \\ &= E[\mathbf{z}'_t (\boldsymbol{\alpha}_t - \mathbf{a}_{t/t-1})(\boldsymbol{\alpha}_t - \mathbf{a}_{t/t-1})' \mathbf{z}_t] + E[\xi_t^2] \\ &\quad + 2E[\mathbf{z}'_t (\boldsymbol{\alpha}_t - \mathbf{a}_{t/t-1}) \xi_t] \end{aligned}$$

$$\begin{aligned}
&= \sigma^2 \mathbf{z}'_t \mathbf{P}_{t/t-1} \mathbf{z}_t + \sigma^2 h_t \\
&= \sigma^2 f_t
\end{aligned}$$

met

$$f_t = \mathbf{z}'_t \mathbf{P}_{t/t-1} \mathbf{z}_t + h_t \quad 3.3.8$$

Die voorspellingsvergelyking word gegee deur 3.3.5 met kovariansiematriks 3.3.6 terwyl 3.3.7 die voorspellingsfout is toe  $y_t$  op tydstep  $y_{t-1}$  voorspel is.

### Die opdateringsvergelyking

Die rol van die opdateringsvergelyking is om, wanneer 'n nuwe waarneming gemaak word se  $\hat{y}_t$ , die bykomende kennis te gebruik om die optimale beramer,  $\mathbf{a}_{t/t-1}$  op te dateer.

Nou het ons 'n analoë situasie as die in Afdeling 3.2. Daar was die voorafgaande inligting, vervat in  $\mathbf{a}_0$ , gekombineer met die steekproefinligting vervat in 3.2.1. Nou is die voorafgaande inligting vervat in die beramer,  $\mathbf{a}_{t/t-1}$ , en die steekproef bestaan uit die enkele waarneming afgelei van die waarnemingsvergelyking  $y_t = \mathbf{z}'_t \boldsymbol{\alpha}_t + \xi_t$ .

Die aangevulde model is dus

$$\begin{bmatrix} \mathbf{a}_{t/t-1} \\ y_t \end{bmatrix} = \begin{bmatrix} \mathbf{I} \\ \mathbf{z}'_t \end{bmatrix} \boldsymbol{\alpha}_t + \begin{bmatrix} \mathbf{a}_{t/t-1} - \boldsymbol{\alpha}_t \\ \xi_t \end{bmatrix}$$

Die versteuringsterm  $\xi_t$  het verwagtingswaarde nul en kovariansiematriks:

$$E \left[ \begin{bmatrix} \mathbf{a}_{t/t-1} - \boldsymbol{\alpha}_t \\ \xi_t \end{bmatrix} \begin{bmatrix} \mathbf{a}'_{t/t-1} - \boldsymbol{\alpha}'_t & \xi'_t \end{bmatrix} \right]$$

$$= \sigma^2 \begin{bmatrix} P_{t/t-1} & 0 \\ 0 & h_t \end{bmatrix}$$

Die MGLKB van  $\alpha_t$  word nou gegee deur (verwys terug na Afdeling 3.2)

$$\mathbf{a}_t = P_t (P_{t/t-1}^{-1} \mathbf{a}_{t/t-1} + \mathbf{z}_t y_t / h_t)^{-1} \quad 3.3.9$$

$$P_t = (P_{t/t-1}^{-1} + \mathbf{z}_t' \mathbf{z}_t / h_t)^{-1} \quad 3.3.10$$

Dus  $(\mathbf{a}_t - \alpha_t)$  het gemiddeld nul en kovariansiematriks  $\sigma^2 P_t$ .

Met behulp van die volgende **matriks-inversielemma** (Jazwinski(1970)) kan  $\mathbf{a}_t$  en  $P_t$  geskryf word in 'n vorm wat geen matriksinversies vereis nie.

Laat D 'n  $n \times n$  matriks wees naamlik

$$D = [A + BCB']^{-1}$$

waar A en C nie-singuliere matrikse van orde n en m respektiewelik is en B  $n \times m$  is.

Dan kan D geskryf word as:

$$D = A^{-1} - A^{-1} B [C^{-1} + B' A^{-1} B]^{-1} B' A^{-1}$$

Stel nou

$$D = P_t, \quad A = P_{t/t-1}^{-1}, \quad B = \mathbf{z}_t \quad \text{en} \quad C = h_t^{-1}$$

Dan word 3.3.10:

$$P_t = P_{t/t-1} - P_{t/t-1} \mathbf{z}_t [h_t + \mathbf{z}_t' P_{t/t-1} \mathbf{z}_t]^{-1} \mathbf{z}_t' P_{t/t-1}$$

$$P_t = P_{t/t-1} - P_{t/t-1} \mathbf{z}_t \mathbf{z}_t' P_{t/t-1} / f_t \quad 3.3.11$$

met

$$f_t = \mathbf{z}_t' P_{t/t-1} \mathbf{z}_t + h_t. \quad 3.3.12$$

Substitusie van 3.3.11 in 3.3.9 lewer

$$\begin{aligned}
 \mathbf{a}_t &= (\mathbf{P}_{t/t-1} - \mathbf{P}_{t/t-1} \mathbf{z}_t' \mathbf{z}_t' \mathbf{P}_{t/t-1} / f_t) (\mathbf{P}_{t/t-1}^{-1} \mathbf{a}_{t/t-1} + \mathbf{z}_t y_t / h_t) \\
 &= \mathbf{a}_{t/t-1} + \mathbf{P}_{t/t-1} \mathbf{z}_t (y_t / h_t - \mathbf{z}_t' \mathbf{a}_{t/t-1} / f_t \\
 &\quad - \mathbf{z}_t' \mathbf{P}_{t/t-1} \mathbf{z}_t y_t / h_t f_t) \\
 &= \mathbf{a}_{t/t-1} + f_t^{-1} \mathbf{P}_{t/t-1} \mathbf{z}_t (y_t f_t / h_t - \mathbf{z}_t' \mathbf{a}_{t/t-1} \\
 &\quad - \mathbf{z}_t' \mathbf{P}_{t/t-1} \mathbf{z}_t y_t / h_t) \\
 &= \mathbf{a}_{t/t-1} + f_t^{-1} \mathbf{P}_{t/t-1} \mathbf{z}_t (y_t \mathbf{z}_t' \mathbf{P}_{t/t-1} \mathbf{z}_t / h_t + y_t - \mathbf{z}_t' \mathbf{a}_{t/t-1} \\
 &\quad - \mathbf{z}_t' \mathbf{P}_{t/t-1} \mathbf{z}_t y_t / h_t) \\
 &= \mathbf{a}_{t/t-1} + \mathbf{P}_{t/t-1} \mathbf{z}_t (y_t - \mathbf{z}_t' \mathbf{a}_{t/t-1}) / f_t
 \end{aligned} \tag{3.3.13}$$

Opmerkings:

1. Vergelykings 3.3.11 tot 3.3.13 is die basiese opdateringsvergelykings.
2. Vergelyking 3.3.13 kan as volg gelees word: "Die beramer op tydstep  $t$  = die voorspeller op tydstep  $t$  (gegee inligting tot op tydstep  $t-1$ ) + die sogenaamde Kalmanvoorsprong (*Kalman gain*) x die Voorspellingsfout".
3. Die Kalmanvoorsprong is die  $m \times 1$  vektor,  $\mathbf{P}_{t/t-1} \mathbf{z}_t' / f_t$ .
4.  $\mathbf{a}_{t/t-1}$  word dus "opgedateer", wanneer  $y_t$  bekend word, via die Kalmanvoorsprong.
5. Let op dat die Kalmanvoorsprong, net soos  $\mathbf{P}_t$  en  $\mathbf{P}_{t/t-1}$ , onafhanklik is van  $y_t$  en vooruit bereken kan word.

6. Bogenoemde afleiding veronderstel dat  $h_t > 0$ . In sommige toestandruimte voorstellings bevat die waarnemingsvergelyking egter geen ruisterme nie. Byvoorbeeld vir 'n ARMA-model kan die waarnemingsvergelyking  $y_t = \mathbf{z}' \boldsymbol{\alpha}_t$  wees (kyk Hoofstuk 3) en is  $h_t = 0$ . Die vraag is nou of die Kalmanvergelykings nog van toepassing is. Harvey beweer die eindresultate is nog steeds geldig en verwys na Theil(1971) p282–287.

### 3.4 ALGEMENE VORM VAN DIE KALMANVERGELYKINGS

Die voorspellings- en opdateringsvergelykings vir die algemene toestandruimte model 3.3.1 en 3.3.2 word op analoë wyse afgelei (Harvey(1981)) as vir die spesiale geval hierbo en slegs die resultate word gegee:

Gegee  $\mathbf{a}_{t-1}$  is die MGLKB van  $\boldsymbol{\alpha}_{t-1}$  op tydstip  $t-1$ , met die gemiddeld en kovariansiematriks van  $\mathbf{a}_{t-1} - \boldsymbol{\alpha}_{t-1}$  gelyk aan nul en  $P_{t-1}$  respektiewelik.

Die voorspellingsvergelykings is dan

$$\mathbf{a}_{t/t-1} = \mathbf{T}_t \mathbf{a}_{t-1} \quad 3.4.1$$

en

$$P_{t/t-1} = \mathbf{T}_t P_{t-1} \mathbf{T}'_t + \mathbf{R}_t \mathbf{Q}_t \mathbf{R}'_t, \quad t = 1, \dots, T \quad 3.4.2$$

Die opdateringsvergelykings word gegee deur:

$$\mathbf{a}_t = \mathbf{a}_{t/t-1} + P_{t/t-1} \mathbf{Z}'_t \mathbf{F}_t^{-1} (\mathbf{y}_t - \mathbf{Z}_t \mathbf{a}_{t/t-1}) \quad 3.4.3$$

en

$$P_t = P_{t/t-1} - P_{t/t-1} \mathbf{Z}'_t \mathbf{F}_t^{-1} \mathbf{Z}_t P_{t/t-1} \quad 3.4.4$$

waar

$$\mathbf{F}_t = \mathbf{Z}_t P_{t/t-1} \mathbf{Z}'_t + \mathbf{S}_t \mathbf{H}_t \mathbf{S}'_t, \quad t = 1, \dots, T \quad 3.4.5$$

Die voorspellingsfout:

$$\mathbf{v}_t = \mathbf{y}_t - \mathbf{Z}_t \mathbf{a}_{t/t-1}, \quad t = 1, \dots, T \quad 3.4.6$$

is nou 'n  $N \times 1$  vektor. Dit besit gemiddeld nul en kovariansiematriks,  $\mathbf{F}_t$ .

Weereens word  $\mathbf{a}_{t/t-1}$  'opgedateer' deur die Kalmanvoorsprong,  $\mathbf{P}_{t/t-1} \mathbf{Z}_t' \mathbf{F}_t^{-1}$  vermenigvuldig met  $\mathbf{v}_t$ .

Opmerking:

Met die aanvangswaardes,  $\mathbf{a}_0$  en  $\mathbf{P}_0$  bekend, gee die Kalmanfilter die MGLKV van  $\alpha_t$  gebaseer op al  $T$  waarnemings.

### 3.5 AANVANGSWAARDES VIR DIE KALMANFILTER

Om die Kalmanrekursies te gebruik, begin die prosedure by 'n aanvangstoestandsvektor,  $\mathbf{a}_0$  en die aanvangskovariansiematriks,  $\mathbf{P}_0$  van die proses.

#### Die Aanvangstoestandsvektor

Per definisie is  $\mathbf{a}_0$  die MGLKB van  $\alpha_0$ . Dit gee dus die beraming van die toestand op tydstip nul gegee alle data tot en met tydstip nul. Anders gestel,  $\mathbf{a}_0$  is die beramer van die toestand op tydstip nul voor enige data verkry is. Vir 'n proses met gemiddeld nul, is dit 'n vektor van nulle:

$$\begin{aligned} \mathbf{a}_0 &= \mathbf{a}_{1/0} \\ &= (0, \dots, 0)' \end{aligned}$$

### Die Aanvangskovariansiematriks

Met  $t=1$  herlei 3.3.6 na

$$P_{1/0} = P_0 = TP_0T' + RQR'$$

Die algoritme van Gardner, Harvey en Phillips(1980) maak gebruik van die volgende oplossing vir bogenoemde vergelyking:

$$\text{vektor}(P_0) = [I - TOT']^{-1} \text{vektor}(RQR')$$

Daar word met hierdie resultaat volstaan. Aan die hand van Jones (1980) word die volgende afleiding vir die aanvangskovariansiematriks gegee:

Die  $m \times m$  matriks  $P_0$  is die onvoorwaardelike kovariansiematriks van

$\alpha_t = (y_t, y_{t+1/t}, \dots, y_{t+m-1/t})$  en kan geskryf word as:

$$P_0 = \begin{bmatrix} E[y_t y_t] & E[y_t y_{t+1/t}] & \dots & E[y_t y_{t+j/t}] & \dots & E[y_t y_{t+m-1/t}] \\ E[y_{t+1/t} y_t] & E[y_{t+1/t} y_{t+1/t}] & & & & E[y_{t+1/t} y_{t+m-1/t}] \\ & & & & & \\ E[y_{t+i/t} y_t] & & & E[y_{t+i/t} y_{t+j/t}] & & \dots \\ & & & & & \\ E[y_{t+m-1/t} y_t] & & & & & E[y_{t+m-1/t} y_{t+m-1/t}] \end{bmatrix}$$

Daar word dus gesoek vir 'n algemene uitdrukking vir die "0,j"-de element naamlik

$$E[y_t y_{t+j/t}]$$

en die "i,j"-de element naamlik

$$E[y_{t+i/t} y_{t+j/t}]$$

Bereken eers die kovariansies tussen die proses en die foutterme,  $\gamma_k$ .

Hier moet telkens onthou word dat die foutterme ongekorreleerd is met die verlede van die proses, of anders gestel:  $y_t$  hang slegs af van die skokterme tot en met tydstep  $t-k$ .

Dus volg dat

$$E[y_{t-k}\eta_t] = 0 \text{ vir } k > 0$$

en

$$E[y_{t-k}\eta_t] \neq 0 \text{ vir } k \leq 0$$

Verder volg dat

$$\begin{aligned} \gamma_0 &= E[y_t\eta_t] \\ &= \sigma^2 \end{aligned}$$

$$\begin{aligned} \gamma_k &= E\left[\left(\sum_{j=1}^p \phi_j y_{t+k-j} + \sum_{j=0}^q \theta_j \eta_{t+k-j}\right)\eta_t\right] \\ &= \sum_{j=1}^p \phi_j E[y_{t+k-j}\eta_t] + \sum_{j=0}^q \theta_j E[\eta_{t+k-j}\eta_t] \\ &= \sigma^2 \theta_k + \sum_{j=1}^p \phi_j \gamma_{k-j}, \quad k > 0 \end{aligned} \tag{3.5.1}$$

met

$$\gamma_k = 0 \text{ as } k < 0 \tag{3.5.2}$$

Wanneer  $\gamma_k$  genormaliseer word deur met  $\sigma^2$  te deel, is daar 'n verband tussen  $\gamma_k$  en die g's van 2.5.9 naamlik

$$\xi_k = \gamma_{k-1}$$

Aangesien enige ARMA-model geskryf kan word as 'n bewegende-gemiddelde(MA)-model, met 'n oneindige aantal terme (kyk byvoorbeeld Box en Jenkins(1976)), volg dat

$$y_t = \sum_{k=0}^{\infty} \gamma_k \eta_{t-k}, \quad 3.5.3$$

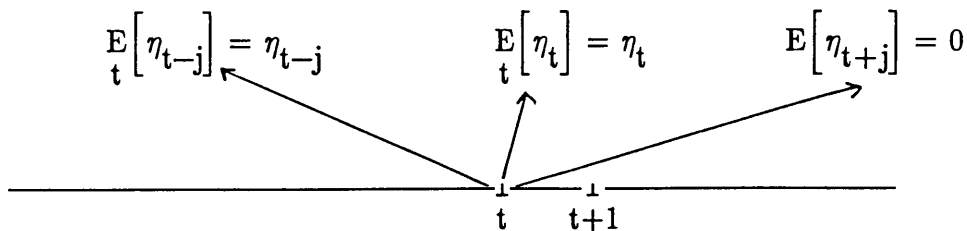
met

$$\gamma_0 = 1$$

Vir  $j \geq 0$ , kan die elemente van die toestandsvektor as volg voorgestel word

$$\begin{aligned} y_{t+j/t} &= E_t[y_{t+j}] \\ &= E_t \left[ \sum_{k=0}^{\infty} \gamma_k \eta_{t+j-k} \right] \\ &= \sum_{k=0}^{\infty} \gamma_k E_t[\eta_{t+j-k}] \end{aligned}$$

Die volgende simboliese voorstelling gee 'n opsomming van die beperkings op die foutterme.



Dus

$$y_{t+j/t} = \sum_{k=0}^{\infty} \gamma_{j+k} \eta_{t-k} \quad 3.5.4$$

Verder is, vir  $j > 0$  (uit 3.5.3)

$$\begin{aligned}
 y_{t+j} &= \sum_{k=0}^{\infty} \gamma_k \eta_{t+j-k} \\
 &= \eta_{t+j} + \gamma_1 \eta_{t+j-1} + \gamma_2 \eta_{t+j-2} + \dots + \gamma_{j-1} \eta_{t+1} \\
 &\quad + \gamma_j \eta_t + \gamma_{j+1} \eta_{t-1} + \dots \\
 &= \sum_{k=-j}^{-1} \gamma_{j+k} \eta_{t-k} + y_{t+j/t}
 \end{aligned} \tag{3.5.5}$$

Vir  $j \geq 0$ , en aangesien die  $\eta$ 's ongekorreleerd is met die  $y$ 's in die verlede, volg dat

$$\begin{aligned}
 E[y_t y_{t+j}] &= E\left[y_t \left( \sum_{k=-j}^{-1} \gamma_{j+k} \eta_{t-k} + y_{t+j/t} \right)\right] \\
 &= E\left[ \sum_{k=-j}^{-1} \gamma_{j+k} y_t \eta_{t-k} + y_t y_{t+j/t} \right] \\
 &= \sum_{k=-j}^{-1} \gamma_{j+k} E[y_t \eta_{t-k}] + E[y_t/t y_{t+j/t}]
 \end{aligned}$$

Die eerste term is nul,

dus

$$E[y_t y_{t+j}] = E[y_{t/t} y_{t+j/t}] \tag{3.5.6}$$

Verder, vir  $j \geq i > 0$ , volg dat

$$\begin{aligned}
 E[y_{t+i}y_{t+j}] &= E\left[\left(\sum_{k=-i}^{-1} \gamma_{i+k}\eta_{t-k} + y_{t+i/t}\right)\left(\sum_{m=-j}^{-1} \gamma_{j+m}\eta_{t-m} + y_{t+j/t}\right)\right] \\
 &= E\left[\sum_{k=-i}^{-1} \gamma_{i+k}\eta_{t-k} \sum_{m=-j}^{-1} \gamma_{j+m}\eta_{t-m} \right. \\
 &\quad \left. + y_{t+j/t} \sum_{k=-i}^{-1} \gamma_{i+k}\eta_{t-k} + y_{t+i/t} \sum_{m=-j}^{-1} \gamma_{j+m}\eta_{t-m} + y_{t+i/t}y_{t+j/t}\right] \\
 &= E\left[\gamma_0\gamma_{j-i}\eta_{t-i}\eta_{t-i} + \gamma_1\gamma_{1+j-i}\eta_{t+i-1}\eta_{t+i-1} + \right. \\
 &\quad \left. \dots + \gamma_{i-1}\gamma_{j-1}\eta_{t+1}\eta_{t+1}\right] + \sum_{k=-i}^{-1} \gamma_{i+k}E\left[y_{t+j/t}\eta_{t-k}\right] \\
 &\quad + \sum_{m=-j}^{-1} \gamma_{j+m}E\left[y_{t+i/t}\eta_{t-m}\right] + E\left[y_{t+i/t}y_{t+j/t}\right] \tag{3.5.6}
 \end{aligned}$$

Vervang 3.5.4 in 3.5.6:

$$\begin{aligned}
 &= \sigma^2 \sum_{k=0}^{i-1} \gamma_k\gamma_{k+j-i} + \sum_{k=-i}^{-1} \gamma_{i+k}E\left[\left[\sum_{n=0}^{\infty} \gamma_{j+n}\eta_{t-n}\right]\eta_{t-k}\right] \\
 &\quad + \sum_{m=-j}^{-1} \gamma_{j+m}E\left[\left[\sum_{p=0}^{\infty} \gamma_{j+p}\eta_{t-p}\right]\eta_{t-m}\right] + E\left[y_{t+i/t}y_{t+j/t}\right] \tag{3.5.7}
 \end{aligned}$$

Aangesien die kovariansies tussen die foutterme nul is, verval die tweede en derde terme en reduseer 3.5.7 na:

$$E\left[y_{t+i}y_{t+j}\right] = \sigma^2 \sum_{k=0}^{i-1} \gamma_k\gamma_{k+j-i} + E\left[y_{t+i/t}y_{t+j/t}\right]$$

Dus

$$E\left[y_{t+i}/t y_{t+j}/t\right] = \sigma^2 \sum_{k=0}^{i-1} \gamma_k \gamma_{k+j-i} - E\left[y_{t+i} y_{t+j}\right] \quad 3.5.8$$

Vergelykings 3.5.6 en 3.5.8 is die benodigde uitdrukkings, en moet nou met behulp van rekursies opgelos word.

Stel

$$\begin{aligned} C_k &= E\left[y_{t+k} y_t\right] \\ &= E\left[\left(\sum_{j=1}^p \phi_j y_{t+k-j} + \sum_{j=0}^q \theta_j \eta_{t+k-j}\right) y_t\right] \\ &= \sum_{j=1}^p \phi_j E\left[y_{t+k-j} y_t\right] + \sum_{j=0}^q \theta_j E\left[y_t \eta_{t+k-j}\right] \end{aligned}$$

Vanaf 0  $\rightarrow k-1$  is  $E\left[y_t \eta_{t+k-j}\right] = 0$

$$= \sum_{j=1}^p \phi_j C_{k-j} + \sum_{j=k}^q \theta_j \gamma_{j-k} \quad 3.5.9$$

Vanuit 3.5.1 kan die genormaliseerde kovariansies rekursief bereken word:

$$\begin{aligned} \gamma_0 &= 1 \\ \gamma_1 &= \phi_1 \gamma_0 + \theta_1 \\ \gamma_2 &= \phi_1 \gamma_1 + \phi_2 \gamma_0 + \theta_2 \\ \gamma_3 &= \phi_1 \gamma_2 + \phi_2 \gamma_1 + \phi_3 \gamma_0 + \theta_3 \\ &\vdots \end{aligned}$$

Vanuit 3.5.9 kan die kovariansies bereken word deur die volgende stelsel lineêre vergelykings op te los:

$$\begin{aligned}
 C_0 &= \phi_1 C_{-1} + \phi_2 C_{-2} + \dots + \phi_p C_p + \gamma_0 + \theta_1 \gamma_1 \\
 &\quad + \theta_2 \gamma_2 + \dots + \theta_q \gamma_q \\
 C_1 &= \phi_1 C_0 + \phi_2 C_{-1} + \dots + \phi_p C_{1-p} + \theta_1 \gamma_0 + \theta_2 \gamma_1 \\
 &\quad + \dots + \theta_q \gamma_{q-1} \\
 &\quad \vdots \\
 C_p &= \phi_1 C_{p-1} + \phi_2 C_{p-2} + \dots + \phi_p C_0 + \theta_p \gamma_0 \\
 &\quad + \theta_{p+1} \gamma_1 + \dots + \theta_q \gamma_q
 \end{aligned}$$

Deur manipulering van die vergelykings kan 'n stelsel van  $p-1$  lineêre vergelykings verkry word waaruit  $C_0, C_1, \dots, C_p$  in terme van  $\phi$ 's,  $\theta$ 's en  $\gamma$ 's opgelos kan word.

Indien  $q > p$  word bloot die bykomstige waardes van  $C_{p+1}$  tot  $C_q$  bereken.

Die "0,j"-de element van die aanvangsmatriks naamlik  $E\left[y_t/t y_{t+1}/t\right]$  word in hierdie geval gegee deur:

$$E\left[y_t y_{t+j}\right] = C_j$$

en die "i,j"-de element deur

$$\begin{aligned}
 E\left[y_{t+i}/t y_{t+j}/t\right] &= E\left[y_t y_{t+(j-i)}\right] \\
 &= C(j-i) - \sum_{k=0}^{i-1} \gamma_k \gamma_{k+j-i}
 \end{aligned}$$

### 3.6 ILLUSTRASIE VAN DIE VERSKILLENDE TEGNIEKE AAN DIE HAND VAN 'N MA(1)-PROSES

Vervolgens word die gebruik van die resultate, aangegee in die voorafgaande afdelings, geïllustreer deur te aanvaar dat die waarnemings deur 'n MA(1)-proses:

$$y_t = \eta_t + \theta\eta_{t-1},$$

waarby  $\eta_t \sim (0, \sigma^2)$ , gegengereer word.

Die toestandsruimte vorm word gegee deur:

$$\begin{aligned}\alpha_{t+1} &= (y_{t+1}, y_{t+2}/t+1)' \\ &= \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \alpha_t + \begin{bmatrix} 1 \\ \theta \end{bmatrix} \eta_t\end{aligned}$$

met

$$T = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, R = \begin{bmatrix} 1 \\ \theta \end{bmatrix}, Q = 1, \alpha_t = (y_t, \theta\eta_t)'$$

Die aanvangstoestandsvektor is:

$$\mathbf{a}_0 = \mathbf{a}_{1/0} = (0 \ 0)'$$

Die aanvangsmatriks  $P_0$  word vervolgens op verskillende maniere bereken:

1. Direk:

$$\begin{aligned}P_0 &= \sigma^{-2} E[\alpha_t \alpha_t'] \\ &= \sigma^{-2} E \begin{bmatrix} y_t \\ \theta\eta_t \end{bmatrix} [y_t, \theta\eta_t] \\ &= \sigma^{-2} \begin{bmatrix} \text{var}(y_t) & \text{okov}(y_t \eta_t) \\ \text{okov}(y_t \eta_t) & \theta^2 \text{var}(\eta_t) \end{bmatrix}\end{aligned}$$

$$= \begin{bmatrix} 1 + \theta^2 & \theta \\ \theta & \theta^2 \end{bmatrix}$$

2. Met behulp van Jones(1980) se afleiding:

$$\begin{aligned} \gamma_0 &= 1 \\ \gamma_1 &= \theta \\ C_0 &= \gamma_0 + \theta\gamma_1 \\ &= 1 + \theta^2 \\ C_1 &= \theta\gamma_0 \\ &= \theta \\ P_{00} &= C_0 = 1 + \theta^2 \\ P_{01} &= C_1 = \theta \\ P_{10} &= C_1 = \theta \\ P_{11} &= C_0 - \gamma_0\gamma_0 \\ &= 1 + \theta^2 - 1 \\ &= \theta^2 \end{aligned}$$

3. Met behulp van vektor( $P_0$ ) =  $[I - T \otimes T]^{-1}$  vektor( $RQR'$ ):

met

$$TOT' = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \otimes \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$I - T \otimes T = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \text{ sodat } (I - T \otimes T')^{-1} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$RQR' = RR' = \begin{bmatrix} 1 \\ \theta \end{bmatrix} [1 \ 0] = \begin{bmatrix} 1 & \theta \\ \theta & \theta^2 \end{bmatrix}$$

$$\begin{aligned} \text{vektor}(P_0) &= \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ \theta \\ \theta \\ \theta^2 \end{bmatrix} \\ &= \begin{bmatrix} 1 + \theta^2 \\ \theta \\ \theta \\ \theta^2 \end{bmatrix} \end{aligned}$$

Deur gebruikmaking van die algemene uitdrukking (3.3.7)

$$v_t = y_t - \bar{y}_{t/t-1}$$

volg dat die voorspellingsfout op tydstep  $t=1$  gegee word deur

$$\begin{aligned} v_1 &= y_1 - \bar{y}_{1/0} \\ &= y_1 \end{aligned}$$

aangesien

$$\bar{y}_{1/0} = z_1' a_{1/0} = 0$$

Met behulp van 3.3.8 en  $t = 1$  volg dat

$$\begin{aligned} f_1 &= z_1' P_{1/0} z_1 + 0 \\ &= [ \ 1 \ 0 \ ] \begin{bmatrix} 1 + \theta^2 & \theta \\ \theta^2 & \theta^2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ &= 1 + \theta^2 \end{aligned}$$

Toepassing van die opdateringsformule (3.3.11) gee:

$$\begin{aligned}
 \mathbf{a}_1 &= \mathbf{a}_{1/0} + \mathbf{P}_{1/0} \mathbf{z}'_1 (y_1 - \mathbf{z}'_1 \mathbf{a}_{1/0}) / f_1 \\
 &= 0 + \begin{bmatrix} 1+\theta^2 & \theta \\ \theta & \theta^2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} [y_1 - 0] / (1 + \theta^2) \\
 &= \begin{bmatrix} 1 + \theta^2 \\ \theta \end{bmatrix} y_1 / (1 + \theta^2) \\
 &= \begin{bmatrix} y_1 \\ \theta y_1 \\ \hline 1 + \theta^2 \end{bmatrix}
 \end{aligned}$$

Verder volg uit 3.3.10 dat

$$\begin{aligned}
 \mathbf{P}_1 &= \mathbf{P}_{1/0} - \mathbf{P}_{1/0} \mathbf{z}'_t \mathbf{z}'_t \mathbf{P}_{1/0} / f_1 \\
 &= \begin{bmatrix} 1+\theta^2 & \theta \\ \theta & \theta^2 \end{bmatrix} - \begin{bmatrix} 1+\theta^2 & \theta \\ \theta & \theta^2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1+\theta^2 & \theta \\ \theta & \theta^2 \end{bmatrix} / (1+\theta^2) \\
 &= \begin{bmatrix} 1+\theta^2 & \theta \\ \theta & \theta^2 \end{bmatrix} - \begin{bmatrix} (1+\theta^2)^2 & \theta(1+\theta^2) \\ \theta(1+\theta^2) & \theta^2 \end{bmatrix} / (1+\theta^2) \\
 &= \begin{bmatrix} 0 & 0 \\ 0 & \theta^4 / (1+\theta^2) \end{bmatrix}
 \end{aligned}$$

Die voorspellingsvergeljking vir  $\alpha_2$ , uit 3.3.5 en  $t = 2$ , word gegee deur:

$$\begin{aligned} \mathbf{a}_{2/1} &= \mathbf{T}\mathbf{a}_1 \\ &= \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ \theta y_1 / (1 + \theta^2) \end{bmatrix} \\ &= \begin{bmatrix} \theta y_1 / (1 + \theta^2) \\ 0 \end{bmatrix} \end{aligned}$$

Met behulp van 3.3.6 en met  $t = 2$  volg dat:

$$\begin{aligned} \mathbf{P}_{2/1} &= \mathbf{T}\mathbf{P}_1\mathbf{T}' + \mathbf{R}\mathbf{R}' \\ &= \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & \theta^4 / (1 + \theta^2) \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} + \begin{bmatrix} 1 & \\ \theta & \end{bmatrix} \begin{bmatrix} 1 & 0 \end{bmatrix} \\ &= \begin{bmatrix} \theta^4 / (1 + \theta^2) & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 1 & \theta \\ \theta & \theta^2 \end{bmatrix} \\ &= \begin{bmatrix} (1 + \theta^2 + \theta^4) / (1 + \theta^2) & \theta \\ \theta & \theta^2 \end{bmatrix} \end{aligned}$$

Soos voorheen word uitdrukkings vir  $v_2$  en  $f_2$  uit 3.3.7 en 3.3.8 afgelei naamlik:

$$\begin{aligned} v_2 &= y_2 - \bar{y}_{2/1} \\ &= y_2 - \mathbf{z}'_t \mathbf{a}_{2/1} \\ &= y_2 - \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \theta y_1 (1 - \theta^2) \\ 0 \end{bmatrix} \\ &= y_2 - \theta y_1 / (1 + \theta^2) \end{aligned}$$

en

$$\begin{aligned}
 f_2 &= \mathbf{z}'_t \mathbf{P}_{2/1} \mathbf{z}_t \\
 &= [1 \ 0] \begin{bmatrix} (1+\theta^2+\theta^4)/(1+\theta^2) & \theta \\ \theta & \theta^2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\
 &= (1+\theta^2+\theta^4)/(1+\theta^2) \\
 &= 1 + \theta^4/(1+\theta^2)
 \end{aligned}$$

Wanneer die proses herhaal word, is dit duidelik dat die Kalmanfilter die voorspellingsfoute bereken vanuit die volgende rekursies:

$$v_t = y_t - \theta v_{t-1} / f_t, \quad t = 1, \dots, T$$

met

$$v_0 = 0$$

en

$$f_t = 1 + \theta^{2t} / (1 + \theta^2 + \dots + \theta^{2(t-1)})$$

### 3.7 SAMEVATTING

Die voorafgaande afdeling se werk word saamgevat aan die hand van 'n ARMA(p,q)-proses

Veronderstel 'n ARMA(p,q)-model van die vorm

$$y_t = \sum_{k=1}^p \phi_k y_{t-k} + \eta_t + \sum_{k=1}^q \theta_k \eta_{t-k} \quad t=1, \dots, T$$

waarby  $\eta_t \sim N(0, \sigma^2)$  verdeel is, word beskou en die parameters  $\phi$  en  $\theta$  moet beraam word.

Die outoregressiewe-bewegendegemiddelde proses word vervolgens eers in die toestandruimte vorm

$$\alpha_t = T \alpha_{t-1} + R \eta_t$$

geskryf (Afdeling 2.5) waarby die metingsvergelyking gegee word deur

$$y_t = z_t' \alpha_t + \xi_t, t = 1, \dots, T$$

Wanneer daar veronderstel word dat die ruisterme,  $\xi_t$  en  $\eta_t$  normaal verdeel is, met  $\alpha_0 \sim N(\mathbf{a}_0, \sigma^2 P_0)$ ,  $\mathbf{a}_0$  en  $P_0$  bekend, het die  $T \times 1$  vektor van waarnemings  $\mathbf{y}$  'n meerveranderlike normaalverdeling.

Met die kennis gegee in Afdeling 3.5, kan die aanvangswaardes vir die beramer  $\mathbf{a}_0$  van  $\alpha_0$ , en die gepaardgaande  $P_0$  bereken word en kan met die Kalmanrekursies begin word.

Soos in Afdeling 3.3 uiteengesit, lewer die sistematiese toepassing van die Kalmanfilter die MGKB van  $y_t$ , gegee die vorige waarnemings (vir  $t = 1, \dots, T$ ). Vir elke waarneming wat bekend word,  $y_t$ , word daar 'n gepaardgaande foutterm  $v_t$  gegenereer met sy geassosieerde variansie,  $\sigma^2 f_t$ .

Die herhaaldelike toepassing van die Kalmanfilter lewer dus 'n stel foutterme. Die terme kan nou direk in die voorspellings–fout–ontbinding van die aanneemlikheidsfunksie (2.4.5) vervang word. Vervolgens kan beramers vir  $\phi$  en  $\theta$  bereken word deur (2.4.5) te maksimaliseer.

Die kennis word in Hoofstuk 4 toegepas waar die Kalmanfilter gebruik word om tydreeks met verlore data te hanteer.

## HOOFSTUK 4

### HANTERING VAN TYDREEKSE MET VERLORE WAARNEMINGS MET BEHULP VAN DIE TOESTANDRUIMTE BENADERING EN DIE KALMANFILTER.

#### 4.1 INLEIDING

Soos reeds in Hoofstuk 1 uiteengesit word daar in die praktyk baie te doen gekry met verlore data. Statistisi word veral deur navorsers in omgewingsake genader om datastelle te analiseer. Vir tydreeksanalise gaan die data gewoonlik bevredigend ver terug in die verlede maar die groot kopseer is die onreëlmatigheid van die data. Die onreëlmatigheid kan te wyte wees aan verskeie redes soos wisseling van personeel, die breek van meettoerusting of die onbereikbaarheid van meetaparaat weens weersomstandighede.

Daar is verskeie metodes om die probleem te benader soos byvoorbeeld vooruit- en terugberaming, outoregressie, latfunksies en nog meer. In Hoofstuk 5 word die metodes verder toegelig wanneer daar 'n vergelyk getref word tussen die verskillende metodes en die resultate verkry deur gebruik te maak van die toestandruimte benadering soos uiteengesit in Afdeling 4.2.

In Afdeling 4.3 word gekyk na die algoritme van Gardner, Harvey en Phillips(1980) wat gewysig is om verlore data te kan hanteer terwyl Afdeling 4.4 wys op die tekortkomings van die metode na gelang die gaping, wat veroorsaak is deur die vermiste data, te groot raak.

## 4.2 Verlore data en die Kalmanfilter

In Hoofstuk 3 is aangetoon dat die Kalmanfilter-vergelykings vir die toestandruimte model

$$\alpha_t = T\alpha_{t-1} + R\eta_t$$

verdeel word in voorspellings- en opdateringsvergelykings.

Wanneer 'n reeks verlore data het, doen die rekursie wel die voorspellingsvergelykings naamlik:

$$3.3.5 \quad \mathbf{a}_{t/t-1} = T\mathbf{a}_{t-1}$$

$$3.3.6 \quad \mathbf{P}_{t/t-1} = T\mathbf{P}_{t-1}T' + RQR'$$

Aangesien die werklike waardes ontbreek, kan die geskatte waarde nie "opgedateer" word nie en is daar nie sprake van 'n voorspellingsfout nie. Die opdateringsvergelykings:

$$3.3.11 \quad \mathbf{P}_t = \mathbf{P}_{t/t-1} - \mathbf{P}_{t/t-1}\mathbf{z}_t\mathbf{z}_t'\mathbf{P}_{t/t-1}/f_t$$

en 
$$3.3.13 \quad \mathbf{a}_t = \mathbf{a}_{t/t-1} + \mathbf{P}_{t/t-1}\mathbf{z}_t(y_t - \mathbf{z}_t'\mathbf{a}_{t/t-1})/f_t$$

reduseer dan na

$$\mathbf{P}_t = \mathbf{P}_{t/t-1}$$

en

$$\mathbf{a}_t = \mathbf{a}_{t/t-1}$$

Beskou byvoorbeeld 'n reeks wat volledig is tot op tydstip  $t-1$ , 'n ontbrekende waarde op tydstip  $t$  het en wat weer hervat word op tydstip  $t+1$ .

Dan het ons

$$\mathbf{a}_{t/t-1} = T\mathbf{a}_{t-1}$$

Met  $y_t$  nie beskikbaar nie, word daar oorgegaan na die twee-stap-vooruitskatting:

$$\begin{aligned} \mathbf{a}_{t+1} &= \mathbf{a}_{t+1/t} \\ &= T\mathbf{a}_t \end{aligned}$$

$$\begin{aligned} \mathbf{a}_{t+1} &= T\mathbf{a}_{t/t-1} \\ &= T(T\mathbf{a}_{t-1}) \end{aligned}$$

In die algemeen, met  $m$  waarnemings verlore word

$$3.3.5 \quad \mathbf{a}_{t+m/t} = T\mathbf{a}_{t+m-1/t}$$

$$\text{en} \quad 3.3.6 \quad P_{t+m/t} = TP_{t+m-1/t}T' + RQR'$$

Die skatting van die model met behulp van die voorspellings–fout–ontbinding van die aanneemlikheidsfunksie word nie veel beïnvloed nie in die sin dat, vir die verlore punt, word daar bloot geen voorspellingsfout in berekening gebring nie.

#### 4.3 DIE ALGORITME VAN GARDNER, HARVEY EN PHILLIPS(1980)

Die algoritme van Gardner, Harvey en Phillips is geskryf sodat die aanneemlikheidsfunksie van 'n stasionêre ARMA–proses deur middel van die Kalmanfilter bereken kan word. Die onbekende parameters van die model word beraam deur die aanneemlikheidsfunksie te maksimeer. Die algoritme bestaan uit die subroetines STARMA, KARMA en KALFOR. STARMA word gebruik om die ARMA–proses in sy toestandruimte vorm te skryf en die aanvangswaardes vir  $\mathbf{a}$  en  $P$  te bereken, KARMA voer die Kalmanrekursies uit en KALFOR maak vooruitskattings moontlik.

Die gewysigde subroetine (wysigings is opsigtelik, "bold", uitgedruk) met verduidelikende opmerkings is verderaan in hierdie afdeling opgeneem. Die toetsprogram is geloop op 'n stel van 400 ARMA(1,1)–gegenereerde data, met  $\phi = 0.8$  en  $\theta = 0.3$ . Vervolgens is 'n gaping van 10 verlore data geskep en die data vervang met 'n verlore–waarde–kode naamlik 555.5 om te bepaal watter deel van die rekursie oorgeslaan moet word.

### Opmerking

1. Die resultate van die "aanneemlikheidsfunksie" verkry deur die Kalmanfilter en die verkry deur Akaike se inligtingskriterium (Akaike's Information Criterion), wanneer van SAS gebruik gemaak word, verskil ten opsigte van sekere konstante waardes.

Vir die toetslopie op 400 ARMA(0.8;0.3) gegengereerde waardes, lewer die Kalmanfilter 'n  $-2\log(\text{"aanneemlikheidsfunksie"})$  van 2361.6621, terwyl AIC 'n  $-2\log(\text{aanneemlikheidsfunksie})$  van 1100.15 lewer. Om die verskil uit te klaar word die volgende inligting gegee.

In Hoofstuk 2 is die voorspellings-fout-ontbinding van die aanneemlikheidsfunksie gegee as

2.4.5:

$$-2\ln L(y) = T\ln 2\pi + T\ln \sigma^2 + \sum_{t=1}^T \ln f_t + \sigma^{-2} \sum_{t=1}^T v_t^2/f_t \quad 4.3.1$$

Wanneer 4.3.1 parsieel ten opsigte van  $\sigma^2$  gedifferensieer en gelyk aan nul gestel word, word die volgende beramer van  $\sigma^2$  gevind:

$$\hat{\sigma}^2 = 1/T \sum_{t=1}^T v_t^2/f_t$$

Wanneer  $\sigma^2$  deur sy beramer vervang word reduseer 4.3.1 na

$$-2\ln L(y) = (T - T\ln T) + T\ln 2\pi + T\ln \sum_{t=1}^T v_t^2/f_t + \sum_{t=1}^T \ln f_t$$

In die uitvoer van die Kalmanfilter-program word die uitdrukking

$$L^*(y) = T\ln \sum_{t=1}^T v_t^2/f_t + \sum_{t=1}^T \ln f_t$$

aangegee as die "log-likelihood function"

Om die aanneemlikheidsfunksie te maksimaliseer word waardes van  $\phi$  en  $\theta$  gevind wat  $L^*(y)$  minimaliseer.

Wanneer die resultate van die toetslopie as volg gewysig word:

$$\begin{aligned} -2\ln L(y) &= (400 - 400\ln 400) + 400\ln 2\pi + 2361.6621 \\ &= 1100.2271 \end{aligned}$$

vergelyk dit goed met:

$$\text{AIC} = -2\ln L(y) + 2(\text{aantal parameters wat gepas is})$$

dus

$$\begin{aligned} -2\ln L(y) &= 1104.15 - 4 \\ &= 1100.15 \end{aligned}$$

2. Om rekenaartyd te bespaar is daar sogenaamde "quick recursions" in die program ingebou. Die vinnige rekursies kom neer op 'n benadering van die aanneemlikheid en geskied as volg. Wanneer 'n sekere aantal waarnemings,  $\hat{t}^*$ , deur die Kalmanfilter geprosesseer is, word die verdere waardes van  $\hat{v}_t$  beraam deur  $\hat{v}_t$  wat verkry word direk van die ARMA-vergelyking naamlik

$$v_t = y_t - \phi_1 y_{t-1} - \dots - \phi_p y_{t-p} - \theta_1 \hat{v}_{t-1} - \dots - \theta_q \hat{v}_{t-q}, \quad t = \hat{t}^* + 1,$$

$$\text{met } \hat{v}_t = \bar{v}_t, \quad t = \hat{t}^*, \dots, \hat{t}^* - q + 1$$

Die waarde van  $\hat{t}^*$  word outomaties bepaal sodra  $f_t < 1 + \delta$ , met  $\delta$  'n klein positiewe waarde,  $\hat{t}^* \leq 0.01$  of  $0.001$ . Sou  $\delta$  gelyk gestel word aan 'n negatiewe getal, word die Kalmanfilter uitgevoer vir die volledige stel waarnemings en die eksakte aanneemlikheid bereken. In die studie was dit nie nodig om van die vinnige rekursies gebruik te maak nie en is  $\delta$  deurgangs negatief gestel.

**Die Hoofprogram:**

Die program is geloop op 'n stel van 400 ARMA(1,1)-gegenereerde data, met  $\phi=0.8$  en  $\theta=0.3$ . 'n Gaping van 10 verlore data is geskep deur die data te vervang met 'n verlore-data-kode naamlik 555.5.

In die program is gebruik gemaak van die subroetines KARMA en STARMA (Gardner, Harvey en Phillips) wat gewysig is om die verlore data te kan hanteer, UMINF (IMSL-Biblioteek) en FUNCT (self verskaf).

```

DIMENSION XGUESS(10),XSCALE(10),IPARAM(7),RPARAM(7),X(10)
1  ,W(1000),EST(1000)
COMMON /ARMA/IP,IQ,IR,NP,NRBAR,N,NMISS,W,EST
EXTERNAL FUNCT,UMINF
IP=1
IQ=1
N=400
XGUESS(1)=0.8
XGUESS(2)=0.3
XSCALE(1)=1.0
XSCALE(2)=1.0
FSCALE=1.0
IPQ=IP+IQ
IPARAM(1)=0
DO 10 I=1,N
READ(5,100) W(I)
100 FORMAT(8X,F10.5)
10 CONTINUE
IR=MAXO(IP,IQ+1)
NP=IR*(IR+1)/2
NRBAR=NP*(NP-1)/2

CALL UMINF(FUNCT,IPQ,XGUESS,XSCALE,FSCALE,IPARAM,RPARAM,X,FVALUE)

```

Die subroetine UMINF is geneem uit die IMSL – wiskunde biblioteek (IMSL, Inc. MATH/LIBRARY Volume3 p802) en word gebruik om  $L^*(y)$  te minimaliseer.

Die invoerinstruksies is die volgende:

### UMINF/DUMINF (Single/Double precision)

**Purpose:** Minimize a function of N variables using a quasi-Newton method and a finite-difference gradient.

**Usage:** CALL UMINF (FCN, N, XGUESS, XSCALE, FSCALE, IPARAM, RPARAM, X, FVALUE)

#### Arguments

FCN - User-supplied SUBROUTINE to evaluate the function to be minimized. The usage is  
CALL FCN (N, X, F), where

N - Length of X. (Input)

X - The point at which the function is evaluated. (Input)

X should not be changed by FCN.

F - The computed function value at the point X.  
(Output)  
FCN must be declared EXTERNAL in the calling program.

N - Dimension of the problem. (Input)

XGUESS - Vector of length N containing an initial guess of the  
computed solution. (Input)

XSCALE - Vector of length N containing the diagonal scaling matrix  
for the variables. (Input)  
In the absence of other information, set all entries  
to 1.0.

FSCALE - Scalar containing the function scaling. (Input)  
In the absence of other information, set FSCALE to 1.0.

IPARAM - Parameter vector of length 7. (Input/Output)  
See Remarks.

RPARAM - Parameter vector of length 7. (Input/Output)  
See Remarks.

X - Vector of length N containing the computed solution.  
(Output)

FVALUE - Scalar containing the value of the function at the  
computed solution. (Output)

```

WRITE(6,500) FVALUE,(X(I),I=1,IPQ)
500 FORMAT(' FVALUE',E20.12/' X ',(5E12.5))
ITER=0
DO 20 I=1,N
IF(ABS(W(I)-555.5).LE.1.E-5) THEN
ITER=ITER+1
W(I)=EST(ITER)
ENDIF
WRITE(3,101) I,W(I)
101 FORMAT(I5,3X,F10.5)
20 CONTINUE
STOP
END

```

Die subroetine FUNCT verskaf aan UMINF die funksie  $L^*(y)$  wat geminimaliseer moet word.

```

SUBROUTINE FUNCT(IPQ,X,F)
DIMENSION PHI(5),THETA(5),A(5),P(15),V(15),THETAB(15),
1 XNEXT(15),XROW(15),RBAR(105),W(1000),RESID(1000),E(5),
2 WORK(5),X(5),EST(1000)
COMMON /ARMA/IP,IQ,IR,NP,NRBAR,N,NMISS,W,EST
PHI(1)=X(1)
THETA(1)=X(2)
IF(ABS(PHI(1)).GT.1.0.OR.ABS(THETA(1)).GT.1.0) GO TO 600
CALL STARMA(IP,IQ,IR,NP,PHI,THETA,A,P,V,THETAB,
1 XNEXT,XROW,RBAR,NRBAR,IFAUULT)
C WRITE(6,501) IP,IQ,NP,NRBAR
C 501 FORMAT(' IP= ',I5,' IQ= ',I5,' NP= ',I5,' NRBAR= ',I5)
WRITE(6,502) (PHI(I),I=1,IR),(THETA(I),I=1,IR)
1 A(1),(P(I),I=1,NP),(V(I),I=1,NP)
502 FORMAT(' PHI = ',2E20.8/ 'THETA = ',2E20.8/
1/ ' INITIAL VALUE = ',E20.8/
1/ ' INITIAL COV. = ',3E20.8/ 'RR= ',3E20.8)
C WRITE(6,503) N,(W(I),I=1,N)
C 503 FORMAT(' N = ',I5,' W = '/(10F10.5))
SUMLOG=0.0
SSQ=0.0
IUPD=1
DELTA=-0.1
NIT=1

```

```

CALL KARMA(IP,IQ,IR,NP,PHI,THETA,A,P,V,N,W,
1 RESID,SUMLOG,SSQ,IUPD,DELTA,E,NIT,NMISS,EST)
WRITE(6,504) (A(I),I=1,IR)
504 FORMAT(' PREDICTED VALUES = '/(10F10.5))
WRITE(6,505) NIT
505 FORMAT(' NUMBER OBS. DEALT WITH BY KALMAN = ',I3)
WRITE(6,508) (EST(ITER),ITER=1,NMISS)
508 FORMAT(' GESKATTE WAARDES = '/(4F10.5))
WRITE(6,506) SUMLOG,SSQ
506 FORMAT(' SUMLOG = ',E20.8/ ' SSQ = ',E20.8)
XLIKE=N*ALOG(SSQ)+SUMLOG
WRITE(6,507) XLIKE
507 FORMAT(' LOG-LIKELIHOOD FUNCTION = ',E20.8)
F=XLIKE
RETURN
600 F=1.E5
RETURN
END

```

Die subroetine STARMA skryf die tydreeks in sy toestandsruimtelike vorm en vind die aanvangswaardes van  $a_0$  en  $P_0$ .

```

SUBROUTINE STARMA(IP,IQ,IR,NP,PHI,THETA,A,P,V,THETAB,XNEXT,XROW,
RBAR,NRBAR,IFault)

```

Formal parameters

IP	Integer	input	: the value of p
IQ	Integer	input	: the value of q
IR	Integer	input	: the value of $r = \max(p,q+1)$
NP	Integer	input	: the value of $r(r+1)/2$
PHI	Real array(IR)	input	: the value of $\phi$ in the first p locations
THETA	Real array(IR)	output	: contains the first column of T
A	Real array(IR)	input	: the value of $\theta$ in the first q locations
P	Real array(NP)	output	: on exit contains $P_0$ , stored as a lower triangular matrix, column by column
V	Real array(NP)	output	: on exit contains $RR^*$ , stored as a lower triangular matrix, column by column
THETAB	Real array(NP)	workspace	: used to calculate P
XNEXT	Real array(NP)	workspace	: used to calculate P
XROW	Real array(NP)	workspace	: used to calculate P
RBAR	Real array(NRBAR)	workspace	: used to calculate P
NRBAR	Integer	input	: the value of $NP*(NP-1)/2$
IFault	Integer	output	: a fault indicator, equal to 1 if IP<0 2 if IQ<0 3 if IP<0 and IQ<0 4 if IP=IQ=0 5 if IR $\neq$ MAX(IP,IQ+1) 6 if NP $\neq$ IR*(IR+1)/2 7 if NRBAR $\neq$ NP*(NP-1)/2 8 if IP=1 and IQ=0 (subroutine STARMA is not appropriate for an AR(1) process) 0 otherwise

```

SUBROUTINE STARMA(IP,IQ,IR,NP,PHI,THETA,A,P,V,THETAB,
1 XNEXT,XROW,RBAR,NRBAR,IFAU)
DIMENSION PHI(IR),THETA(IR),A(IR),P(NP),V(NP),THETAB(NP),
1 XNEXT(NP),XROW(NP),RBAR(NRBAR)
IFAU=0
IF(IP.LT.0) IFAU=1
IF(IQ.LT.0) IFAU=IFAU+2
IF(IP*IP+IQ*IQ.EQ.0) IFAU=4
K=IQ+1
IF(K.LT.IP) K=IP
IF(IR.NE.K) IFAU=5
IF(NP.NE.IR*(IR+1)/2) IFAU=6
IF(NRBAR.NE.NP*(NP-1)/2) IFAU=7
IF(IR.EQ.1) IFAU=8
IF(IFAU.NE.0) RETURN
DO 10 I=2,IR
A(I)=0.0
IF(I.GT.IP) PHI(I)=0.0
V(I)=0.0
IF(I.LE.IQ+1) V(I)=THETA(I-1)
10 CONTINUE
A(1)=0.0
IF(IP.EQ.0) PHI(1)=0.0
V(1)=1.0
IND=IR
DO 20 J=2,IR
VJ=V(J)
DO 20 I=J,IR
IND=IND+1
V(IND)=V(I)*VJ
20 CONTINUE
IF(IP.EQ.0) GO TO 300

```

Die algoritme los  $P_0$  op uit die vergelyking:

$$P_0 = TP_0T' + RR'$$

V word gelyk gestel aan  $RR'$  en aangesien elke element van  $V$  'n lineêre kombinasie is van  $P_0$ , word  $P_0$  opgelos uit die vergelyking

$$\text{vektor}(V) = S * \text{vektor}(P_0), \text{ met } S = I - T \otimes T'$$

```

IR1=IR-1
IRANK=0
IFAIL=0
SSQERR=0.0
DO 40 I=1,NRBAR
40 RBAR(I)=0.0
DO 50 I=1,NP
P(I)=0.0
THETAB(I)=0.0
XNEXT(I)=0.0
50 CONTINUE
IND=0
IND1=0
NPR=NP-IR
NPR1=NPR+1
INDJ=NPR1
IND2=NPR
DO 110 J=1,IR
PHIJ=PHI(J)
XNEXT(INDJ)=0.0
INDJ=INDJ+1
INDI=NPR1+J
DO 110 I=J,IR
IND=IND+1
YNEXT=V(IND)
PHII=PHI(I)

```

```

      IF (J.EQ.IR) GO TO 100
      XNEXT(INDJ)=-PHII
      IF (I.EQ.IR) GO TO 100
      XNEXT(INDI)=XNEXT(INDI)-PHIJ
      IND1=IND1+1
      XNEXT(IND1)=-1.0
100  XNEXT(NPR1)=-PHII*PHIJ
      IND2=IND2+1
      IF (IND2.GT.NP) IND2=1
      XNEXT(IND2)=XNEXT(IND2)+1.0
      WEIGHT=1.0

      CALL INCLU2(NP,NRBAR,WEIGHT,XNEXT,XROW,YNEXT,
1  P,RBAR,THETAB,SSQERR,RECRES,IRANK,IFAIL)
      XNEXT(IND2)=0.0
      IF (I.EQ.IR) GO TO 110
      XNEXT(INDI)=0.0
      INDI=INDI+1
      XNEXT(IND1)=0.0
110  CONTINUE

      CALL REGRES(NP,NRBAR,RBAR,THETAB,P)

      IND=NPR
      DO 200 I=1,IR
      IND=IND+1
      XNEXT(I)=P(IND)
200  CONTINUE
      IND=NP
      IND1=NPR
      DO 210 I=1,NPR
      P(IND)=P(IND1)
      IND=IND-1
      IND1=IND1-1
210  CONTINUE
      DO 220 I=1,IR
220  P(I)=XNEXT(I)
      RETURN
300  INDN=NP+1
      IND=NP+1
      DO 310 I=1,IR
      DO 310 J=1,I
      IND=IND-1
      P(IND)=V(IND)
      IF (J.EQ.1) GO TO 310
      INDN=INDN-1
      P(IND)=P(IND)+P(INDN)
310  CONTINUE
      RETURN
      END

```

Vervolgens word die subroetine KARMA uitgevoer wat die Kalmanfilter in werking laat tree. Let op dat by die opdateringsvergelings word die verlore data in aanmerking geneem.

SUBROUTINE KARMA(IP,IQ,IR,NP,PHI,THETA,A,P,V,N,W,RESID,SUMLOG,SSQ,IUPD,DELTA,E,NIT)

Formal parameters

IP	Integer	input	: the value of p
IQ	Integer	input	: the value of q
IR	Integer	input	: the value of $r = \max(p, q+1)$
NP	Integer	input	: the value of $r(r+1)/2$
PHI	Real array(IR)	input	: the first column of T
THETA	Real array(IR)	input	: the value of $\theta$ in the first q locations
A	Real array(IR)	input	: contains $a_0$
		output	: contains $a_t$ , where $t=t^*$
P	Real array(NP)	input	: contains $P_0$
		output	: contains $P_t$ , where $t=t^*$
V	Real array(NP)	input	: contains $RR^*$
N	Integer	input	: n, the number of observations
W	Real array(N)	input	: the observations
RESID	Real array(N)	output	: the corresponding standardized prediction errors
SUMLOG	Real	input	: initial value of $\Sigma \log f_t$ (zero if no previous observations)
		output	: final value of $\Sigma \log f_t$
SSQ	Real	input	: initial value of $\bar{\Sigma} v_t^2$ (zero if no previous observations)
		output	: final value of $\bar{\Sigma} v_t^2$
IUPD	Integer	input	: if IUPD=1 the prediction equations are bypassed for the first observation. This is necessary when the value of $P_0$ has obtained from STARMA. In this case, $P_{1/0} = P_0$ and $a_{1/0} = a_0$ and using the prediction equations as coded in KARMA would lead to erroneous results. For values other than 1, the prediction equations are not bypassed
DELTA	Real	input	: when NIT=0 this parameter determines the level of approximation. Negative DELTA ensures that the Kalman filter is performed while $f_t \geq 1 + \delta$ , "quick recursions" being used thereafter
E	Real array	workspace	: used to store the last q standardized prediction errors

NIT	Integer	input	: when set to zero see description of DELT for the effect of NIT; for non-zero values the "quick recursions" are performed throughout, so that a conditional likelihood is obtained
		output	: number of observations dealt with by the Kalman filter, i.e. t*

```

SUBROUTINE KARMA(IP,IQ,IR,NP,PHI,THETA,A,P,V,N,W,
1 RESID,SUMLOG,SSQ,IUPD,DELTA,E,NIT,ITER,EST)
DIMENSION PHI(IR),THETA(IR),A(IR),P(NP),V(NP),W(N),
1 RESID(N),E(IR),EST(N)
IR1=IR-1
ITER=0
DO 10 I=1,IR
10 E(I)=0.0
INDE=1
IF(NIT.EQ.0) GO TO 600
DO 500 I=1,N
WNEXT=W(I)

```

### Voorspellings

```

IF(IUPD.EQ.1.AND.I.EQ.1) GO TO 300
DT=0.0
IF(IR.NE.1) DT=P(IR+1)
IF(DT.LT.DELTA) GO TO 610
A1=A(1)
IF(IR.EQ.1) GO TO 110
DO 100 J=1,IR1
100 A(J)=A(J+1)
110 A(IR)=0.0
IF(IP.EQ.0) GO TO 200
DO 120 J=1,IP
120 A(J)=A(J)+PHI(J)*A1
200 IND=0
INDN=IR
DO 210 L=1,IR
DO 210 J=L,IR
IND=IND+1
P(IND)=V(IND)
IF(J.EQ.IR) GO TO 210
INDN=INDN+1
P(IND)=P(IND)+P(INDN)
210 CONTINUE

```

### Opdatering

```

300 FT=P(1)
IF(ABS(WNEXT-555.5).LE.1.E-5) THEN
ITER=ITER+1
EST(ITER)=A(1)
ENDIF

```

met die volgende stelling word toegesien dat geen foutterm intree nie:

```

UT=0
IF(ABS(WNEXT-555.5).GT.1.E-5) UT=WNEXT-A(1)
C PRINT*, ' I ', I, WNEXT, A(1), UT
IF(IR.EQ.1) GO TO 410
IND=IR
DO 400 J=2,IR
G=P(J)/FT
A(J)=A(J)+G*UT
DO 400 L=J,IR
IND=IND+1
P(IND)=P(IND)-G*P(L)
400 CONTINUE
410 CONTINUE

```

```

      IF (ABS(WNEXT-555.5).GT.1.E-5) A(1)=WNEXT
      DO 420 L=1,IR
420  P(L)=0.0
      RESID(I)=UT/SQRT(FT)
      E(INDE)=RESID(I)
      INDE=INDE+1
      IF (INDE.GT. IQ) INDE=1
      SSQ=SSQ+UT*UT/FT
      SUMLOG=SUMLOG+ALOG(FT)
500  CONTINUE
      NIT=N
      RETURN

```

"quick recursions"

```

600  I=1
610  NIT=I-1
      DO 650 II=I,N
      ET=W(II)
      INDW=II
      IF (IP.EQ. 0)GO TO 630
      DO 620 J=1,IP
      INDW=INDW-1
      IF (INDW.LT. 1) GO TO 630
      ET=ET-PHI(J)*W(INDW)
620  CONTINUE
630  IF (IQ.EQ. 0) GO TO 645
      DO 640 J=1,IQ
      INDE=INDE-1
      IF (INDE.EQ. 0) INDE=IQ
      ET=ET-THETA(J)*E(INDE)
640  CONTINUE
645  E(INDE)=ET
      RESID(II)=ET
      SSQ=SSQ+ET*ET
      INDE=INDE+1
      IF (INDE.GT. IQ) INDE=1
650  CONTINUE
      RETURN
      END

```

## AUXILIARY ALGORITHMS

The subroutine STARMA calls the auxiliary algorithms INCLU2(Farebrother(1976)) and REGRES(Gentleman(1974)). These algorithms were originally presented as Algol 60 procedures. The following modified Fortran 66 versions of these procedures are listed:

```

      SUBROUTINE INCLU2(NP,NRBAR,WEIGHT,XNEXT,XROW,YNEXT,
1  D,RBAR,THETAB,SSQERR,RECRES,IRANK,IFAU)
      DIMENSION XNEXT(NP),XROW(NP),D(NP),RBAR(NRBAR),THETAB(NP)
      Y=YNEXT
      WT=WEIGHT
      DO 10 I=1,NP
10  XROW(I)=XNEXT(I)
      RECRES=0.0
      IFAU=1
      IF (WT.LE.0.0) RETURN
      IFAU=0
      ITHISR=0
      DO 50 I=1,NP
      IF (XROW(I).NE.0.0)GO TO 20
      ITHISR=IThisR+NP-I
      GO TO 50

```

```

20 XI=XROW(I)
   DI=D(I)
   DPI=DI+WT*XI*XI
   D(I)=DPI
   CBAR=DI/DPI
   SBAR=WT*XI/DPI
   WT=CBAR*WT
   IF (I.EQ.NP) GO TO 40
   I1=I+1

   DO 30 K=I1,NP
     ITHISR=ITHISR+1
     XK=XROW(K)
     RBTHIS=RBAR(ITHISR)
     XROW(K)=XK-XI*RBTHIS
     RBAR( ITHISR)=CBAR*RBTHIS+SBAR*XK
30 CONTINUE
40 XK=Y
   Y=XK-XI*THETAB(I)
   THETAB(I)=CBAR*THETAB(I)+SBAR*XK
   IF (DI.EQ.O.O) GO TO 100
50 CONTINUE
   SSQERR=SSQERR+WT*Y*Y
   RECRES=Y*SQRT(WT)
   RETURN
100 IRANK=IRANK+1
   RETURN
   END

SUBROUTINE REGRES(NP,NRBAR,RBAR,THETAB,BETA)
DIMENSION RBAR(NRBAR),THETAB(NP),BETA(NP)
ITHISR=NRBAR
IM=NP
DO 50 I=1,NP
BI=THETAB(IM)
IF (IM.EQ.NP) GO TO 30
I1=I-1
JM=NP
DO 10 J=1,I1
BI=BI-RBAR( ITHISR ) *BETA(JM)
ITHISR=ITHISR-1
JM=JM-1
10 CONTINUE
30 BETA( IM )=BI
IM=IM-1
50 CONTINUE
RETURN
END

```

In die studie is nie van die subroetine KALFOR gebruik gemaak nie. Die subroetine maak vooruitskattings moontlik en word vir volledigheidshalwe ingesluit.

**SUBROUTINE KALFOR(M,IP,IR,NP,PHI,A,P,V,WORK)**

Formal parameters

M	Integer	input	: the value of m, the number of steps ahead for which predictor is required
IP	Integer	input	: the value of p
IR	Integer	input	: the value of r
NP	Integer	input	: $r(r+1)/2$
PHI	Real array(IR)	input	: contains the first column of T, the transition matrix
A	Real array(IR)	input	: current value of $a_t$
		output	: predicted value of $a_{t+m}$

P	Real array(NP)	input	: current value of $P_t$ , stored in lower triangular form, column by column
		output	: predicted value of $P_{t+m}$
V	Real array(NP)	input	: contains $RR'$ stored in lower triangular form, column by column
WORK	Real array(IR)	workspace	

```

SUBROUTINE KALFOR(M,IP,IR,NP,PHI,A,P,V,WORK)
DIMENSION PHI(IR),A(IR),P(NP),V(NP),WORK(IR)
IR1=IR-1
DO 300 L=1,M
A1=A(1)
IF (IR .EQ. 1)GO TO 110
DO 100 I=1,IR1
100 A(I)=A(I+1)
110 A(IR)=0.0
IF(IP.EQ.0) GO TO 200
DO 120 J=1,IP
120 A(J)=A(J)+PHI(J)*A1
200 DO 210 I=1,IR
210 WORK(I)=P(I)
IND=0
IND1=IR
DT=P(1)
DO 220 J=1,IR
PHIJ=PHI(J)
PHIJDT=PHIJ*DT
DO 220 I=J,IR
IND=IND+1
PHII=PHI(I)
P(IND)=V(IND)+PHII*PHIJDT
IF (J.LT.IR) P(IND)=P(IND)+WORK(J+1)*PHII
IF (I.EQ.IR) GO TO 220
IND1=IND+1
P(IND)=P(IND)+WORK(I+1)*PHIJ+P(IND1)
220 CONTINUE
300 CONTINUE
RETURN
END

```

### Opmerking

Namate die gaping van verlore data vergroot, verflou die inligting (omtrent die verlede) wat in die toestandsvektor vervat is. Die invloed word waargeneem in die vektor  $a_{t+j/t}$  wat neig na nul soos  $j \rightarrow \infty$  en  $P_{t+j/t} \rightarrow P_0$ , die aanvangskovariansiematriks. Wanneer die rekursie dus oor 'n groot blok verlore data uitgevoer word, is dit ekwivalent aan 'n rekursie wat opnuut, na die gaping, begin word.

Ter illustrasie word gekyk na die volgende voorbeeld:

Beskou 'n outoregressiewe model van orde een [AR(1)] vir 'n tydreeks:

$$y_t = \phi y_{t-1} + \eta_t$$

$$\begin{aligned} 1. \quad \hat{y}_{t+1} &= y_{t+1/t} \\ &= E_t[y_{t+1}] \\ &= E_t[\phi y_t + \eta_{t+1}] \\ &= \phi y_t \end{aligned}$$

Die enigste inligting wat nodig en voldoende is om die reeks arbitrêr ver in die toekoms te voorspel, is bloot  $y_t$  en per definisie is  $\alpha_t$  dus gelyk aan  $y_t$ .

Die toestandruimte vorm word gegee deur:

$$\alpha_{t+1} = y_{t+1} = \phi \alpha_t + \eta_t$$

$$\text{Dus} \quad T = \phi$$

$$R = 1$$

$$Q = 1$$

2. Met een verlore waarneming, sê  $\hat{y}_t$ , is die twee-stap-vooruitskattings:

$$\begin{aligned} a_{t+1} &= a_{t+1/t} \\ &= T a_t \\ &= \phi a_t \\ &= \phi a_{t/t+1} \\ &= \phi[\phi a_{t-1}] \\ &= \phi^2 a_{t-1} \end{aligned}$$

Met  $j-1$  verlore waarnemings, is die  $j$ -stap-vooruitskating:

$$\begin{aligned} a_{t+j-1} &= \phi^j a_{t-1} \\ &\longrightarrow 0 \text{ soos } j \text{ groot word} \end{aligned}$$

3. Die een-stap-vooruitskatting word gegee deur:

$$\hat{y}_{t/t-1} = a_{t/t-1} = \phi a_{t-1} = \phi y_{t-1}$$

Die een-stap-beramingsfout is:

$$y_t - \phi y_{t-1}$$

Gevolgtik is:

$$\begin{aligned} P_{t/t-1} &\stackrel{\text{def}}{=} E[y_t - \phi y_{t-1}]^2 \\ &= E[y_t^2] - 2\phi E[y_t y_{t-1}] + \phi^2 E[y_{t-1}^2] \\ &= \frac{1}{1-\phi^2} - 2\phi \left[ \frac{\phi}{1-\phi^2} \right] + \frac{\phi^2}{1-\phi^2} \dots (\text{Box en Jenkins}) \\ &\qquad\qquad\qquad \text{p56-57} \\ &= 1 \end{aligned}$$

$$\begin{aligned} P_{t+1/t} &= TP_t T' + RQR' \\ &= \phi P_{t/t-1} \phi + 1 \\ &= \phi^2 + 1 \end{aligned}$$

Netso is

$$\begin{aligned} P_{t+2/t} &= \phi^2(\phi^2 + 1) + 1 \\ &= [\phi^2]^2 + [\phi^2] + 1 \\ &\vdots \\ P_{t+j/t} &= [\phi^2]^j + [\phi^2]^{j-1} + \dots + [\phi^2] + 1 \end{aligned}$$

$$\frac{j}{\omega} \frac{1}{1-\phi^2}, \text{ wat die gewone variansie van die proses is.}$$

**Gevolgtrekking:**

Wanneer  $j$  groot is (die gaping van verlore data dus groot), is die waarde na die gaping te ver voor die waarde voor die gaping, om enige sinvolle inligting te kan verskaf. Die nadeel is egter nie beperk tot net die metode nie, maar is algemeen geldig.

**4.4 SAMEVATTING**

In hierdie hoofstuk is hoofsaaklik gekyk na hoe die Kalmanfilter gebruik kan word om verlore data te kan hanteer. Dit word gedoen aan die hand van die gewysigde algoritme van Gardner, Harvey en Phillips (1980).

## HOOFSTUK 5

### 'N KORT OORSIG OOR ANDER METODEDES WAT VERLORE DATA HANTEER

#### 5.1 INLEIDING

Die onderstaande tabel is 'n uittreksel van 'n datastel en dien as 'n voorbeeld hoe die data, wat aan die Statistikus voorgele word vir analise, daar uitsien. In hierdie geval moet daar diskriminantanalise uitgevoer word wat vereis dat die data stasioner, struktuurvry en Normaal verdeel moet wees. As eerste stap moet die uiters onreëlmatig gepasieerde data sodanig verwerk word dat met die analise voortgegaan kan word. Daar is dan ook verskeie tegnieke waaruit gekies kan word alhoewel sommige rekenaarpakette (byvoorbeeld SAS se nuutste weergawe) reeds al voorsiening maak vir tydreëse met verlore data.

DATE SAMPLED	TIME SAM- PLED	GAUGE PLATE (M)	* * * * *	EC	*TOTAL* *DSLVD* *SALTS*	PH	NA	MG	CA	F	CL	*NO3+ *NO2 *AS N	SO4	PO4	TAL	SI
79-05-07	* 14H05*	0.420	*0*U*	88.0*	*	*	*	*	*	*	*	*	*	*	*	*
79-05-14	* 14H10*	0.358	*0*U*	76.4*	533*	7.41*	59.8*	26.2*	52.7*	0.70*	54.0*	11.75*	126.4*	1.86*	120.7*	7.69*
79-05-21	* 13H55*	0.360	*0*U*	73.7*	*	*	*	*	*	*	*	*	*	*	*	*
79-05-28	* 12H35*	0.370	*0*U*	71.0*	*	*	*	*	*	*	*	*	*	*	*	*
79-06-04	* 14H25*	0.360	*0*U*	76.3*	*	*	*	*	*	*	*	*	*	*	*	*
79-06-11	* 14H15*	0.365	*0*U*	77.2*	539*	7.49*	61.3*	27.5*	53.3*	0.52*	58.6*	11.30*	116.3*	2.11*	128.1*	0.00*
79-06-18	* 14H05*	0.370	*0*U*	78.1*	*	*	*	*	*	*	*	*	*	*	*	*
79-06-25	* 14H20*	0.365	*0*U*	79.2*	557*	7.44*	63.1*	26.6*	56.1*	0.54*	57.9*	13.36*	131.1*	2.04*	121.3*	5.08*
79-07-02	* 14H15*	0.365	*0*U*	77.3*	*	*	*	*	*	*	*	*	*	*	*	*
79-07-04	* 11H00*	0.360	*0*U*	80.8*	560*	8.34*	63.6*	27.6*	57.0*	0.69*	59.5*	12.24*	134.7*	1.39*	123.1*	4.39*
79-07-16	* 14H05*	0.370	*0*U*	77.4*	*	*	*	*	*	*	*	*	*	*	*	*
79-07-30	* 13H40*	0.370	*0*U*	81.4*	*	*	*	*	*	*	*	*	*	*	*	*
79-08-04	* 09H30*	0.360	*0*U*	88.9*	595*	7.74*	72.1*	27.3*	57.7*	0.47*	65.9*	13.99*	159.4*	1.54*	110.9*	4.37*
79-08-13	* 15H50*	0.400	*0*U*	80.5*	*	*	*	*	*	*	*	*	*	*	*	*
79-08-27	* 14H10*	0.425	*0*U*	77.0*	*	*	*	*	*	*	*	*	*	*	*	*
79-09-17	* 14H10*	0.390	*0*U*	89.7*	*	*	*	*	*	*	*	*	*	*	*	*
79-09-24	* 14H05*	0.370	*0*U*	60.5*	*	*	*	*	*	*	*	*	*	*	*	*
79-10-08	* 14H10*	0.360	*0*U*	82.3*	*	*	*	*	*	*	*	*	*	*	*	*
79-10-15	* 13H45*	0.329	*0*U*	77.6*	568*	7.58*	64.5*	26.3*	56.3*	0.96*	63.8*	9.08*	146.1*	1.74*	128.4*	6.37*
79-10-22	* 14H05*	0.900	*0*U*	34.4*	*	*	*	*	*	*	*	*	*	*	*	*
79-11-05	* 13H30*	0.360	*0*U*	60.4*	*	*	*	*	*	*	*	*	*	*	*	*
79-11-12	* 14H05*	0.390	*0*U*	61.6*	430*	7.05*	42.2*	22.1*	46.4*	0.63*	46.5*	8.07*	99.3*	0.88*	105.2*	7.00*
79-11-19	* 13H15*	0.600	*0*U*	33.2*	*	*	*	*	*	*	*	*	*	*	*	*
79-12-03	* 14H10*	0.510	*0*U*	56.0*	*	*	*	*	*	*	*	*	*	*	*	*
79-12-10	* 12H55*	0.410	*0*U*	68.2*	581*	7.00*	49.3*	22.7*	49.8*	0.51*	59.2*	31.53*	98.5*	1.02*	123.9*	8.75*
79-12-17	* 12H50*	0.410	*0*U*	91.5*	*	*	*	*	*	*	*	*	*	*	*	*
79-12-24	* 13H05*	0.605	*0*U*	53.7*	436*	7.00*	40.9*	16.0*	34.2*	0.39*	50.5*	19.44*	63.0*	1.06*	110.5*	6.77*
79-12-31	* 14H15*	0.405	*0*U*	88.0*	709*	6.90*	68.6*	25.5*	61.0*	1.04*	71.2*	22.70*	199.0*	0.51*	138.0*	9.04*
80-01-07	* 12H35*	0.570	*0*U*	91.9*	803*	7.17*	76.3*	30.4*	65.9*	0.98*	75.6*	52.70*	201.4*	1.14*	83.6*	8.78*
80-01-14	* 13H25*	0.455	*0*U*	76.3*	707*	7.17*	58.0*	25.8*	55.2*	0.78*	62.6*	56.53*	145.4*	0.72*	81.5*	9.59*
80-01-21	* 12H05*	0.530	*0*U*	57.2*	454*	6.94*	32.0*	18.0*	40.4*	0.69*	34.5*	31.77*	94.0*	0.45*	70.8*	7.06*
80-01-28	* 12H30*	0.520	*0*U*	76.2*	658*	7.20*	64.7*	23.4*	51.5*	0.65*	72.0*	44.40*	119.0*	0.11*	100.0*	8.97*
80-02-04	* 13H25*	0.590	*0*U*	80.5*	729*	7.20*	80.8*	23.8*	54.3*	0.64*	78.8*	52.04*	124.5*	1.29*	101.7*	8.54*
80-02-11	* 13H35*	0.535	*0*U*	88.4*	563*	7.20*	83.9*	22.5*	55.1*	0.58*	173.3*	9.96*	115.1*	0.86*	89.4*	4.22*
80-02-18	* 13H25*	0.750	*0*U*	66.4*	459*	7.29*	51.0*	20.4*	48.5*	0.60*	56.9*	8.78*	105.2*	0.90*	105.5*	8.06*
80-02-25	* 13H55*	0.660	*0*U*	95.8*	241*	6.88*	25.2*	11.7*	24.7*	0.37*	33.2*	3.25*	46.7*	0.23*	64.5*	5.36*
80-03-03	* 13H50*	0.630	*0*U*	58.7*	404*	6.91*	42.1*	20.6*	43.1*	0.72*	47.0*	9.87*	82.5*	0.77*	94.2*	8.26*
80-03-10	* 13H15*	0.470	*0*U*	63.4*	453*	7.08*	46.6*	23.2*	44.2*	0.60*	48.0*	7.31*	87.2*	1.22*	132.1*	8.22*
80-03-11	* 13H05*	0.460	*1*U*	66.0*	459*	7.30*	48.6*	23.9*	46.7*	0.42*	52.8*	7.06*	87.8*	1.35*	128.4*	7.92*
80-03-17	* 14H00*	0.455	*0*U*	72.0*	505*	7.05*	55.6*	25.5*	50.4*	0.99*	57.9*	10.98*	118.8*	1.31*	111.2*	8.41*
80-03-20	* 13H10*	0.490	*1*U*	44.4*	318*	7.10*	26.3*	16.7*	35.4*	0.42*	35.9*	4.07*	55.7*	0.82*	99.7*	6.31*
80-03-21	* 11H00*	0.640	*1*U*	50.3*	356*	7.30*	33.1*	18.3*	38.5*	0.49*	40.2*	6.63*	66.7*	1.19*	97.6*	7.03*
80-03-22	* 09H30*	0.490	*1*U*	63.9*	435*	7.29*	43.2*	20.3*	44.1*	0.68*	48.4*	9.98*	108.1*	0.87*	93.7*	7.99*
80-03-23	* 10H00*	0.610	*1*U*	28.7*	185*	6.65*	16.0*	9.0*	21.4*	0.43*	24.4*	2.74*	34.8*	0.26*	50.0*	5.25*
80-03-24	* 14H30*	0.500	*1*U*	41.3*	305*	7.20*	26.3*	15.1*	32.5*	0.38*	34.1*	3.45*	50.8*	0.62*	100.2*	6.15*
80-03-25	* 10H30*	0.500	*1*U*	56.0*	383*	7.20*	43.3*	18.2*	39.5*	0.40*	64.8*	4.54*	62.7*	1.05*	101.6*	6.95*
80-03-26	* 10H40*	0.500	*1*U*	102.4*	646*	7.30*	102.1*	27.0*	63.8*	0.58*	174.4*	9.14*	87.9*	0.06*	116.0*	8.07*
80-03-27	* 09H30*	0.500	*1*U*	72.0*	483*	7.55*	58.3*	23.0*	50.3*	0.59*	80.4*	8.42*	85.2*	1.02*	113.5*	8.65*
80-03-28	* 10H50*	0.480	*1*U*	68.2*	465*	7.67*	51.7*	23.4*	49.0*	0.62*	65.1*	8.51*	91.0*	1.02*	112.4*	8.39*
80-03-29	* 10H20*	0.460	*1*U*	67.6*	470*	7.81*	51.2*	24.3*	48.7*	0.64*	60.8*	8.97*	95.9*	1.09*	113.4*	8.39*
80-03-30	* 12H30*	0.470	*1*U*	70.0*	479*	7.67*	53.2*	24.0*	49.2*	0.70*	57.7*	9.42*	104.5*	1.12*	112.3*	8.20*
80-03-31	* 13H30*	0.420	*0*U*	72.5*	436*	7.32*	52.9*	22.1*	47.3*	0.76*	57.3*	2.15*	113.1*	1.10*	100.7*	6.08*

In Afdeling 5.2 word 'n kort oorsig gegee oor verskillende metodes om ongelyk gespasiëerde en verlore data te hanteer terwyl in Afdeling 5.3 vier van hierdie metodes gebruik word om hulle te vergelyk teen die Kalmanfilter-metode.

Ten slotte word die resultate van hierdie hoofstuk kortliks saamgevat en bespreek in Afdeling 5.4.

## 5.2 'N KORT OORSIG OOR VERSKEIE TEGNIEKE OM VERLORE EN ONGELYK GESPASIËERDE DATA TE HANTEER.

Die tegnieke kan in twee groepe verdeel word naamlik die metodes wat die data voorberei sodat standaard tydreeksanalise-tegnieke toegepas kan word, en die tweede groep wat die verlore data beraam en invul.

### 5.2.1 METODES WAT DIE TYDREEKS EERS VERWERK

Die grootste nadeel van hierdie metodes is dat hulle nie die kovariansiestruktuur in aanmerking neem nie en dit kan 'n ernstige uitwerking  $\hat{h}$  op die beraming van die model.

#### a) Die neem van gemiddeldes

Die eenvoudigste metode om beide die probleem van verlore en die van ongelyk gespasiëerde data te omseil, is om gemiddeldes te neem totdat 'n min of meer volledige reeks gevind is. Wanneer na die uittreksel van die datastel soos gegee in die inleiding, gekyk word, sien 'n mens dat daar sprake is van daaglikse, weeklikse en maanddata. Deur maandgemiddeldes te neem word 'n betreklik volledige datastel gevind. Vir September 1979 kort daar wel 'n inskrywing vir die maand, maar dit kan met oordeel ingevul word, aangesien 'n enkele waarneming nie die modelpassing noemenswaardig sal beïnvloed nie.

Die neem van gemiddeldes het egter 'n gelykstrykende effek en kan tot gevolg hê dat die inherente veranderlikheid onder beraam word. Die punte wat so verkry word het verskillende presisie alhoewel standaard metodes van analise alle punte hanteer asof hulle dieselfde presisie het. Die beraming van die kovariansiestruktuur word ook nadelig beïnvloed.

#### b) Die skuif van datapunte

Veronderstel data word weekliks maar op verskillende dae ingewin, dan kan die punte hernoem word na se week1, week2, ensovoorts. Tensy daar 'n spesifieke dag-van-die-week effek is, sal dit die beraming nie noemenswaardig beïnvloed nie. Daar word dus vaste, gelyk gespasiëerde punte gekies. Die waarneming naaste aan die punt word dan gekies asof hy gemeet is by daardie punt. Ekstra punte word geïgnoreer en die waarnemings moet in 'n sekere interval val, anders word dit as verlore beskou. Enkele gapings kan weer ingevul word met behulp van byvoorbeeld regressie of lineêre interpolasie. Alhoewel die resulterende punte gelyke presisie het, gaan daar 'n groot hoeveelheid inligting verlore, veral as data soms daagliks, dan weekliks ingewin was, sodat data gekies moet word om 'n weeklikse of selfs 'n maandreeks te vorm.

#### c) "Seisoenale aanpassing"

Die tegniek word besryf in 'n artikel deur McLeod, Hipel en Comancho(1983) en transformeer ongelyk gespasiëerde waarnemings na gelyk gespasiëerde beramers. In hierdie metode word verskeie iterasies uitgevoer waartydens die tendens, seisoenale en onreëlmatige variasie komponente beraam en aangepas word. Mediane word hoofsaaklik gebruik en potensiele uitskieters aangepas. Die metode lewer 'n volledige maar 'n uiters gelykgestrykte reeks waarvan die variansie- en kovariansiestruktuur betreklik baie verstuur is.

In 'n studie deur Swart(1989) is bogenoemde drie metodes toegepas in 'n analise van waterkwaliteitdata van die Vaalrivier. Oor die algemeen het die resulterende tydreeksmodelle, gepas op die data gegenerer deur die drie metodes, goed vergelyk.

### 5.2.2 METODES WAT VERLORE DATA BERAAM EN INVUL.

#### a) Vooruit- en Terugberaming

Die metode het betrekking op gevalle waar daar, verkieslik, 'n enkel gaping (van nie te veel waarnemings nie) met lang, ononderbroke reekse waarnemings aan beide kante, is. Daar mag meer as een gaping wees, maar met die voorbehoud dat daar voldoende waarnemings tussenin is.

Die voorwaarde van baie waarnemings is nodig om 'n voldoende ARIMA-model te kan pas. Die waarnemings voor die gaping word geneem, waarna met behulp van die SAS ARIMA-prosedure 'n model gepas en die verlore data vooruitgeskat word. Die data na die gaping word omgekeer, weereens 'n model gepas (vandaar die vereiste van genoeg waardes na die gaping) waarna die gaping weer "terugberaam" word. Vir die gaping is daar dus twee stelle beramings. Vir die studie is die voor- en terugskattings geneem; die gemiddeld van die twee asook die geweegde gemiddeld van die twee. Vir laasgenoemde is as gewigte geneem die inverse van die betroubaarheidsintervalbreedtes.(Sien Bylaag A)

#### Opmerkings

1. Ferriero(1986) haal 'n interpolasiemetode aan van Brubacher en Wilson(1976) wat gebruik kan word om die verlore data te beraam. Waar daar gebruik gemaak is van die SAS-prosedure (gebaseer op die Box-Jenkins benadering) het Brubacher en Wilson dan ook 'n ander benadering voorgestel om die parameters van die ARMA-model te beraam.

2. Damsleth(1980) ontwikkel nog 'n metode om die "voor- en terugskattings" te kombineer in "tussenskattings" met 'n minimum benaderingsfout.

3. Abraham(1981) gebruik weer 'n geometriese benadering tot die probleem om die "voor- en terugskattings" optimaal te gebruik.

### b) Outoregressie

Soos reeds genoem in Hoofstuk3 kan enige ARIMA-model geskryf word as 'n MA-model met 'n oneindige aantal terme.

So kan die ARMA(1,1)-model geskryf word as die volgende bewegende-gemiddelde reeks:

$$y_t = \mu + \eta_t + (\phi - \theta)\eta_{t-1} + (\phi - \theta)\phi\eta_{t-2} + (\phi - \theta)\phi^2\eta_{t-3} + \dots$$

Die reeks kan gesien word as 'n regressievergelyking waar die foutterme gekorreleerd is, met ander woorde die foutterme het 'n kovariansiematriks  $\sigma^2V$ , waar  $V$  afhang van beide die AR en MA parameters van die oorspronklike model:

$V$  het as diagonaalelemente die variansie van  $y_t$  nl.

$$(1 - 2\phi\theta + \theta^2)/(1 - \phi^2)$$

en die ander terme word gegee deur die kovariansie tussen  $y_t$  en  $y_{t-j}$  nl.

$$\phi^{j-1}(\phi - \theta)(1 - \phi\theta)/(1 - \phi^2)$$

Aangesien, vir stasionêriteit, die parameters  $\phi$  en  $\theta$  in absolute waarde kleiner as 1 moet wees, sterf die kovariansies eksponensieel met toenemende sloering,  $j$  (Sien Box en Jenkins p76).

Sou die waardes van die AR- en MA-parameters (wat in  $V$  voorkom) bekend gewees het, kon die probleem opgelos word via die veralgemeende regressiemetode. Dit is egter gewoonlik nodig om die parameters te beraam om sodoende 'n beraming van die kovariansiematriks  $V$  te vind. Foutiewe beraming van die kovariansiematriks kan tot gevolg hê dat 'n swak beraming van die model (wat op sy beurt weer gebruik gaan word

om die verlore data te voorspel), verkry word.

Die probleem kan opgelos word deur 'n iterasie tussen die beraming van die elemente van  $V$  en beraming van die veralgemeende regressiemodel.

Vir die beraming van  $V$  moet 'n beraming van die aantal sloerings wat gebruik gaan word in die model, gemaak word.

Om die metode te implimenter is die SAS AUTOREG—prosedure gebruik.(sien Bylaag B) Verskillende sloerings was gespesifiseer (van lengte 8, 20, 30 en 50) om die invloed van die parameter te bepaal. Na 'n sloering van 20 was daar nie 'n noemenswaardige verandering in die skatting van die parameters of 'n noemenswaardige verbetering in die residue som van vierkante nie, en is daarmee volstaan.

#### c) Latfunksies ( *Splines* )

Latfunksies kan by enige datastel eksak of benaderd gepas word. Hierdie funksies word nie beperk tot gelyk gespasieerde data nie en kan ook oor gapings gepas word. Die funksies kan dan gebruik word om enige waarde by enige punt te interpoleer. Sodoende kan die verlore punte beraam word.

Aangesien omgewingsdata, in die besonder, gevoelig is vir uitskieters en gladstryking in 'n mate sinvol is, is na die klas van B— en kubieselatfunksies gekyk. Dit is gevind dat die CSAKM—subroetine in die IMSL—biblioteek, wat 'n kubiese Akima latfunksie implimenter, die beste vir die datastelle werk(sien Bylaag C).

#### d)"Kriging"

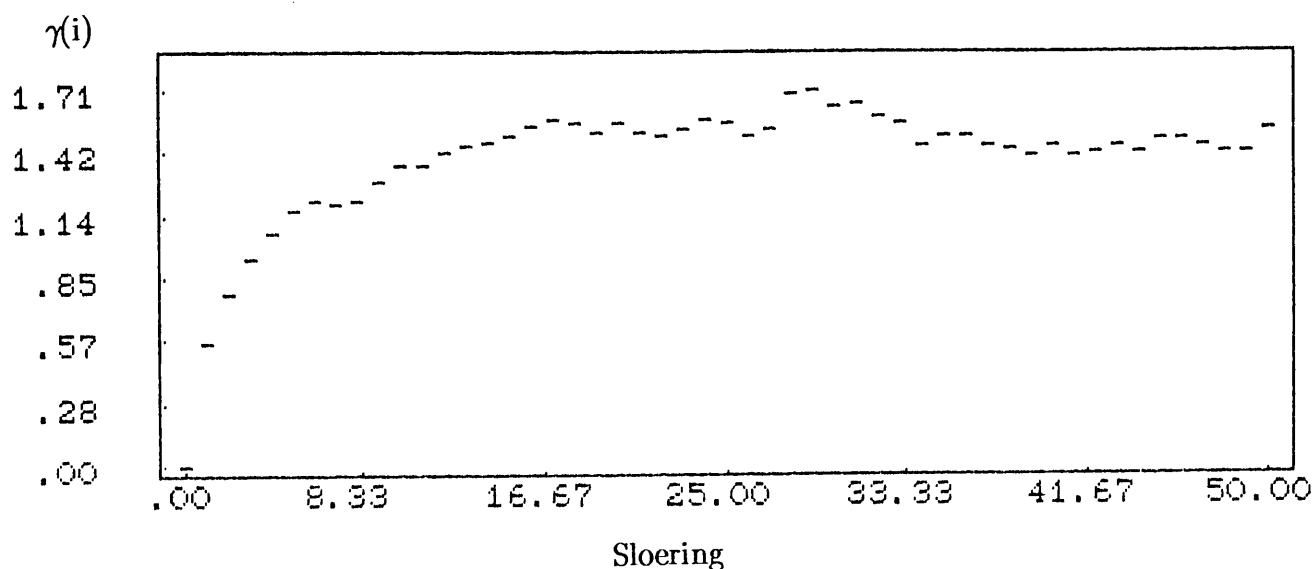
'n Suid—Afrikaanse myningeneur, D.G.Krige, was een van die eerstes wat gebruik gemaak het van ruimtelike korrelasie en lineêr onsydige beraming in die gebied van delfstofbronne evaluering. Die geostatistiese metodes is gerig op die probleem van ruimtelik gekorreleerde data en kan onreëlmatig gespasieerde data hanteer. Die metodes is dan ook gekies aangesien die data wat in die studie gebruik is, beskou kan word as 'n veralgemening van

die ruimtelike opset daar dit slegs eendimensioneel gekorreleerd is.(Sien Galpin en Basson(1990)).Die metode begin deur die berekening van 'n semivariogram. Die semivariogram is 'n grafiek van  $\gamma(i) = [\text{konstante} - \text{kovariansie}(\text{sloering } i)]$  teenoor die sloering (i).

Figuur 1 is die semivariogram vir die Box–Jenkins data met sloerings geneem tot 50.

Figuur 1

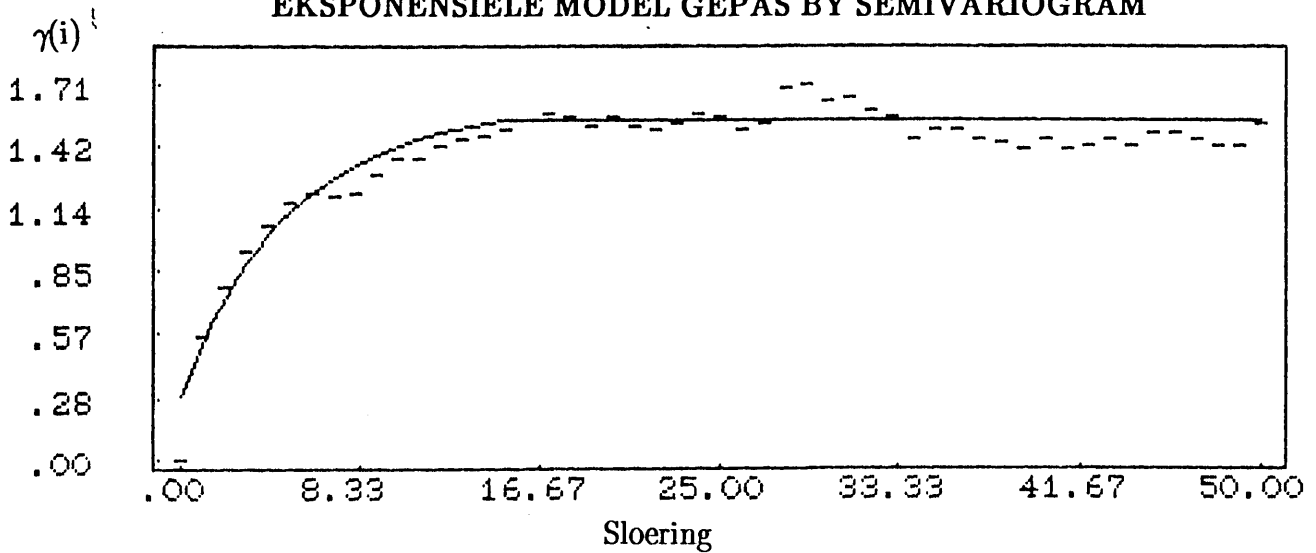
SEMIVARIOGRAM VIR BOX–JENKINS (REEKS A)–DATA



'n Belangrike aspek in Geostatistiek is die beraming van die variogram wat neerkom op die beraming van  $\sigma^2V$  soos in 'n regressieprobleem. Sou die data afkomstig gewees het van 'n bekende model, word die teoretiese vorm van  $\sigma^2V$  gebruik. Dit was egter nie die geval nie en daar is gevind dat die eksponensiele model 'n goeie passing gee...sien Figuur 2.

Figuur2

### EKSPONENSIELE MODEL GEPAS BY SEMIVARIOGRAM

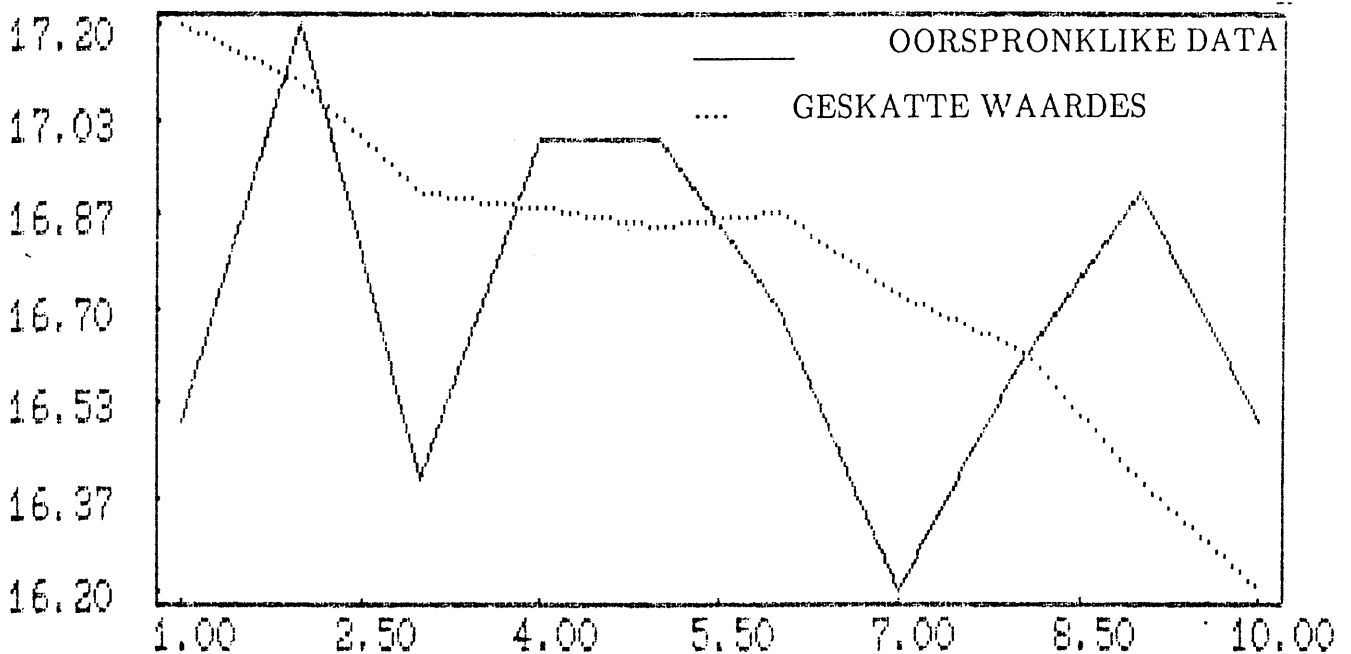


Die model word vervolgens in 'n Kriging-program gebruik en gee die geskatte waardes vir die verlore data.

Figuur 3 toon die oorspronklike punte en die geskatte waardes vir 'n gaping van 10 geneem oor die punte.

Figuur 3.

### GRAFIEK VAN OORSPRONKLIKE DATA EN DIE GESKATTE WAARDES VIR 'N GAPING VAN 10 (BOX-JENKINS-DATA)



Ander metodes sluit die gebruik in van die EM("Expectation Maximization")-algoritme. Die metode alterneer tussen die beraming van die verlore waardes en die beraming van die parameters. Alhoewel die skatting van die parameters wel maksimum aanneemlikheid (MA)-beramers is, word die skatting van die verlore data nie noodwendig gedoen met behulp van die MA nie. Die enigste voordeel van die metode blyk rekenkundig te wees in die sin dat 'n matriks se inverse nie gevind, en die aanneemlikheidsfunksie nie bereken hoef te word nie. Vir die studie is die metode nie verder bestudeer nie en die leser wat meer hiervan te wete wil kom, word verwys na Sargan en Drettakis(1983) en Harvey en Mckenzie(1983).

### 5.3 VERGELYKING VAN METODES

Van die metodes soos beskryf in Afdeling 5.2 is "Vooruit- en Terugberaming", Outoregressie, Latfunksies en "Kriging" geneem om 'n numeriese vergelyking te tref met die Kalmanfilter-metode.

Die metodes is toegepas op 'n gegengereerde datastel van 400 ARMA(1,1) -waardes met  $\phi=0.8$  en  $\theta=0.3$ , asook op die REEKS A ("Chemical process concentration readings")-datastel van Box en Jenkins(1976). Verlore data is geskep deur gapings van 10 en 50 op verskillende plekke in die reeks weg te laat. Voorsorg is getref dat daar altyd voldoende data aan beide kante van die gaping gelaat word om die metode "Vooruit- en Terugberaming" te akkommodeer. Vir die gegengereerde datastel is die resultate met die gaping in die begin van die reeks gegee, terwyl vir die Box-en-Jenkins-data (200 waardes) die gaping in die middel geneem is. Die volgende drie metodes is gebruik om 'n vergelyking te kan tref.

1. Vir die gegengereerde datastel is die parameters  $\phi=0.8$  en  $\theta=0.3$  gekies. Vervolgens is die "verlore data" deur die verskillende metodes ingevul, die model weereens gepas en die

Vir die Box-en-Jenkins-data is die ARMA(1,1)-model met  $\phi=0.9$  en  $\theta=0.56$ , op die volledige reeks gepas en die ander "ingevulde" modelle daarteen vergelyk. Die resultate word weergegee in Tabela 5.1 en 5.2 en die verskillende metodes lewer merkwaardige eenderse resultate.

**TABEL 5.1**  
**GEGENEREERDE ARMA(1,1) WAARDES**

$$\phi = 0.8 \quad \theta = 0.3$$

	10 VERLORE PUNTE		50 VERLORE PUNTE	
	AR(1)	MA(1)	AR(1)	MA(1)
VOOR- EN TERUGSKATTING	0.8128	0.3337	0.7941	0.3364
OUTOREGRESSIE	0.8144	0.3376	0.7969	0.3427
LATFUNKSIES	0.8136	0.3292	0.8165	0.3443
"KRIGING"	0.8131	0.3338	0.7999	0.3350
KALMANFILTER	0.8126	0.3334	0.7932	0.3351

**TABEL 5.2**  
**BOX-EN-JENKINS-DATA**  
**(ARMA(1,1)-MODEL MET  $\phi = 0.9$  EN  $\theta = 0.56$ )**

	10 VERLORE PUNTE		50 VERLORE PUNTE	
	AR(1)	MA(1)	AR(1)	MA(1)
VOOR- EN TERUGSKATTING	0.8717	0.4837	0.8773	0.5158
OUTOREGRESSIE	0.8772	0.5053	0.8955	0.5439
LATFUNKSIES	0.8848	0.5069	0.9012	0.5673
"KRIGING"	0.8856	0.5140	0.9019	0.5663
KALMANFILTER	0.8801	0.5097	0.9053	0.5721

2. Vir die tweede metode is die gemiddeld van die som van vierkante van die residue geneem. Weereens is die "verlore data" deur die verskillende metodes ingevul, 'n model gepas en die residue geneem van die werklike waardes en die voorspelde waardes. Die resultate word in Tabel 5.3 opgesom. Die metode waar van latfunksies gebruik gemaak is, lewer die kleinste waarde van die gemiddeld van die som van vierkante van die residue, maar weereens lewer die verskillende metodes nie noemenswaardig verskillende resultate nie.

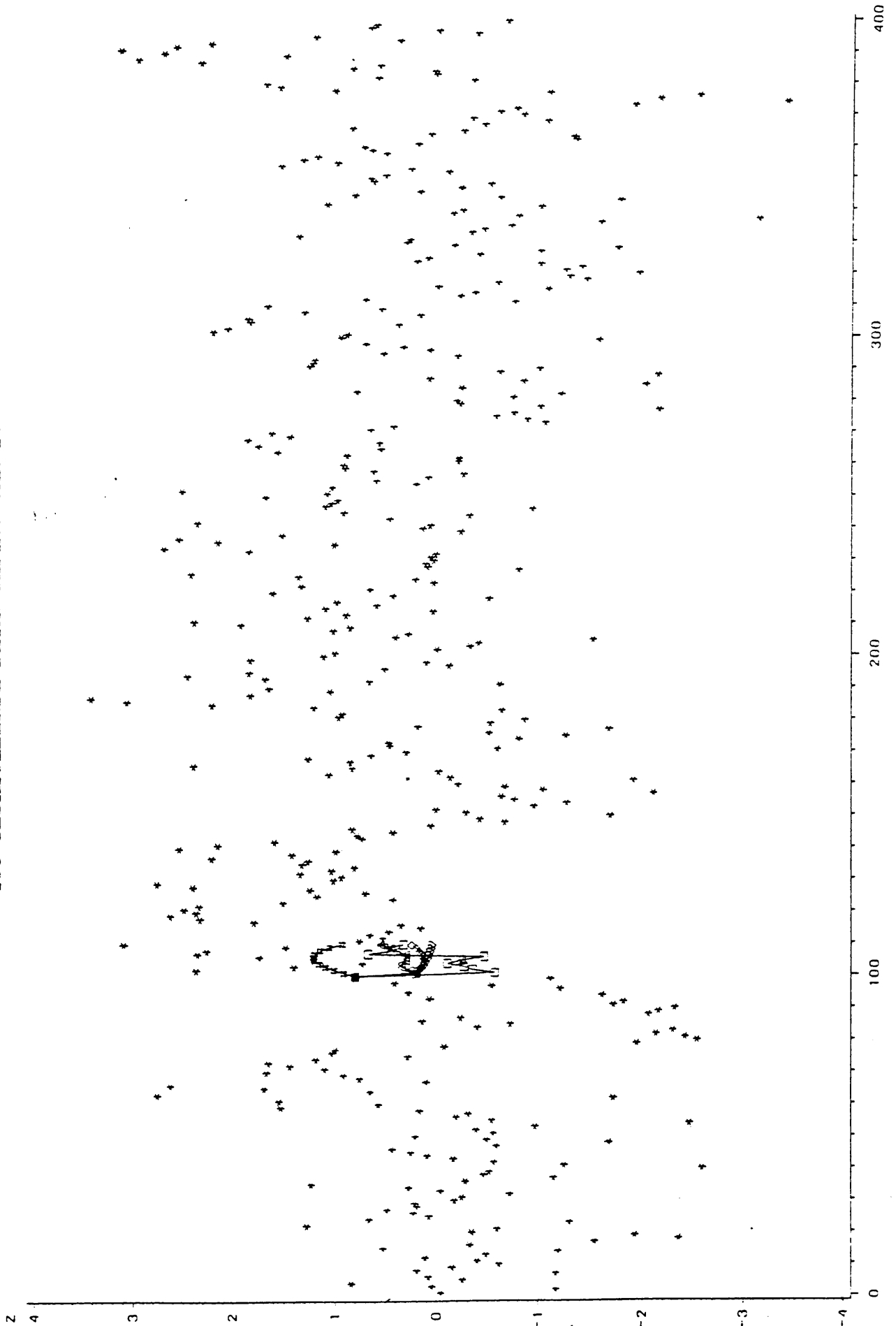
**TABEL 5.3**  
**GEMIDDELD VAN DIE VIERKANTE VAN DIE RESIDUE**

GAPING	GEGENEREERDE ARMA(1,1)		BOX-EN-JENKINS	
	10	50	10	50
VOOR- EN TERUGSKATTING	377.84	444.65	19.68	22.26
OUTOREGRESSIE	381.32	455.95	19.98	27.70
LATFUNKSIES	363.84	363.84	19.09	21.83
"KRIGING"	374.87	420.06	19.24	20.62
KALMANFILTER	378.90	441.82	19.72	20.61
<b>OORSPRONKLIKE REEKS:</b>	<b>340.06</b>		<b>19.42</b>	

3. Figure 5.1 tot 5.8 is die grafiese voorstelling van die verskillende reekse met die verskillende gapings. Die volledige reeks met die gaping van 10 en 50 respektiewelik, word telkens gevolg deur 'n vergroting van die gaping sodat die verskillende tegnieke van nader beskou kan word. In die grafieke kan die tendens dat die ingevulde waardes (by die gapings van 50) konvergeer na die gemiddeld duidelik by sommige van die metodes gesien word.

FIGUUR 5.1

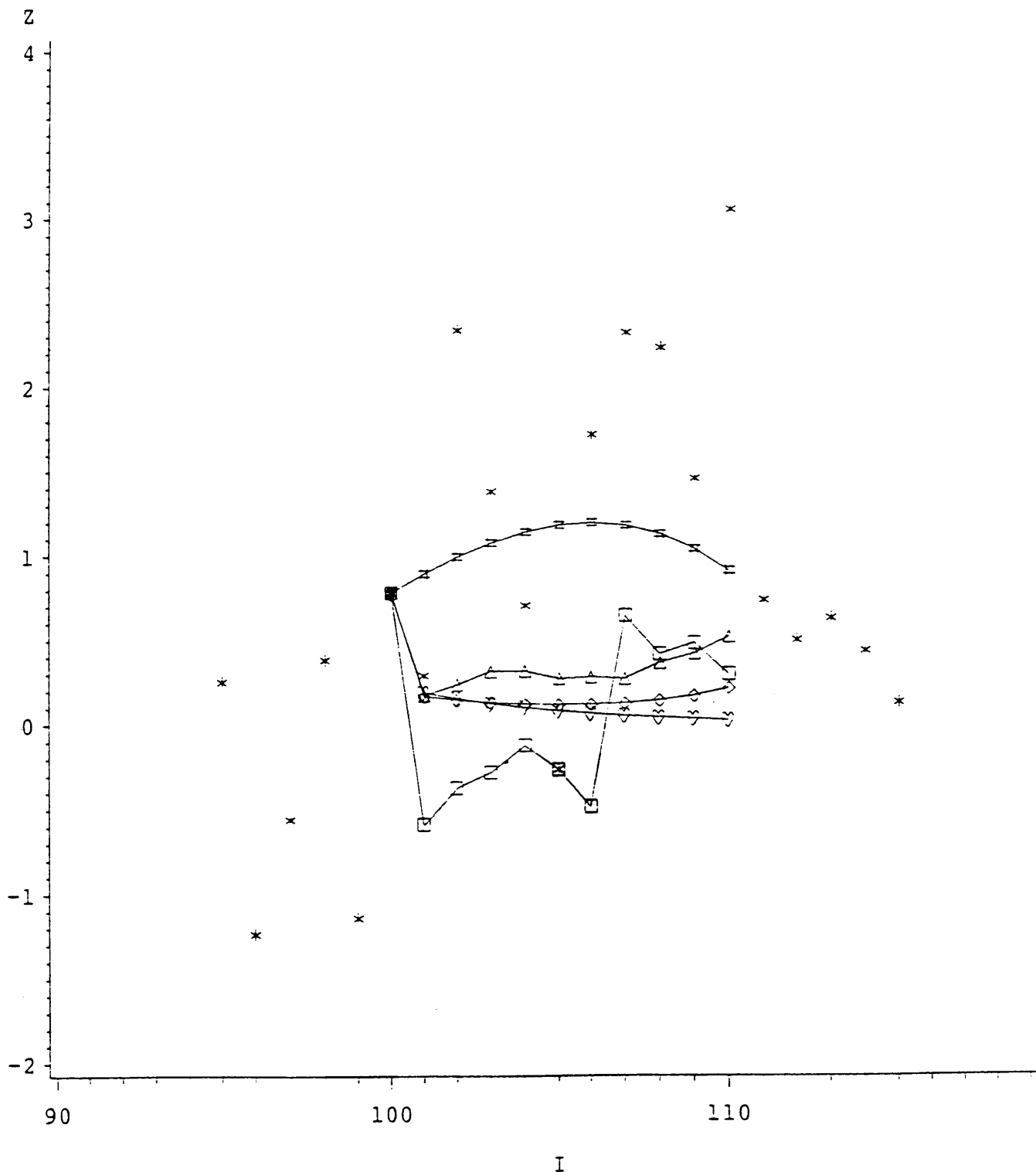
**VERGELYKING VAN METODEDES  
400 GESIMULEERDE DATA: GAPING VAN 10**



FIGUUR 5.2

## VERGELYKING VAN METODEDES (Vergroting)

400 GESIMULEERDE DATA: GAPING VAN 10



**STER: OORSPRONKLIKE DATA**

**VIERTANT: OUTOREGRESSIE**

**DIAMANT: VOOR- EN TERUGSKATTING**

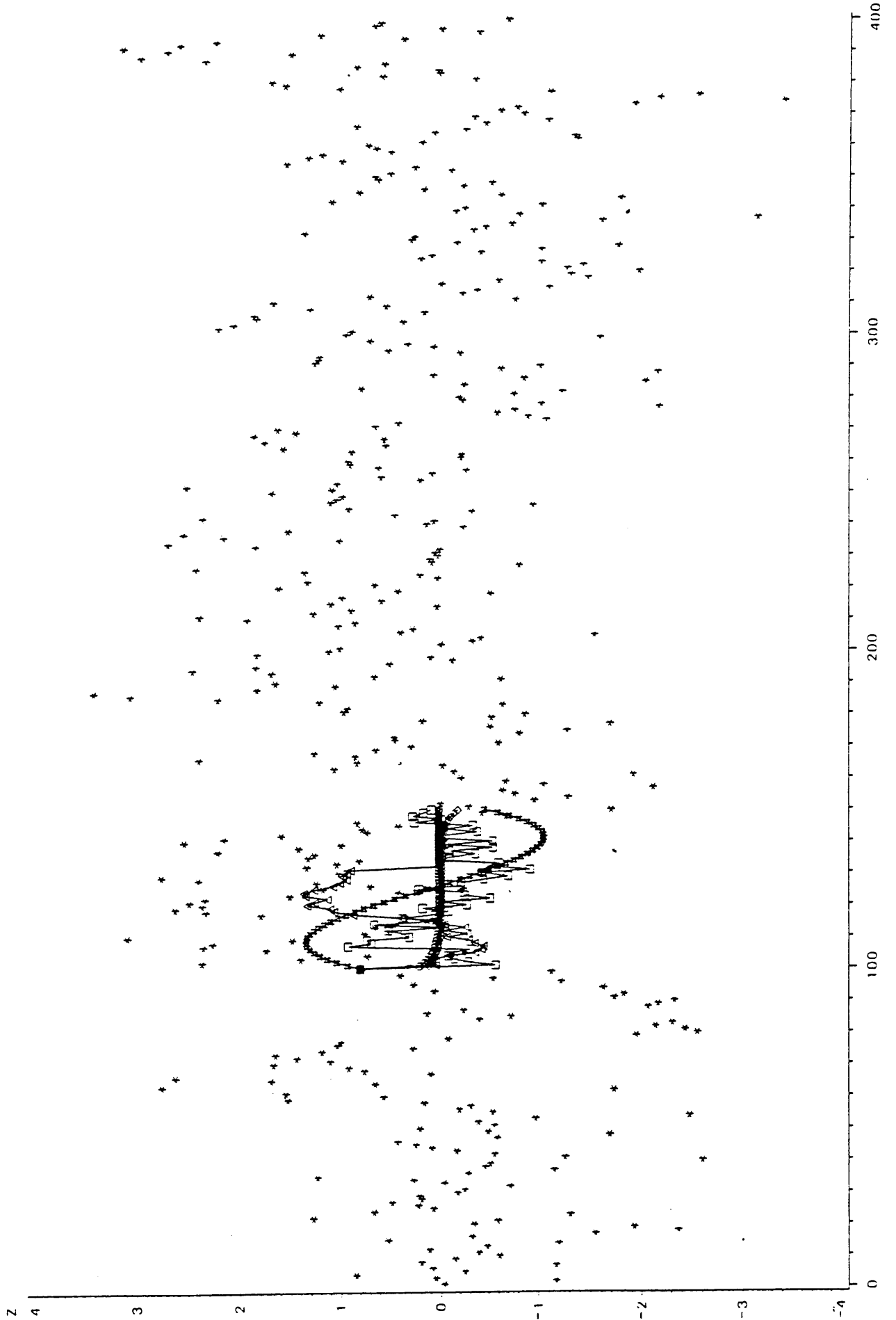
**DRIEHOEK: KRIGING**

**#: LATFUNKSIE**

**HART: MAKSIMUMAANNEEMLIKHEID (KALMAN)**

FIGUUR 5.3

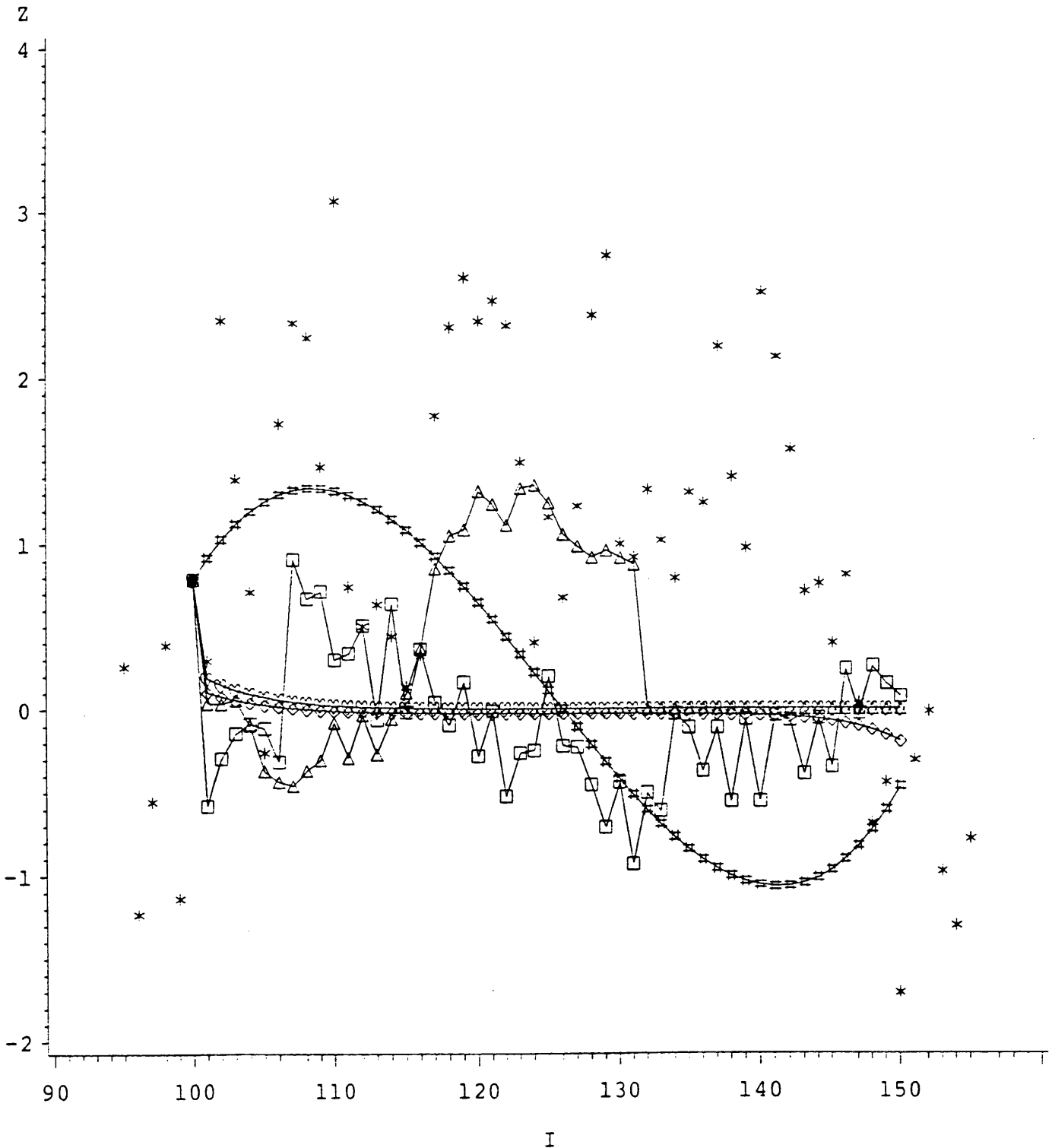
**VERGELYKING VAN METODES  
400 GESIMULEERDE DATA: GAPING VAN 50**



FIGUUR 5.4

# VERGELYKING VAN METODEDES (Vergroting)

400 GESIMULEERDE DATA: GAPING VAN 50



**STER: OORSPRONKLIKE DATA**

**VIERKANT: OUTOREGRESSIE**

**DIAMANT: VOOR- EN TERUGSKATTING**

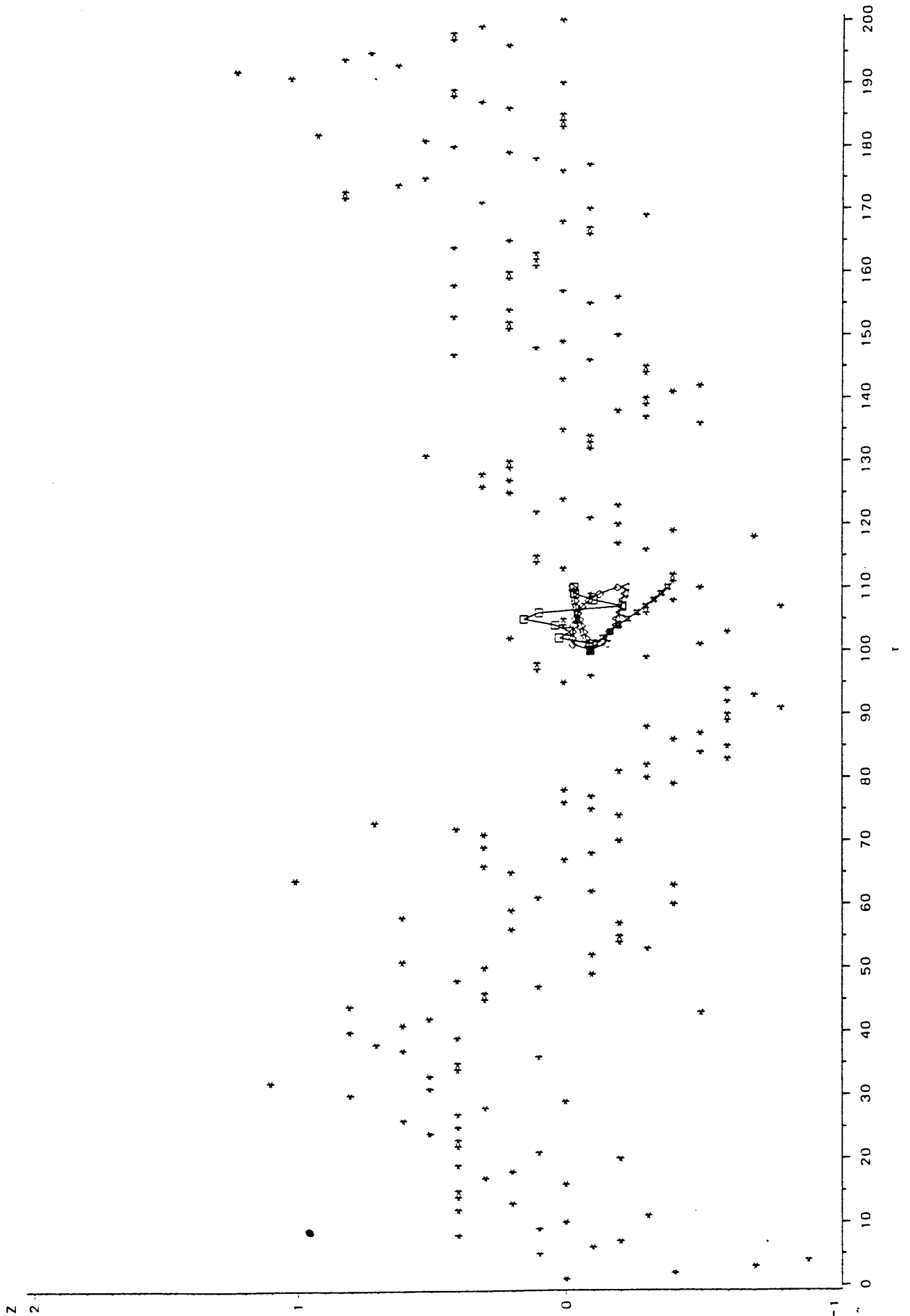
**DRIEHOEK: KRIGING**

**#: LATFUNKSIE**

**HART: MAKSIMUMAANNEEMLIKHEID (KALMAN)**

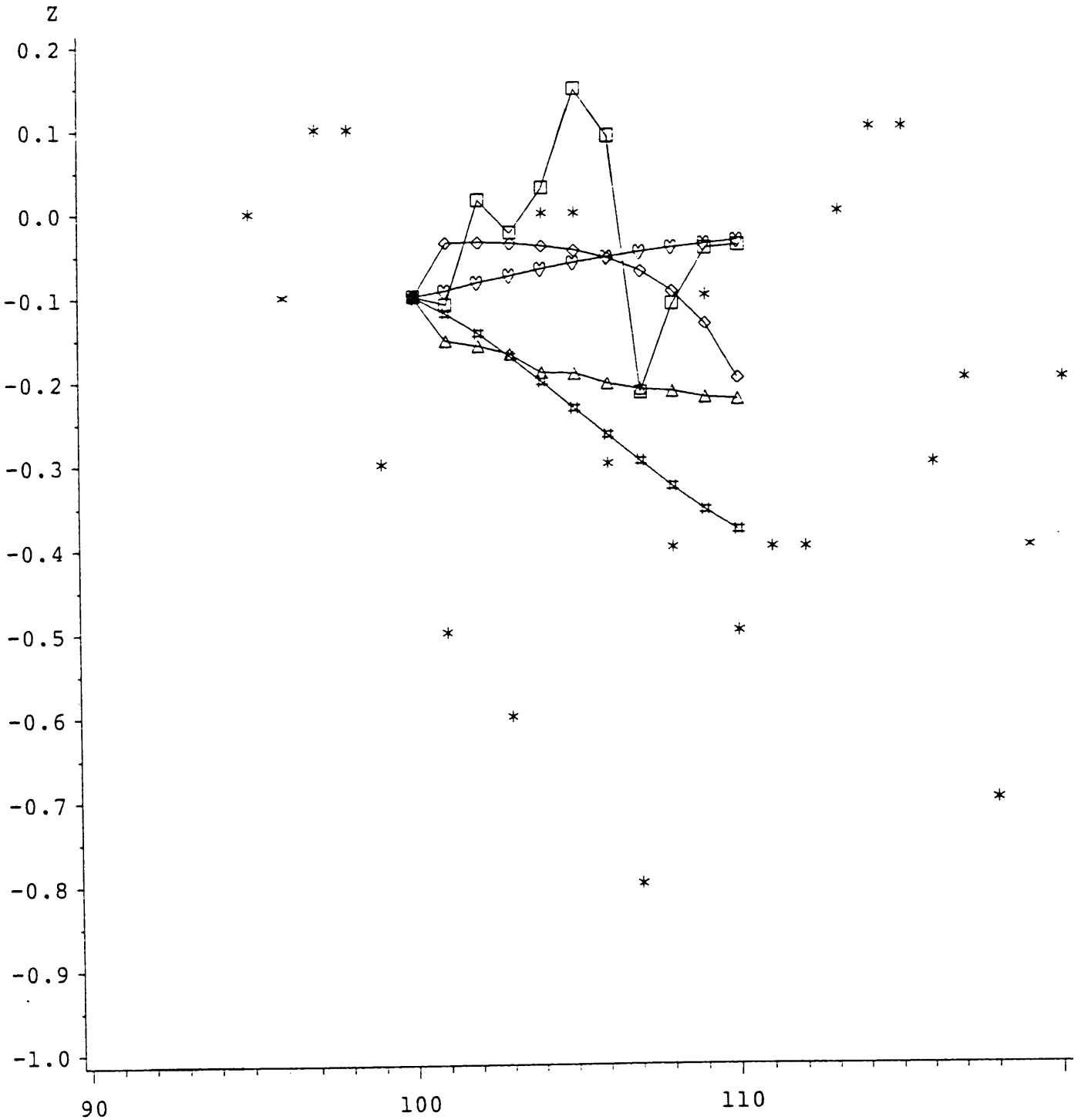
FIGUUR 5.5

**VERGELYKING VAN METODEDES  
BOX EN JENKINS (SERIE A) - DATA: GAPING VAN 10**



FIGUUR 5.6 |

# VERGELYKING VAN METODEDES (Vergroting) BOX EN JENKINS (SERIE A)-DATA: GAPING VAN 10



**STER: OORSPRONKLIKE DATA**

**VIKANT: OUTOREGRESSIE**

**DIAMANT: VOOR- EN TERUGSKATTING**

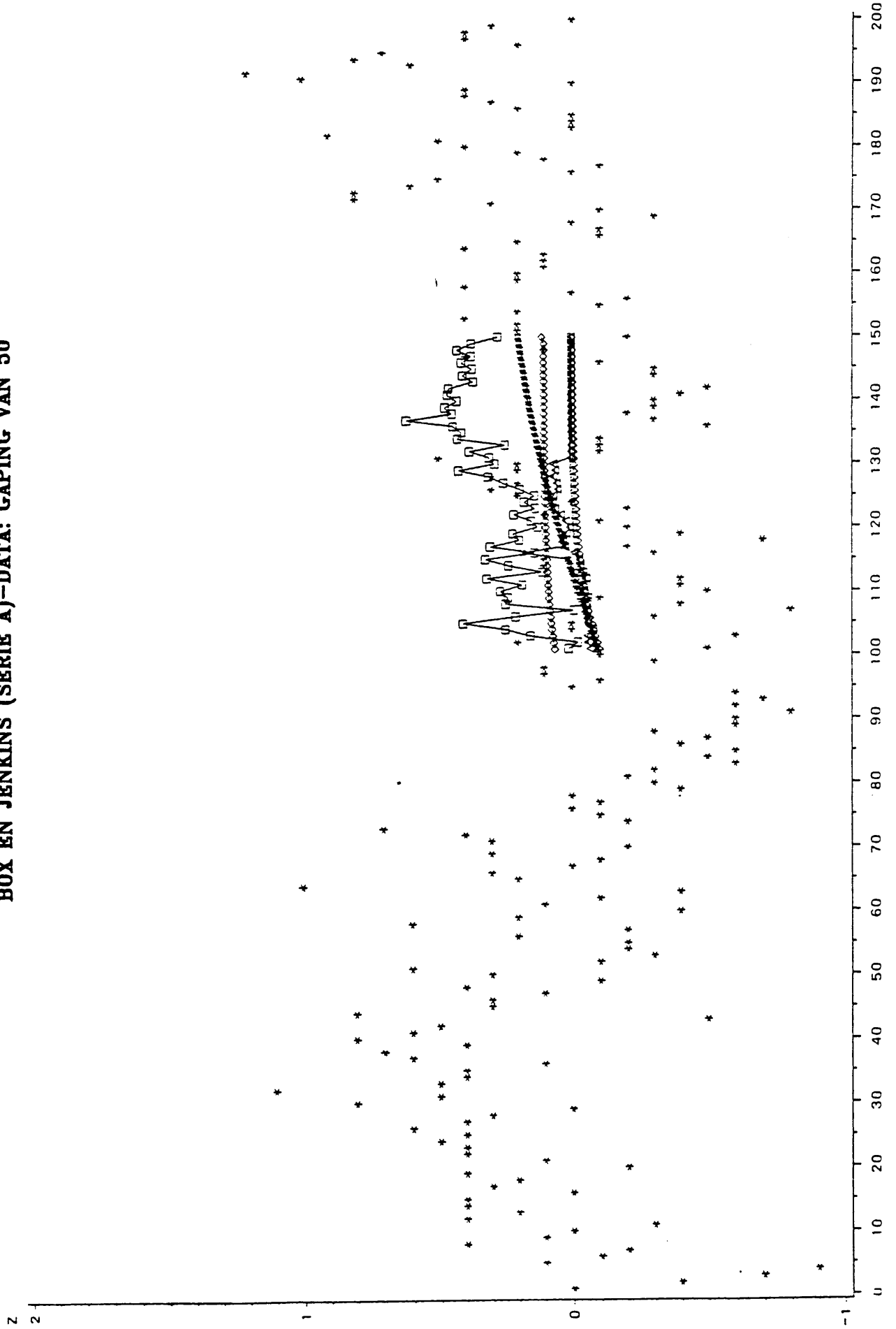
**DRIEHOEK: KRIGING**

**#: LATFUNKSIE**

**HART: MAKSIMUMAANNEELIKHEID (KALMAN)**

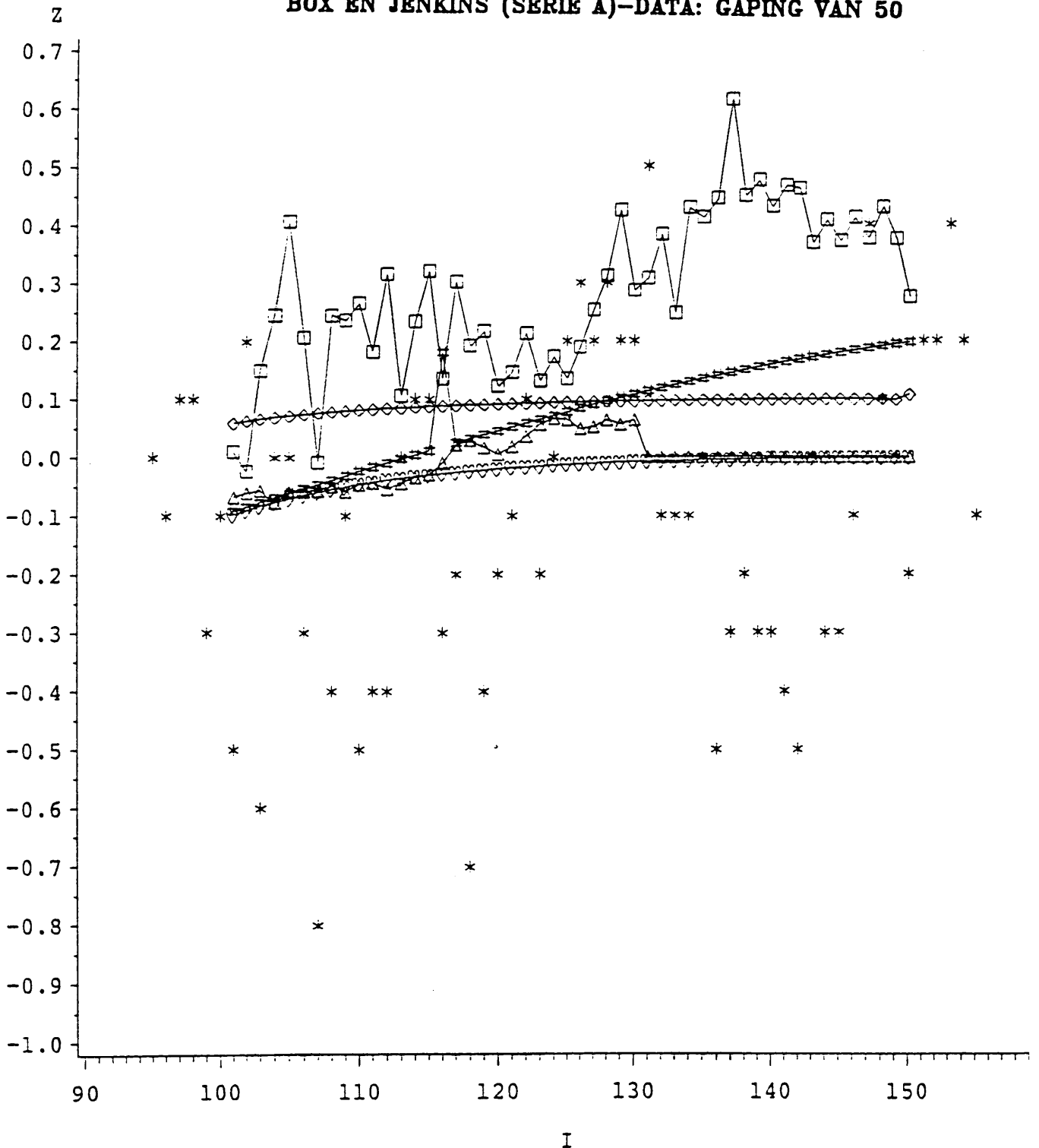
FIGUUR 5.7

**VERGELYKING VAN METODES  
BOX EN JENKINS (SERIE A)-DATA: GAPING VAN 50**



# VERGELYKING VAN METODEDES (Vergroting)

BOX EN JENKINS (SERIE A)-DATA: GAPING VAN 50



**STER: OORSPRONKLIKE DATA**

**VIERKANT: OUTOREGRESSIE**

**DIAMANT: VOOR- EN TERUGSKATTING**

**DRIEHOEK: KRIGING**

**#: LATFUNKSIE**

**HART: MAKSIMUMAANNEEMLIKHEID (KALMAN)**

#### 5.4 OPSOMMING

Daar is tot die gevolgtrekking gekom dat die verskeie tegnieke (insluitend die Kalmanfilter metode) ewe goed presteer vir klein gapings en ewe sleg vir groot gapings met die uitsondering van latfunksies en "Kriging".

Gedurende die opskryf van die navorsingsprojek het die opgedateerde weergawe van SAS voor die dag gekom met die prosedure EXPAND. Die prosedure hanteer tydreeks met verlore data deur kubiese latfunksies op die beskikbare data te pas en met behulp daarvan word kontinue-tyd-beramings van die ingevulde punte gemaak. Baie rekenaarpakette vereis nog steeds dat tydreeks gelykgespasieerde tydsintervalle gebruik en wanneer die gapings nie te groot is nie, kan die Kalmanfilter met sukses gebruik word.

## 5.5 BYLAE

## Bylaag A

## Vooruit-en-terugberaming

```

DATA RAB2;
CMS FILEDEF UIT DISK BOXJ1 DATA A1;
INFILE UIT;
INPUT I Z;
PROC ARIMA DATA=RAB2;
TITLE '200 DATA:BOX-JENKINS-DATA';
I VAR=Z;
E P=1 Q=1 METHOD=ML;
F LEAD=0 ID=I OUT=A111;
DATA A111;
SET A111;
I=_N_;
Z=Z;
PROC MEANS DATA=A111 N USS;
VAR RESIDUAL;
DATA A1;
SET RAB2;
IF _N_ LE 100;
I=_N_;
Z=Z;
PROC ARIMA DATA=A1;
TITLE 'DATA=BOX-J...EERSTE 100 DATA';
I VAR=Z;
E P=1 Q=1 MAXIT=100;
F LEAD=10 ID=I OUT=B11;
F LEAD=50 ID=I OUT=B14;

DATA B11;
SET B11 (FIRSTOBS=101);
I=_N_+100;
PROC PRINT DATA=B11;
DATA B14;
SET B14 (FIRSTOBS=101);
I=_N_+100;
PROC PRINT DATA=B14;

DATA A2;
PROC SORT DATA=RAB2 OUT=RARMAO;
BY DESCENDING I;
DATA RARMAO;
SET RARMAO;
I=_N_;
Z=Z;
PROC ARIMA DATA=RARMAO;
TITLE 'OMGEKEERDE WAARDES';
I VAR=Z;
E P=1 Q=1;
DATA A21;
SET RARMAO;
IF _N_ LE 90;
I=_N_;
Z=Z;
PROC ARIMA DATA=A21;
TITLE 'LAASTE 90 WAARDES';
I VAR=Z;

```

```
E P=1 Q=1 MAXIT=100;
F OUT=B21 LEAD=10 ID=I;
```

```
DATA B21;
SET B21 (FIRSTOBS=91);
I=_N_;
PROC SORT DATA=B21;
BY DESCENDING I;
DATA B21;
SET B21;
I= _N_+100;
PROC PRINT DATA=B21;
DATA A24;
SET RARMAO;
IF _N_ LE 50;
I= _N_;
Z=Z;
PROC ARIMA DATA=A24;
TITLE 'LAASTE 50 WAARDES';
I VAR=Z;
E P=1 Q=1 MAXIT=100;
F OUT=B24 LEAD=50 ID=I;
DATA B24;
SET B24 (FIRSTOBS=51);
I= _N_;
PROC SORT DATA=B24;
BY DESCENDING I;
DATA B24;
SET B24;
I= _N_+100;
PROC PRINT DATA=B24;
```

```
DATA B31;
MERGE B11(RENAME=(FORECAST=F1 U95=U1 L95=L1)) B21(RENAME=(FORECAST=F2
U95=U2 L95=L2));
BY I;
GEM=(F1+F2)/2;
RWYTE1=1/(U1-L1);
RWYTE2=1/(U2-L2);
GGEM=(RWYTE1*F1+RWYTE2*F2)/(RWYTE1+RWYTE2);
VERS1=Z-GEM;
VERS2=Z-GGEM;
DATA B34;
MERGE B14(RENAME=(FORECAST=F1 U95=U1 L95=L1)) B24(RENAME=(FORECAST=F2
U95=U2 L95=L2));
BY I;
GEM=(F1+F2)/2;
RWYTE1=1/(U1-L1);
RWYTE2=1/(U2-L2);
GGEM=(RWYTE1*F1+RWYTE2*F2)/(RWYTE1+RWYTE2);
VERS1=Z-GEM;
VERS2=Z-GGEM;
```

```
DATA B41;
MERGE A1(RENAME=(Z=Z1)) B31;
BY I;
DO J=1 TO 200;
IF 100<I<111 THEN Z1=GGEM;
ELSE Z1=Z1;
END;
PROC ARIMA DATA=B41;
TITLE 'DATA:BOX-J...MET 10 WAARDES VERVANG DEUR HULLE GEWEEGDE GEM.';
I VAR=Z1;
E P=1 Q=1;
DATA B44;
MERGE A1(RENAME=(Z=Z1)) B34;
BY I;
DO J=1 TO 200;
IF 100<I<151 THEN Z1=GGEM;
ELSE Z1=Z1;
END;
PROC ARIMA DATA=B44;
TITLE 'DATA:BOX-J...MET 50 WAARDES VERVANG DEUR HULLE GEWEEGDE GEM.';
I VAR=Z1;
E P=1 Q=1;
```

```

DATA B51;
MERGE RAB2 B41;
BY I;
IF I LE 110 THEN Z=Z1;
ELSE Z=Z;
PROC PRINT DATA=B51;
PROC ARIMA DATA=B51;
TITLE '10 INGELASTE WAARDES';
I VAR=7;
E P=1 Q=1;
F LEAD=0 ID=I OUT=B551;
DATA B551;
SET B551;
I=_N_;
Z=Z;
DATA REG1;
MERGE RAB2 B551(RENAME=(Z=Z1));
BY I;
DATA REG1;
SET REG1;
RES=Z-FORECAST;
DROP Z1 STD L95 U95 RESIDUAL;
PROC MEANS DATA=REG1 N USS;
VAR RES;
TITLE 'GESKATTE WAARDES:10 INGELASTE WAARDES(PHI=0.8 THETA=0.3)';
DATA B54;
MERGE RAB2 B44;
BY I;
IF I LE 150 THEN Z=Z1;
ELSE Z=Z;
PROC ARIMA DATA=B54;
TITLE '50 INGELASTE WAARDES';
I VAR=Z;
E P=1 Q=1;
F LEAD=0 ID=I OUT=B554;
DATA B554;
SET B554;
I=_N_;
Z=Z;
DATA REG4;
MERGE RAB2 B554(RENAME=(Z=Z1));
BY I;
DATA REG4;
SET REG4;
RES=Z-FORECAST;
DROP Z1 STD L95 U95 RESIDUAL;
PROC MEANS DATA=REG4 N USS;
VAR RES;
TITLE 'OORSP. DATA PLUS GESKAT INGELASTE DATA(50)';

DATA BFBJ10;
CMS FILEDEF UIT2 DISK BFBJ10 DATA A1;
MERGE RAB2(RENAME=(Z=Z1)) B31 A111(RENAME=(FORECAST=F3)) B551(RENAME=
(FORECAST=F4));
BY I;
FILE UIT2;
PUT I ( Z1 GGEM ) (10.5);
DATA BFBJ50;
CMS FILEDEF UIT5 DISK BFBJ50 DATA A1;
MERGE RAB2(RENAME=(Z=Z1)) B34 A111(RENAME=(FORECAST=F3)) B554(RENAME=
(FORECAST=F4));
BY I;
FILE UIT5;
PUT I ( Z1 GGEM ) (10.5);

```

## Bylaag B

## Outoregressie

```

DATA A1;
CMS FILEDEF BET DISK BOXJ1 DATA A1;
INFILE BET;
INPUT I Z;

PROC AUTOREG DATA=A1;
TITLE 'AUTOREG(8)...BOX-JENKINS';
MODEL Z=I / NLAG=8;
OUTPUT OUT=T1 PREDICTED=P1 RESIDUAL=RES;
PROC MEANS DATA=T1 N USS;
TITLE 'AUTOREG(8)...BOX-JENKINS';
VAR RES;
PROC AUTOREG DATA=A1;
TITLE 'AUTOREG(20).';
MODEL Z=I / NLAG=20;
OUTPUT OUT=T2 PREDICTED=P2 RESIDUAL=RES;
PROC MEANS DATA=T2 N USS;
TITLE 'AUTOREG(20)...ARMA(1,1)...PHI=0.8,THETA=0.3';
VAR RES;
PROC AUTOREG DATA=A1;
TITLE 'AUTOREG(30)...ARMA(1,1)...PHI=0.8,THETA=0.3';
MODEL Z=I / NLAG=30;
OUTPUT OUT=T3 PREDICTED=P3 RESIDUAL=RES;
PROC MEANS DATA=T3 N USS;
TITLE 'AUTOREG(30)...ARMA(1,1)...PHI=0.8,THETA=0.3';
VAR RES;
PROC AUTOREG DATA=A1;
TITLE 'AUTOREG(50)...ARMA(1,1)...PHI=0.8,THETA=0.3';
MODEL Z=I / NLAG=50;
OUTPUT OUT=T4 PREDICTED=P4 RESIDUAL=RES;
PROC MEANS DATA=T4 N USS;
TITLE 'AUTOREG(50)...ARMA(1,1)...PHI=0.8,THETA=0.3';
VAR RES;

DATA RARMAM1;
SET A1;
IF 50<I<81 THEN Z=.;

PROC AUTOREG DATA=RARMAM1;
TITLE 'AUTOREG(8)...BOX-JENKINS...GAPING=30';
MODEL Z=I / NLAG=8;
OUTPUT OUT=T31 PREDICTED=P31 RESIDUAL=RES;
DATA T3M1;
MERGE A1(RENAME=(Z=Z1)) T31;
RESS=Z1-P31;
PROC MEANS DATA=T3M1 N USS;
TITLE 'AUTOREG(8).';
VAR RES RESS;

PROC AUTOREG DATA=RARMAM1;
TITLE 'AUTOREG(20)';
MODEL Z=I / NLAG=20;
OUTPUT OUT=T41 PREDICTED=P41 RESIDUAL=RES;
DATA T4M1;

MERGE A1(RENAME=(Z=Z1)) T41;
RESS=Z1-P41;
PROC MEANS DATA=T4M1 N USS;
TITLE 'AUTOREG(20).';
VAR RES RESS;

```

```

PROC AUTOREG DATA=RARMAM1;
TITLE 'AUTOREG(30)';
MODEL Z=I / NLAG=30;
OUTPUT OUT=T51 PREDICTED=P51 RESIDUAL=RES;
DATA T5M1;
MERGE A1(RENAME=(Z=Z1)) T51;
RESS=Z1-P51;
PROC MEANS DATA=T5M1 N USS;
TITLE 'AUTOREG(30)';
VAR RES RESS;

PROC AUTOREG DATA=RARMAM1;
TITLE 'AUTOREG(50).';
MODEL Z=I / NLAG=50;
OUTPUT OUT=T61 PREDICTED=P61 RESIDUAL=RES;
DATA T6M1;
MERGE A1(RENAME=(Z=Z1)) T61;
RESS=Z1-P61;
PROC MEANS DATA=T6M1 N USS;
TITLE 'AUTOREG(50)';
VAR RES RESS;

DATA BJAUT8;
CMS FILEDEF AUTO DISK BJAUT8 DATA A1;
MERGE A1(RENAME=(Z=Z1)) T1 T31;
BY I;
FILE AUTO;
PUT I (Z1 P31) (10.5);
PROC PLOT DATA=BJAUT8 (FIRSTOBS=40 OBS=90);
PLOT Z1*I='*' P1*I='8' P31*I='M' / OVERLAY;
TITLE 'AUTOREG(8)...BOX-JENKINS..GAPING=310:"M"';

DATA BJAUT30;
CMS FILEDEF AUTO2 DISK BJAUT30 DATA A1;
MERGE A1(RENAME=(Z=Z1)) T2 T41;
BY I;
FILE AUTO2;
PUT I (Z1 P41) (10.5);
PROC PLOT DATA=BJAUT30 (FIRSTOBS=40 OBS=90);
PLOT Z1*I='*' P2*I='2' P41*I='M' / OVERLAY;
TITLE 'AUTOREG(20)...BOX-JENKINS...GAPING=30:"M"';
DATA C31;
MERGE A1(RENAME=(Z=Z1)) T3 T51;
BY I;
PROC PLOT DATA=C31 (FIRSTOBS=40 OBS=90);
PLOT Z1*I='*' P3*I='8' P51*I='M' / OVERLAY;
TITLE 'AUTOREG(30)..BOX-JENKINS..GAPING=30:"M"';
DATA C41;

MERGE A1(RENAME=(Z=Z1)) T4 T61;
BY I;
PROC PLOT DATA=C41 (FIRSTOBS=40 OBS=90);
PLOT Z1*I='*' P4*I='2' P61*I='M' / OVERLAY;
TITLE 'AUTOREG(50)...BOX-JENKINS...GAPING=30:"M"';

```

## Bylaag C

## Latfunksies

```

PROGRAM SPLINE
C
C   PARAMETER (N=151,NMIS=30,NDATA=N-NMIS,KORDER=3
1   ,NKNOT=NDATA+KORDER,KORD2=4,NCOEF=NDATA,NKNOT2=NCOEF+KORD2)
C
C   INTEGER SPOINT,EPOINT
REAL FDATA(NDATA),XDATA(NDATA),CSCOEF(4,NDATA),BREAK(NDATA)
1   ,BSCOEF(NDATA),ORD(1000),ORD2(1000),XKNOT(NKNOT),ABSC(1000)
2   ,WK1(15000),WK2(NDATA),WK3(NDATA),IWK(NDATA),ABSC2(1000)
3   ,XGUESS(NKNOT2),WEIGHT(NDATA),BSCOF(NCOEF)
4   ,XKNOT2(NKNOT2),XGUES2(700),XGUES3(700),XKNOT3(700)
5   ,BSCOF2(700)
EXTERNAL CSAKM,CSVAL,B2INT,BSVAL,BSNAK,B2VLS
C
90 FORMAT(2I4)
100 FORMAT(F3.0,F4.1)
110 FORMAT(13X,'X',9X,'Y',9X,'INTERPOLANT',5X,'ERROR')
120 FORMAT(' AKIMA CUBIC SPLINE INTERPOLANT ')
130 FORMAT(' CUBIC SPLINE INTERPOLANT WITH NOT-A-KNOT CONDITION ')
140 FORMAT(' B-SPLINE INTERPOLATION ')
150 FORMAT(' LEAST SQUARES APPROXIMATION ')
160 FORMAT(' LEAST SQUARES APP. WITH CHANGED INPUT KNOTS',/,
& 'EVERY SECOND OF ORIGINAL KNOTS USED')
170 FORMAT(13X,'X',9X,'Y',9X,'APPROXIMATION',7X,'ERROR')
180 FORMAT(' SQUARE ROOT OF THE SUM OF SQUARES : ',F9.4)
190 FORMAT(' SUM OF SQUARED ERRORS : ',F15.6)
NINIV = NDATA - 1
C
C   LEES ROU DATA IN, SKEP DATA VEKTOR VIR SPLINE ROETINES DEUR VOORAF
C   BEPAALDE PUNTE UIT TE LAAT EN OOK VEKTOR WAT DIE VOORAF BEPAALDE
C   BEVAT VIR DIE EVALUERING VAN SPLINES
C   NORMAALWEG SAL ONS NIE NODIG HE OM DIT TE DOEN AANGESIEN DAAR
C   VERLORE PUNTE IN DATA IS.DIE PUNTE IS DAN DIE EVALUASIE PUNTE.
C
READ(1,90)SPOINT,EPOINT
WRITE(2,*)' STARTPOINT = ',SPOINT,' ENDPOINT = ',EPOINT
DO 10 I = 1,N
READ(1,100) ABSC(I),ORD(I)
10 CONTINUE
K=0
NN = SPOINT - 1
DO 20 J = 1,NN
K = K + 1
FDATA(K) = ORD(J)
XDATA(K) = ABSC(J)
20 CONTINUE
MM = EPOINT + 1
DO 30 J = MM,N
K = K + 1
FDATA(K) = ORD(J)
XDATA(K) = ABSC(J)
30 CONTINUE
KK = 0
DO 40 J = SPOINT,EPOINT
KK = KK + 1
ABSC2(KK) = ABSC(J)
ORD2(KK) = ORD(J)
40 CONTINUE

```

```

C   DRUK OPSKRIFTE
C
C   WRITE(2,120)
C   WRITE(2,110)
C
C   BEREKEN KUBIESE SPLINE INTERPOLANT (AKIMA)
C
C   CALL CSAKM(NDATA,XDATA,FDATA,BREAK,CSCOE)
C
C   DRUK INTERPOLANTE EN FOUTE BY EVALUASIE PUNTE
C
C   ER = 0.0
C   DO 60 J = 1,NMIS
C   X = ABSC2(J)
C   Y = ORD2(J)
C   BT = Y - CSVAL(X,NINTV,BREAK,CSCOE)
C   WRITE(2,'(3F15.3,F15.6)') X,Y,CSVAL(X,NINTV,BREAK,CSCOE),
1   BT
C   ER = ER + BT**2
60 CONTINUE
C   WRITE(2,190) ER
C
C   DRUK OPSKRIFTE
C
C   WRITE(2,140)
C   WRITE(2,*)' ORDER = ',KORDER
C   WRITE(2,110)
C
C   GENEREER KNOOP REEKS
C
C   CALL BSNAM(NDATA,XDATA,KORDER,XKNOT)
C
C   INTERPOLEER (B-SPLINE)
C
C   CALL B2INT(NDATA,XDATA,FDATA,KORDER,XKNOT,BSCOE,WK1,WK2,WK3,IWK)
C
C   DRUK INTERPOLANTE EN FOUTE BY EVALUASIE PUNTE
C
C   ER = 0.0
C   DO 70 J = 1,NMIS
C   X = ABSC2(J)
C   Y = ORD2(J)
C   BT = Y - BSVAL(X,KORDER,XKNOT,NDATA,BSCOE)
C   WRITE(2,'(3F15.3,F15.6)') X,Y,BSVAL(X,KORDER,XKNOT,NDATA,BSCOE),
1   BT
C   ER = ER + BT**2
70 CONTINUE
C   WRITE(2,190) ER
C
C   COMPUTATION OF THE VARIABLE KNOT B-SPLINE LEAST SQUARES APPROX TO
C   GIVEN DATA
C   DRUK OPSKRIFTE
C
C   WRITE(2,150)
C   WRITE(2,*)' ORDER OF LEAST SQUARES = ',KORD2
C   WRITE(2,170)
C
C   GENEREER KNOOP RY
C
C   CALL BSNAM(NDATA,XDATA,KORD2,XGUESS)
C
C   SKEP VEKTOR WAT GEWIGTE BEVAT
C
C   DO 200 J = 1,NDATA
C   WEIGHT(J) = 1.0
200 CONTINUE
C

```

```

C      COMPUTE LEAST SQUARES B-SPLINE REPRESENTATION WITH KORD2,NCOEF ANC
C      XGUESS
C
C      CALL B2VLS(NDATA,XDATA,FDATA,WEIGHT,KORD2,NCOEF,XGUESS,XKNOT2
1      ,BSCOF,SSQ,IWK,WK1)
C
C      PRINT B-SPLINE REPRESENTATION
C
C      WRITE(2,180)SSQ
C      ER = 0.0
C      DO 61 J = 1,NMIS
C      X = ABSC2(J)
C      Y = ORD2(J)
C      BT = Y - BSVAL(X,KORD2,XKNOT2,NCOEF,BSCOF)
C      WRITE(2,'(3F15.3,F15.6)') X,Y,BSVAL(X,KORD2,XKNOT2,NCOEF,BSCOF),
&      BT
C      ER = ER + BT**2
61  CONTINUE
C      WRITE(2,190) ER
C
C      VERANDER INVOER VAN KNOOPPUNTE XGUESS, SODAT XGUES2 SLEGS ELKE
C      TWEEDE VAN DIE OORSPRONKLIKE KNOPE BEVAT + KORD2 KNOPE
C
C      DRUK OPSKRIFTE
C
C      WRITE(2,160)
C      WRITE(2,170)
C
C      BEREKEN NUWE NCOEF EN SKEP VEKTOR WAR SLEGS ELKE TWEEDE KNOOP BEVA
C
C      S = NDATA/2.
C      IF(S.EQ.0.0) THEN
C      NCO = INT(NDATA/2)
C      K = 0
C      DO 700 M = 2,NDATA,2
C      K = K + 1
C      XGUES2(K) = XDATA(M)
700  CONTINUE
C      ELSE
C      NCO = INT(NDATA/2) + 1
C      K = 0
C      DO 999 J = 1,NDATA,2
C      K = K + 1
C      XGUES2(K) = XDATA(J)
999  CONTINUE
C      ENDIF
C
C      GENEREER KNOOPPUNTE
C
C      CALL BSNK(NCO,XGUES2,KORD2,XGUES3)
C
C      ROEP WEER SUBROETINE BSVLS
C
C      CALL B2VLS(NDATA,XDATA,FDATA,WEIGHT,KORD2,NCO,XGUES3,XKNOT3
&      ,BSCOF2,SSQ,IWK,WK1)
C      WRITE(2,180)SSQ
C      ER = 0.0
C      DO 62 J = 1,NMIS
C      X = ABSC2(J)
C      Y = ORD2(J)
C      BT = Y - BSVAL(X,KORD2,XKNOT3,NCO,BSCOF2)
C      WRITE(2,'(3F15.3,F15.6)') X,Y,BSVAL(X,KORD2,XKNOT3,NCO,BSCOF2),
1      BT
C      ER = ER + BT**2
62  CONTINUE
C      WRITE(2,190) ER
C      STOP
C      END

```

## HOOFSTUK 6

### GEVOLGTREKKINGS EN VOORSTELLE VIR VERDERE STUDIE

Maksimum aanneemlikheids beramers vir 'n outoregressiefbewegende model kan doeltreffend bereken word deur gebruik te maak van die toestandruimte benadering en Kalman rekursies. Met behulp van die voorspellings-fout-ontbinding van die aanneemlikheidsfunksie kan klein gapings van verlore data suksesvol hanteer word.

Die toestandruimte benadering is getoets vir 'n ARMA(1,1)-proses en die gedrag van hierdie benadering vir hoer orde modelle kan nog nagevors word. Waar die EXPAND-prosedure in SAS beskikbaar is, kan die resultate daarteen vergelyk word.

Die metode is toegepas op 'n eenveranderlike tydreeks. Die toepassingsmoontlikhede in die gebied van meerveranderlike tydreekse met verlore waarnemings kan nog aandag geniet.

Deurgaans is veronderstel dat die aantal verlore data min genoeg is, en die hoeveelheid beskikbare data voldoende is om 'n model te identifiseer. Sou daar egter baie gapings van verlore data met min beskikbare data wees, sal van ander metodes gebruik gemaak moet word en bied dit weer 'n aparte studieveld.

# Hantering van tydreekse met verlore waarnemings met behulp van die toestandruimte benadering en die Kalmanfilter

deur

Elizabeth M Basson

Leier: Professor S.H.C. Du Toit

Departement Statistiek

Voorgelê ter vervulling van 'n deel van die vereistes vir die graad  
Magister in Wiskundige Statistiek

## SAMEVATTING

Die probleem is om die parameters van 'n outoregressiefbewegende gemiddelde (ARMA) proses, wat 'n tydreeks met verlore data beskryf, te beraam. 'n Metode om die probleem op te los is om gebruik te maak van die toestandruimte benadering van 'n tydreeks. Wanneer 'n tydreeks in die vorm geskryf is, kan die Kalman rekursies gebruik word om, met behulp van die voorspellings-fout-ontbinding van die aanneemlikheidsfunksie, maksimum aanneemlikheids beraamers van die ARMA proses te vind.

Vervolgens is ander metodes wat tydreekse met verlore data ook kan hanteer, bespreek. Vier van hierdie metodes is gekies om 'n numeriese vergelyk te kan tref tussen hulle en die resultate verkry deur van die toestandruimte benadering gebruik te maak.

Daar is tot die gevolgtrekking gekom dat die verskeie tegnieke (insluitend die toestandruimte benadering) ewe goed presteer vir klein gapings en ewe sleg vir groot gapings.

**Analysis of time series with missing data using the  
state space approach and the Kalman filter**

by

Elizabeth M Basson

Supervisor: Professor S.H.C. Du Toit

Statistics Department

Submitted in fulfilment of part of the requirements for the degree of  
Master of Mathematical Statistics

**SUMMARY**

The problem of estimating the parameters of an autoregressive moving average (ARMA) process based on a time series with missing observations, is considered. This paper describes a solution of the problem by using the state space approach. The method of calculating the exact likelihood function of a ARMA time series based on the state space representation and using Kalman recursive estimation, is modified to accommodate the missing values. This is accomplished via the prediction error decomposition of the likelihood function.

Other possible methods for handling time series with missing data are discussed. Of these, four are chosen for numerical comparison of the results obtained by the state space approach.

The main conclusion that is drawn is that several techniques, including the state space approach, appear to perform equally well for shorter stretches of missing data, and equally poor for longer stretches.

## LITERATUURVERWYSINGS

- Abraham, B. (1981). Missing observations in time series.  
*Commun Statist Theory Meth* 10 (16), 1643–1653.
- Abraham en Ledolter (1983). Statistical methods for forecasting.  
*John Wiley & Sons*
- Akima, H (1970). A new method of interpolation and smooth curve fitting based on local procedures.  
*Journal of the ACM*, 17, 589–602.
- Anderson (1966). An introduction to multivariate statistical analysis  
*John Wiley & sons, Inc New York*
- Box, G.E.P. en G.M. Jenkins (1976). Time series analysis: forecasting and control.  
*Holden– Day San Francisco*
- Brockwell, P.J. en R.A. Davis (1986). Time series: theory and methods.  
*Springer– Verlag New York*
- Brubacher, S.R. en G.T. Wilson (1976). Interpolating time series with application to the estimation of holiday effects on electricity demand.  
*Journal of the Royal Statistical Society ( Applied Statistics)* 25(2), 107–116.
- Damsleth, E. (1980). Interpolating missing values in a time series.  
*Scandinavian Journal of Statistics* 7(7), 33–39.
- Ferreiro, O. (1986). Methodologies for the estimation of missing observations in time series.  
*Statistics & Probability Letters* 5 (1987) 65–69.
- Galpin, J.S. en B. Basson (1990). Some aspects of analysing irregularly spaced time dependent data  
*South African Journal of Science* 86, 458–461.

Gardner, G., Harvey, A.C. en G.D.A. Phillips (1980). An algorithm for exact maximum likelihood estimation of autoregressive–moving average models by means of Kalman filter.

*Applied Statistics* 29, 311–322.

Harrison, P.J. en C.F. Stevens (1976). Bayesian forecasting.

*Journal of the Royal Statistical Society, Series B*, 38, 205–247.

Harvey, A.C. (1981). Time series models.

*Philip Allan Oxford*

Harvey, A.C en C.R. McKenzie (1983). Missing observations in dynamic econometric models: a partial synthesis.

*Time series analysis of irregularly observed data, Proceedings, College Station*

Harvey, A.C. en R.G. Pierse (1984). Estimating missing observations in economic time series

*J. Amer. Statist. Ass.*, 79, 125–131.

Jazwinski (1970). Stochastic processes and filtering theory.

*Academic Press, New York*

Jones, R.H. (1980). Maximum likelihood fitting of ARMA models to time series with missing observations.

*Technometrics* 22(3), 389–395.

Kalman, R.E. (1960). A new approach to linear filtering and prediction problems.

*Transactions A.S.M.E. Journal of Basic Engineering*, 82, 389–395.

Kalman, R.E. en R.S. Bucy (1961). New results in linear filtering and prediction theory.

*Transactions A.S.M.E. J of Basic Engineering*, 83, 95–108.

McLeod, I.A., Hipel, K.W. en F. Comancho (1983). Trend assessment of water quality time series.

*Water resources bulletin, Vol 19, no 4*, 537–547.

Roode, D.L.(1987). 'n Oorsig oor diskrete en kontinue tydreeks en die gebruik van die maksimum—aanneemlikheidsmetode in die ontleding daarvan.

( *M.Sc—verhandeling , Universiteit van Pretoria* )

Sargan en Drettakis (1983). In time series analysis of irregularly observed data.

*Proceedings, College Stasios, Ed D. Brillinger, S. Fienberg, J. Gani, J. Hartigan, J. Krikberg*

Shumway, R.H.(1985). Deconvolution of multiple time series.

*Technometrics* 4,385—393

Shumway, R. H. en D. S. Stoffer (1982). An approach to time series smoothing and forecasting using the EM algorithm.

*J. Time Series Anal* 3, 253—264.

Swart, E.M. (1989). Analysis of water quality time series data from the Vaal River.

*Internal report, CSIR, CROM.89/35.*

Theil, H. (1971). Principles of Econometrics.

*John Wiley, New York*

Wincek. M.A. en C.R. Reinsel (1986). An exact maximum likelihood estimation procedure for regression ARMA time series models with possible nonconsecutive data.

*J. R. Statist Soc B*(1986), 48, no 3, 303—313.