RIVM rapport 550002002/2002

Detectie van milieuveranderingen
Een toepassing van Structurele Tijdreeksmodellen
en het Kalmanfilter

H. Visser

Dit onderzoek werd verricht in opdracht en ten laste van MAP-SOR, in het kader van project S/550002/01/TO, Tools ten behoeve van Onzekerheidsanalyse.

RIVM, Postbus 1, 3720 BA Bilthoven, telefoon: 030 - 274 91 11; fax: 030 - 274 29 71

Abstract

Are there significant trends in temperatures and precipitation over the past hundred years? And show these series some cyclic behaviour corresponding to sun spot numbers? Or, can we detect significant downward trends in concentrations of Particulate Matter? And what is the role of meteorological conditions? Are observed trends due to reduced emissions?

In this report we describe a generic statistical tool dealing with these type of questions. The technique for analysing environmental time series is based on *Structural Time Series Analysis and the Kalman filter*. These techniques are well-known in fields as Econometrics and Signal processing and Control, but are relatively unknown in Environmental research. Structural Time-Series models can be seen as a modular 'toolkit': we can estimate trends, cycles and the influence of explanatory variables (also called 'regressors' or 'predictors'). Also combinations of these components can be chosen. Moreover, confidence limits are given for all estimation results.

The associated software is called *TrendSpotter* and has been made available for both UNIX and PC. Early versions of TrendSpotter were developed at KEMA, under the name *KALFIMAC*. This report gives elaborate simulated examples illustrating the unique features of the methodology. These features are (among others) (i) estimation of flexible trends with elaborate uncertainty estimates, (ii) estimation of cycles where the form of the cycle may evolve over time, and (iii) estimation of time-varying weighing factors for explanatory variables.

The modelling approach is applied to two environmental issues: (i) the estimation of trends and cycles in climatological time series, and (ii) the influence of meteorological conditions to concentrations of Particulate Matter (PM_{10}). The former issue has great relevance in the light of greenhouse-gas-induced climate change. The latter issue has great policy relevance due to the hypothesized link between policy-driven emission reductions and corresponding trends in concentrations.

Voorwoord

De auteur wil Arthur Beusen (RIVM) bedanken voor het transfromeren van de oude UNIX-software (KALFIMAC) naar een nog beter werkende PC-versie (TrendSpotter). Verder is het rapport, en op onderdelen de software, sterk verbeterd dankzij het gedegen commentaar van Peter Heuberger, Peter Janssen en Anton van der Giessen (allen RIVM).

De auteur wil KEMA Power Generation and Sustainables (KPS) bedanken voor de toestemming om de software aan te passen en de mogelijkheden ervan uit te breiden.

Inhoud

Samenvatting

Onderzoek aan milieuproblemen levert een grote variëteit aan meetreeksen. Bij analyse van deze gegevens komen onder andere de volgende vragen naar voren:
- is er sprake van een trend in de data en is de stijging of daling statistisch significant?
- zijn er cyclische signalen aanwezig en hoe zien die er uit?
- wat is de invloed van externe variabelen op de metingen, zoals bijvoorbeeld meteorologische omstandigheden of economische ontwikkelingen?
- wat te doen met ontbrekende waarnemingen?
- hoe kunnen we voorspellingen genereren, en hoe (on)zeker zijn deze?

Antwoorden op deze vragen kunnen gevonden worden door modellering van de relevante fysische, chemische, biologische of meteorologische relaties ('white box modelling'), of met statistische data-georiënteerde modellen ('black box modelling'). Ook tussenvormen zijn mogelijk ('grey box modelling').

In dit rapport wordt de beschrijving gegeven van een *generieke statistische techniek* waarmee antwoord kan worden gegeven op bovenstaande vragen. De methode is gebaseerd op **Structurele Tijdreeksmodellen** en het **Kalmanfilter**. Het bijbehorende softwarepakket heet **TrendSpotter,** en is recentelijk beschikbaar gekomen voor toepassing op PC's. De methode die gebruik maakt van empirische gegevens (meetgegevens) en statistische analyses, moet niet gezien worden als een benadering die staat tegenover fysisch-georiënteerde modellen, zoals bijvoorbeeld het Euros- of het OPS-model, maar als een die juist aanvullende inzichten verschaft. Daarnaast hebben we vaak te maken met situaties waarbij nog geen adequaat fysisch-georiënteerd causaal verklarend model bestaat. Dit is bijvoorbeeld het geval bij de relatie tussen ziekenhuisopnames/mortaliteit en luchtverontreiniging. In dit soort situaties zijn statistische modellen onontbeerlijk.

TrendSpotter werd aanvankelijk ontwikkeld bij KEMA onder de naam *KALFIMAC* voor de analyse van milieu-meetreeksen. Het pakket is door RIVM aangekocht in 1996, en momenteel met toestemming van KEMA verbeterd en aangepast aan de RIVM-praktijk (implementatie op PC en gebruik van S-PLUS).

De methode heeft een aantal unieke kenmerken met hoge relevantie voor milieu-onderzoek. We noemen drie van zulke kenmerken:
- het schatten van trends met één of meer buigpunten in de tijd, waarbij steeds alle onzekerheidsinformatie beschikbaar is;
- het schatten van een cyclus waarbij de vorm van deze cylus in de tijd mag evolueren;
- het schatten van weegfactoren voor verklarende variabelen waarbij deze weegfactoren in de tijd mogen veranderen.

Essentieel bij deze punten is dat niet de modelleur maar de *methode* aangeeft of er buigpunten zijn in de trend, of dat de vorm van een cyclus al dan niet constant is in de tijd.

Het praktisch gebruik van Structurele Tijdreeksmodellen wordt in dit rapport uitgewerkt aan de hand van twee gesimuleerde tijdreeksen die samengesteld zijn uit bekende componenten (trend, cyclus, invloed van een externe factor en ruis). Hierbij worden bovengenoemde kenmerken van de methode geïllustreerd.

Een tweetal toepassingen uit de praktijk van het MilieuNatuurPlanbureau (MNP) van het RIVM worden gegeven, namelijk:
- het detecteren van klimaatveranderingen als gevolg van het broeikaseffect, en
- het schatten van meteo-correcties voor luchtverontreinigende componenten.

De analyse van honderd jaar temperatuur- en neerslagmetingen in De Bilt laat zien dat beide reeksen een significant stijgende tendens vertonen over de afgelopen 102 jaar. Echter het tempo van deze stijging verschilt. De temperatuurreeks vertoont een lichte stijging over de periode 1901-1940, stabiliseert over de periode 1941-1960, en vertoont een sterke toename vanaf 1970. Vanaf 1990 bedraagt de toename gemiddeld 0.04 ºC per jaar en in het eindjaar 2002 zelfs 0.05 ºC. De jaarlijkse toenames zijn statistisch significant vanaf het jaar 1975. Verder blijkt de trendwaarde in 2002 significant hoger dan alle trendwaarden in de voorafgaande eeuw.

Neerslag vertoont een constante trendmatige stijging over de hele periode van 102 jaar. Deze toename bedraagt 101 mm, ofwel bijna 1.0 mm per jaar. Ondanks de grote variabiliteit in de jaarlijkse neerslag is deze toename statistische significant (de totale neerslag in het jaar 1921 bedroeg nog geen 400 mm, maar in 1998 meer dan 1200 mm).

Naast jaargemiddelde temperatuur en neerslag is ook onderzocht hoe de jaarcyclus in deze reeksen verandert in de tijd. Hiertoe zijn maandgemiddelde data gebruikt over de periode 1901-2002. Veranderingen in de jaarcyclus zijn van groot belang voor plantengroei en daarmee samenhangde landbouwopbrengsten. Door de langjarige trend in de temperatuurreeks zette het groeiseizoen aan het eind van de meetreeks (1996-2002) gemiddeld twee weken eerder in dan aan het begin van de meetreeks (1901-1905). Bovendien eindigde het groeiseizoen gemiddeld ruim een week later. Hiermee wordt geconcludeerd dat het groeiseizoen in Nederland in honderd jaar tijd verlengd is met drie weken.

Deze verschuivingen zijn geheel toe te wijzen aan de trendmatige temperatuurstijging over de periode 1901-2002. Voor de jaargang van temperatuur (de cyclus uit het model) is geen vormverandering opgetreden. De neerslagreeks bevat slechts een geringe jaarcyclus en deze cyclus heeft daarom geen betekenis voor plantengroei.

De analyse van een reeks fijn-stof-metingen in Eindhoven laat zien dat een groot deel van de maand-op-maand concentratievariaties verklaard kunnen worden uit meteorologische condities. De variabelen temperatuur, neerslag en windsnelheid verklaren samen 72% van de variaties rond een dalende trend. Voor de variabelen temperatuur en neerslag bezit de relatie tot PM_{10} een niet-lineair karakter.

De berekende langjarige trend in de PM_{10}-concentraties wordt niet beïnvloed door meteorologische condities (deze condities waren opgenomen in het Structurele Tijdreeksmodel). Over de trend wordt geconcludeerd dat:
- over de meetperiode 1993-2001 de trendmatige daling statistisch significant is. De daling bedraagt 8.5 $\mu g/m^3$;
- de gemiddelde concentratie in 1993 *boven* de EU-richtlijn voor 2005 ligt. Deze concentratie ligt op de grens van statistische significantie (trendwaarde is 43 $\mu g/m^3$ en richtlijn is 40 $\mu g/m^3$). In 2001 is de gemiddelde concentratie significant gedaald tot ruim *onder* de richtlijn (trendwaarde is 34.5 $\mu g/m^3$).

1. INLEIDING

1.1 Statistiek versus fysica?

Binnen het RIVM MilieuNatuurPlanbureau (MNP) wordt een grote verscheidenheid aan wiskundige modellen toegepast. Deze modellen zijn gebaseerd op fysische, chemische, meteorologische dan wel biologische relaties. Met deze modellen wordt de werkelijke situatie zo goed mogelijk benaderd. Wat 'zo goed mogelijk' betekent, kan worden geverifieerd aan de hand van milieumetingen.

Deze modelmatige benadering zullen we hierna aanduiden met 'white box modelling'. In veel gevallen zal *white box modelling* een deterministische benadering geven van de werkelijkheid. Een voorbeeld is het OPS-model (Jaarsveld, 1995) waarmee verspreiding van een aantal luchtverontreinigende componenten wordt beschreven als functie van emissies en meteorologie.

Tegenover *white box modelling* staat de zogenaamde *black box modelling*, het modelleren van meetreeksen op basis van statistische principes. Het te analyseren signaal wordt opgevat als de som van een deteministisch signaal en een stochastisch restsignaal, de 'ruis'. Een meetreeks kan daarmee gezien worden als een *mogelijke realisatie* van de werkelijkheid. Iets andere uitkomsten waren dus ook mogelijk geweest. We duiden in deze filosofie een meetreeks aan met de term 'tijdreeks'.

Binnen de statistische benadering kunnen relaties (associaties) opgespoord worden via berekening van correlaties. Correlaties wijzen op overeenkomst in patronen, maar bewijzen geen oorzakelijk verband. Vandaar de benaming 'black box'. Voorbeelden van *black box modelling* zijn Multiple regressiemodellen, ARIMA-modellen of de in dit rapport toegepaste Structurele Tijdreeksmodellen (zie resp. Montgomery en Peck (1983), Box en Jenkins (1976) en Harvey (1989)).

Er bestaat ook een mengvorm tussen *white box modelling* en *black box modelling*, namelijk *grey box modelling*. Als een fysisch model slechts een deel van de werkelijkheid beschrijft, dan kan het verschil tussen modelvoorspellingen en werkelijkheid (de metingen) verklaard worden via statistische modellering.

Een voorbeeld is de modellering van PM_{10}-concentraties in Nederland met het OPS-model. Het OPS-model beschrijft de antropogene bijdrage aan de gemeten PM_{10}-concentraties. Maar hiermee wordt slechts de helft van de metingen verklaard. Uit onderzoek is gebleken dat het restant te verklaren is uit de bijdrage van natuurlijke bronnen (Visser, Buringh en Breugel, 2001). Men zou nu het verschil tussen metingen en OPS-voorspellingen statistisch kunnen beschrijven en voorspellen, bijvoorbeeld met geostatistische technieken als Universal Kriging. In een latere fase zou het OPS-model uitgebreid kunnen worden door verdiscontering van natuurlijke bronnen.

Als men niet de beschikking heeft over een fysisch model, dan is het duidelijk dat statistische modellen een belangrijke rol spelen. Maar ook als er wèl een fysisch model aanwezig is, blijft statistische modellering van belang. Statistiek is bij uitstek geschikt als een 'diagnostic tool' om oorzaken dan wel mogelijke verklaringen aan te dragen voor verschillen tussen metingen de voorspellingen van fysisch-georiënteerde milieumodellen.

Een eerste voorbeeld is het corrigeren van milieumetingen voor meteorologische condities. Dekkers en Noordijk (1997), en Visser en Noordijk (2002) beschrijven een generieke methode gebaseerd op Regressieboom-Analyse waarmee tijdreeksen van luchtverontreinigende componenten gecorrigeerd kunnen worden voor meteorologie. Zo'n correctie is van groot belang voor het beleid omdat na correctie met meer zekerheid uitspraken kunnen worden gedaan over het doorwerken van emissiereducties op heersende concentraties. Zie ook hoofdstuk 7 van dit rapport voor een meteo-correctie met Structurele Tijdreeksmodellen. Meteo-correcties kunnen met de huidige versies van het OPS- en EUROS-model niet uitgevoerd worden. Beide modellen kunnen zo'n correctie in principe wel uitvoeren, maar dit zou een aantal aanpassingen in de software vergen.

Bovenstaand voorbeeld geeft ook aan hoe statistiek inzichten verschaft die later in fysische relaties kunnen worden omgezet. Visser en Noordijk (2002) geven een voorbeeld voor PM_{10}-concentraties. Zij vonden dat maanden met extreme droogte en koude corresponderen met maanden met zeer hoge PM_{10}-concentraties. In dit soort situaties blijkt opwaaiend stof uit Nederland of vanover de grens uit Duitsland een onverdacht belangrijke *natuurlijke* bron van stof te zijn. De winterperiode is van belang omdat landerijen er dan kaal bij liggen. De lange droogte zorgt dan voor uitdroging van de bodem. Dit soort relaties zijn nog niet in een verklarend model voor PM_{10} verwerkt.

Een tweede voorbeeld is het doen van voorspellingen. Voor componenten als PM_{10} en O_3 moeten vanwege smogvoorspellingen op Teletekst voorspellingen worden gedaan, één en twee dagen vooruit. Op dit terrein is al vaak aangetoond dat eenvoudige statistische modellen evengoed of zelfs beter voorspellen dan ingewikkelde 'fysische' modellen.

Omdat binnen de MNP-praktijk *white box modelling* sterk ontwikkeld is, maar *black box modelling* juist niet, worden binnen het projekt Tools ten behoeve van Onzekerheidsanalyse (nr. S/550002/01/TO) methodes ontwikkeld om deze lacune op te vullen. Dit rapport beschrijft één van zulke tools, namelijk het softwarepakket TrendSpotter.

1.2 Structurele tijdreeksmodellen, een nieuwe trend

Onderzoek aan milieuproblemen levert een grote variëteit aan meetreeksen. Belangrijke vragen zijn: (i) zijn er trends aanwezig in de data, (ii) zijn er cycli aanwezig, (iii) bestaat er een eenvoudig verklarend verband tussen de verschillende gemeten grootheden, (iv) wat te doen met ontbrekende waarnemingen en (v) hoe goed kan de toekomst voorspeld worden?

Een voorbeeld is de analyse van fijn-stof-metingen (PM_{10}) in relatie tot meteorologie. Het RIVM beschikt over een meetnet van 19 stations verspreid over Nederland, waar op uurbasis PM_{10}-concentraties worden gemeten. Twaalf van deze 19 stations functioneren sinds 1992. We zouden nu graag willen weten

- wat de langjarige trend in de concentraties is (vraag (i));
- of meteorologie invloed heeft op deze trend en wellicht de invloed van emissiereducties maskeert (vraag (iii));
- hoe het dagverloop van concentraties eruit ziet (vraag (ii));
- of er een significante weekcyclus aanwezig is (voor stadstations geldt bijvoorbeeld dat er minder verkeeremissies zijn in het weekend dan door-de-week);
- of er een jaarcyclus aanwezig is (vraag (ii));
- of we ontbrekende uur- en dagwaarden mogen weglaten of dat we ze zo goed mogelijk moeten reconstrueren door middel van interpolatie (vraag (iv));
- hoe nauwkeurig we PM_{10}-concentraties een en twee dagen vooruit kunnen voorspellen (smogvoorspellingen op Teletext)? En wat is dan de voorspelnauwkeurigheid (vraag (v))?

Om bovenstaande vragen te beantwoorden, wordt gebruik gemaakt van tijdreeksanalyse. Een meetreeks wordt hierin opgevat als een realisatie van een kansproces dat bestaat uit een deterministisch signaal, bijvoorbeeld een trend, en een stochastisch ruisproces. Bekende technieken uit de tijdreeksanalyse zijn Multiple regressiemodellen en ARIMA-modellen.

In dit rapport wordt het softwarepakket TrendSpotter beschreven dat gebaseerd is op een andere aanpak, namelijk die van *Structurele Tijdreeksmodellen*. Dit type modellen is populair geworden in de econometrie door toedoen van Harvey (1984, 1989, 1993) en later Koopmans (2001, en http://staff.feweb.vu.nl/koopman/rede.pdf). Buiten de econometrie zijn deze modellen weinig toegepast, maar zeer ten onrechte!

Het voordeel van structurele tijdreeksmodellen is dat componenten als een trend, een cyclisch signaal en de invloed van verklarende variabelen additief gemodelleerd worden en als zodanig rechtstreeks beschikbaar komen. Bij ARIMA-modellen is dit bijvoorbeeld niet het geval.

Verder mogen parameters in het model veranderen over de tijd. Hiermee wordt bedoeld dat het regressiemodel zoals dat hier gebruikt wordt, tijdsafhankelijke weegfactoren c.q. parameters bevat. Om het eerder genoemde voorbeeld te volgen, het zou kunnen zijn dat de invloed van meteorologie op PM_{10}-concentraties niet constant is over een groot aantal jaren. Als de nadruk bij de emissies van PM_{10} verschuift van locale bronnen naar buitenlandse bronnen, dan vervalt bijvoorbeeld de invloed van menghoogte-variaties.

Literatuurverwijzingen naar toepassing van TrendSpotter op milieu-thema's zijn: Brakel en Visser (1996), Van der Wal en Janssen (1996, 2000), Visser (1994, 1995, 2000), Visser en Molenaar (1995), Visser et al. (1999), Visser en De Koningh (2000), en Visser en Rőmer (1999, 2000).

1.3 Ontstaansgeschiedenis software

De eerste opzet van het hier beschreven softwarepakket TrendSpotter werd gemaakt door H. Visser en J. Molenaar (T.U. Eindhoven) in 1984. De eerste algemeen bruikbare versie werd geprogrammeerd door M.A.C. Mettes (HTS Arnhem) onder de naam **KALFIMAC**. Zijn initialen vormden de laatste drie letters van de naam van het pakket. Een tweede release werd gemaakt door R. Leene (HTS Arnhem). Hij breidde het aantal opties van het pakket uit en maakte een 'turbo-versie' voor de optimalisatie-routine. Het pakket werd een factor 2 tot 10 sneller. De derde release werd gemaakt door M. Habets (HTS Heerlen) en bevatte het overzetten van de software van de KEMA mainframe, een UNIVAC 1100, naar een APOLLO-netwerk. Voor een beschrijving van release 3.1 zie Visser, Habets en Leene (1990). W.C.A. Maas en K. Friso breidden de APOLLO-versie uit met een tweetal routines om variabelen te selecteren in de context van regressiemodellen met tijdsafhankelijke parameters. Dit leidde tot release 4.0.

Bij de overgang van release 4.0 naar 5.0 zijn een drietal wijzigingen doorgevoerd. In de eerste plaats is het pakket overgezet van APOLLO-UNIX naar HPUX. Ten tweede is het plotpakket SIMPLEPLOT vervangen door het interactieve plotpakket UNIGRAPH. Ten derde zijn de mogelijkheden van trendschatting uitgebreid met ARIMA-modellen. De aanpassingen zijn uitgevoerd door A. Binzer van de Deense Technische Universiteit in Lyngby.

De eerste versie van TrendSpotter is ontwikkeld op een UNIVAC-1100-systeem, onder de naam KALFIMAC. De batch-achtige opbouw van input- en outputfiles herinnert hier nog aan. Een overzetting van KALFIMAC naar PC is de laatste ontwikkeling. De PC-software heet TrendSpotter. *Foto: UNISYS Corporation.*

Bij de overgang van release 4.0 naar 5.0 zijn een drietal wijzigingen doorgevoerd. In de eerste plaats is het pakket overgezet van APOLLO-UNIX naar HPUX. Ten tweede is het plotpakket SIMPLEPLOT vervangen door het interactieve plotpakket UNIGRAPH. Ten derde zijn de mogelijkheden van trendschatting uitgebreid met ARIMA-modellen. De aanpassingen zijn uitgevoerd door A. Binzer van de Deense Technische Universiteit in Lyngby.

De huidige release 6.0 van mei 2002 is speciaal voor RIVM aangepast (RIVM kocht het pakket van KEMA in 1996). Overtollige statements uit de oude versies zijn weggelaten. Mede hierdoor is de file met opties compacter en daardoor duidelijker geworden. Verder is het selecteren van regressoren binnen de context van tijdsafhankelijke weegfactoren in de versie 6.0 niet meegenomen (de selectieprocedures zijn niet voldoende uitgekristalliseerd).

Als laatste stap in de ontwikkeling van het softwarepakket is een PC-versie ontwikkeld voor KALFIMAC door A. Beusen (CIM, RIVM). De PC-versie heeft de naam **TrendSpotter** gekregen.

De ontwikkeling van relase 6.0 (UNIX) en de PC-versie, release 1.0, is uitgevoerd in het kader van het deelproject Tools ten behoeve van Onzekerheidsanalyse (S/550002/01/TO).

1.4 Leeswijzer

Dit rapport kan op drie niveau's van detaillering gelezen worden.

1. Het eerste niveau is voor wie alleen in de toepassingsmogelijkheden van de gepresenteerde tijdreeksmodellen geïnteresseerd is. Op dit leesniveau is jargon zoveel mogelijk vermeden. **Hoofdstuk 2** geeft een globale indruk van de methode en laat zien welk type problemen gemodelleerd kan worden. Aan bod komen de verschillende componenten 'trend', 'cyclus' en 'verklarende variabelen'. Twee voorbeelden uit de praktijk van het MNP worden gegeven in de **Hoofdstukken 6 en 7**.

   Hoofdstuk 6 beschrijft trends en cycli in klimatologisch meetreeksen. Deze trends en cycli hebben grote ecologische en economische gevolgen. Hoofdstuk 7 geeft een voorbeeld van een actueel beleidsrelevant probleem: in hoeverre kunnen concentraties van luchtverontreinigende stoffen verklaard worden uit meteorologische variabiliteit? Als fijn-stof-concentraties dalen, komt dat dan door dalende emissies of is er ook een samenhang met een veranderd klimaat in Nederland?

2. Het tweede niveau is voor diegenen die ook zelf trends in data willen gaan schatten met de software. Om het inzicht te verhogen geeft **Hoofdstuk 3** een korte beschrijving van de schattingsmethode, het Kalmanfilter. Jargon is hier zoveel mogelijk vermeden. **Hoofdstuk 4** gaat daarna in op enkele aspecten van de bij het RIVM ontwikkelde software, genaamd TrendSpotter. Om de mogelijkheden van de methode en de software te verkennen, is in **Hoofdstuk 5** een simulatie uitgewerkt. Aan de hand van pseudoreeksen worden alle mogelijkheden van de methode en software geïllustreerd. Optiefiles van TrendSpotter zijn gegeven in **Appendix B**. Pseudoreeksen zijn reeksen die samengesteld zijn uit zelfgekozen componenten 'trend', 'cyclus' en 'verklarende variabelen'.

3. Het derde niveau is voor de specialisten die geïnteresseerd zijn in de exacte wiskundig formuleringen. Alle Structurele Tijdreeksmodellen uit dit rapport zijn wiskundig uitgeschreven in **Appendix A.3**. Het volledige Kalmanfilter-algoritme is gegeven in **Appendix A.4**.

In het vervolg van het rapport komt de kleurencode groen, blauw en rood terug in paragraaftitels, figuuronderschriften en tabelbovenschriften.

2. STRUCTURELE TIJDREEKSMODELLEN

2.1 Algemeen

Structurele modellen bezitten een modulaire opbouw. Het model voor een meetreeks y_t kan gezien worden als een optelling van vier componenten:

$$y_t = trend_t + cyclus_t + invloed\ verklarende\ variabelen + ruis_t \qquad (1a)$$

De index t geeft de tijd aan, oplopend van 1 tot en met N. Model (1a) is additief. Als de componenten multiplicatief zijn, dus

$$y_t = trend_t * cyclus_t * invloed\ verklarende\ variabelen_t * ruis_t \qquad (1b)$$

dan krijgen we de vorm (1a) door het nemen van de logaritme van y_t. Als y_t niet overal positief is, dan moet y_t verhoogd worden met een constante.

Een structureel model bestaat uit één of meer van deze componenten, al naar gelang de toepassing. De eenvoudigste vorm van model (1) ontstaat als de componenten deterministisch zijn. In dat geval heeft model (1) de vorm van een Multiple Regressiemodel:

$$y_t = \mu_t + \sum_{i=1}^{S} \gamma_i d_{i,t} + \sum_{k=1}^{M} \delta_k \cdot x_{k,t} + \xi_t, \quad t = 1,...,N \qquad (2a)$$

De grootheid μ_t is een constante of lineaire trend:

$$\mu_t = \alpha + \beta * t \qquad (2b)$$

De eerste sommatie staat voor de bijdrage van een cyclisch signaal met een vaste periode lengte S en wordt gemodelleerd door S dummy-variabelen $d_{1,t}$ . ...,$d_{S,t}$ te definiëren (deze variabelen zijn 0 of 1): $\gamma_t = \gamma_1 d_{1,t} + ... + \gamma_S d_{S,t}$. Bijvoorbeeld, als y_t een variabele is met een weekcyclus (dus S = 7), dan is de dummy variabele $d_{1,t}$ 1 op alle maandagen en verder nul. Dummy $d_{2,t}$ is 1 op alle dinsdagen en verder nul, etc. Bovendien geldt:

$$\sum_{i=1}^{S} \gamma_{t-i+1} = 0 \qquad (2c)$$

Verklarende variabelen worden toegevoegd in de vorm van een Multiple Regressiemodel waarbij de weegfactoren δ_i staan voor de regressiecoëfficiënten. Het niet door het model verklaarde gedeelte wordt gerepresenteerd door het ruisproces ξ_t. Dit stochastische proces wordt meestal verondersteld normaal-verdeeld te zijn met opeenvolgende waarden die onderling ongecorreleerd zijn.

Door het toevoegen van 'ruis' aan de componenten in model (2) wordt het model flexibeler. De lineaire trend kan buigpunten vertonen, het cyclisch signaal kan van periode op periode langzaam van vorm veranderen, en het Regressiemodel krijgt tijdsafhankelijke coëfficiënten:

$$y_t = \mu_t + \gamma_t + \sum_{k=1}^{M} \delta_{k,t} \cdot x_{k,t} + \xi_t, \quad t = 1,..., N \tag{3a}$$

$$\mu_t = \alpha_t + \beta_t * t \tag{3b}$$

$$\gamma_t = \gamma_{1,t} d_{1,t} + ... + \gamma_{S,t} d_{S,t} \tag{3c}$$

Elke component (trend, cyclus of verklarende variabele(n)) kan afzonderlijk of in combinatie met andere geschat worden. De afzonderlijke componenten worden in de volgende paragrafen kort toegelicht.

Onder omstandigheden kunnen bij het schatten van model (3) *identificatieproblemen* optreden. Bijvoorbeeld, we schatten een model met een trend μ_t en een verklarende variabele $x_{i,t}$ die *zelf* ook een trend in de tijd vertoont. Het is nu niet eenduidig hoe de trend in de metingen y_t beschreven moet worden: uit μ_t, uit de trend in $x_{i,t}$, of uit een mengvorm van beide (zie verder §3.4).

2.2 Trends

Een trend staat voor het laag-frequente gedeelte in een meetreeks. Zoals vermeld, dit kan een constante of een lineaire trend zijn, maar ook een meer flexibele trend. De laatste situatie komt in de milieupraktijk vaak voor en kan via Structurele Tijdreeksmodellen goed geschat worden. Daarbij kunnen via deze modellen uitspraken gedaan worden over het significant overschrijden van norm- of drempelwaarden, en of de daling of stijging over langere perioden statistisch significant is. Deze laatste uitspraak is een unieke mogelijkheid van de gekozen modellen in geval van trends met één of meerdere buigpunten.

We onderscheiden vier trendmodellen:
- het Stochastic Level (SL) niveaumodel. Dit model bestaat uit een variabel niveau of drift, welke, afhankelijk van het karakter van de data, kan reduceren tot een constante. Het model heeft één instelparameter die maatgevend is voor de toegelaten variabiliteit van de drift;
- het Local Level (LL) trendmodel. Dit model wordt vaak in de literatuur vermeld. Het model heeft twee instelparameters. Het op nul stellen van deze parameters geeft een lineaire trendschatting;
- het Doubly Differenced (DD) trendmodel. Dit trendmodel heeft één instelparameter en voldoet in de meeste praktische situaties. Het op nul stellen van de instelparameter reduceert het trendmodel tot een lineaire trend;
- het ARIMA-trendmodel. Dit trendmodel heeft één ruisfactor en heeft enkele specialistische toepassingen. Het kan gebruikt worden wanneer naast een trend gecorreleerde ruis aanwezig is in de metingen.

Een keuze uit één van deze vier trendmodellen kan gemaakt worden door het karakter van het laagfrequente deel van de metingen y_t te bepalen. Voor details over het kiezen van een trendmodel zie **Appendix A.3.1**.

Een uniek kenmerk van trendschattingen met Structurele Tijdreeksmodellen is dat elke vorm van onzekerheidsinformatie gegeven kan worden over trends, ook wanneer de trend buigpunten vertoont.

2.3 Cyclus

Cyclische signalen kunnen uit een meetreeks gefilterd worden door het opnemen van een cyclische component in het model. De periodelengte moet constant zijn, terwijl slechts één cyclisch signaal tegelijk geschat kan worden. Het periodieke signaal mag elke willekeurige vorm hebben, en kan in meer of mindere mate van vorm veranderen door toevoegen van een ruisbron. De modellering wijkt af van Fourier-decompositie, waar een signaal ontbonden wordt in een aantal sinusfuncties.

Wanneer de periodelengte groot is, bijvoorbeeld groter dan 12 tijdseenheden, dan neemt de rekentijd van de software sterk toe, dit omdat de periodelengte tevens de dimensie van matrices en vectoren in het model bepaalt. Zo zal het op een doorsnee PC niet lukken om in een reeks van uurlijke concentraties een weekcyclus te schatten. De periodelengte is dan 168 uur en leidt tot te lange rekentijden. Wil men toch een weekcylus schatten, dan moet men eerst uurwaarden middelen tot dagwaarden. Voor deze nieuwe data bedraagt de periodelengte 7 en vormt geen probleem.

2.4 Verklarende variabelen

Een regressiemodel kan gekozen worden door het toevoegen van de regressiecomponent. Het bekende Multiple Regressiemodel ontstaat wanneer het SL-trendmodel wordt gekozen zonder ruis, ofwel een constante, en het regressiegedeelte zonder ruis per weegfactor:

$$y_t = \delta_0 + \delta_1 x_{1,t} + \ldots + \delta_M x_{M,t} \qquad (4)$$

Voor een beschrijving van dit model zie bijvoorbeeld Montgomery en Peck (1982).

Meer flexibiliteit ontstaat door toevoeging van ruisfactoren aan de individuele weegfactoren uit het regressiemodel. Er ontstaat nu een regressiemodel met tijdsafhankelijke weegfactoren, zoals in (3a). Voor veel toepassingen is dit model realistischer dan het traditionele Multiple Regressiemodel dat over de hele tijdas onveranderlijk is. Zie ook hoofdstuk 7 voor een praktijkvoorbeeld.

2.5 Transformaties vooraf

Zoals vermeld in §2.1 kan het nodig zijn om data *vooraf* te transformeren. Doel kan zijn om een multiplicatief model te transformeren naar een additief model. De meest gebruikte transformatie is de log-transformatie:

$$y_t' = \log (y_t + \text{constante}) \qquad (5)$$

De constante wordt zo gekozen dat het argument van de log nooit negatief wordt. Deze transformatie kan gekozen worden in TrendSpotter. Ook is een toets toegevoegd om vooraf te kijken of een log-transformatie nodig is. Dit geschiedt met de zogenaamde *Range-Mean plot*.

In een *Range-mean plot* wordt de tijdas in zeg 10 gelijke tijdsintervallen verdeeld. Voor elk tijdsinterval j wordt het gemiddelde m_j en standaarddeviatie S_j berekend. De *Range-mean plot* is nu niets anders dan een *scatterplot* van m_j (op de x-as) tegen S_j (op de y-as). Als er bij benadering een horizontale lijn loopt door de puntenwolk, dan is de spreiding onafhankelijk van het gemiddelde en is geen transformatie nodig. Maar bij concentraties van luchtverontreinigende stoffen bijvoorbeeld, zal S_j lineair oplopen met m_j. Dit betekent dat de spreiding een vast percentage is van het gemiddelde niveau. Voor dit soort situaties is een logaritmische transformatie zeer geschikt. Zie verder McLeod (1983, pag. 11-18 e.v.).

De opbouw van een structureel tijdreeksmodel is modulair. Men kan zelf kiezen voor een gewenste combinatie van componenten. *Foto: Legoland*

2.6 Voldoet het model?

Een belangrijke vraag na het schatten van het gekozen model is: voldoet het model wel aan de eisen? Zijn we klaar?

Een model kan om een aantal redenen niet voldoen:

- er had een transformatie vooraf moeten plaats vinden (§2.5);
- er is het verkeerde trendmodel gekozen (§2.2);
- er is geen cyclisch signaal geschat, terwijl die wel aanwezig is in de data (§2.3);
- het patroon van de reeks wordt door een externe variabele beïnvloed die niet in het model is verdisconteerd (§2.4);
- het model geeft geen eenduidige schattingen voor de gevraagde componenten (§3.4)

Er zijn een groot aantal statistische tests en *diagnostic checks* ontwikkeld om antwoorden te vinden op deze vragen. We noemen hier alleen de belangrijkste controle op het model. Dit een contole op samenhang binnen de restreeks of residureeks van het geschatte model (dus de meting y_t minus de door het model geschatte waarde op tijdstip t). Deze residureeks is belangrijk omdat hij vertelt hoe goed of hoe slecht het model de metingen heeft kunnen volgen (verklaren).

Op de residureeks berekenen we de zogenaamde autocorrelatiefunctie, afgekort als ACF. Een ACF bestaat uit een reeks van correlaties ρ_1, ρ_2, ... , ρ_M. Hierbij is ρ_1 de bekende correlatiecoëfficiënt tussen data die precies één tijdstap verwijderd zijn: (y_1,y_2), (y_2,y_3), ... , (y_{N-1}, y_N). Evenzo staat ρ_2 voor de correlatie tussen data die twee tijdstappen verschillen: (y_1, y_3), (y_2,y_4), (y_3,y_5), ..., (y_{N-2},y_N). Enzovoorts tot en met ρ_M. De *schattingen* voor ρ_1, ..., ρ_M duiden we aan met de notatie R_1, ..., R_M.

Als er in de resdureeks nog een trendmatig signaal aanwezig is, dan zal de reeks R_1, R_2, R_3, ... langzaam uitdempen. Als er in de data nog een cyclus aanwezig is, bijvoorbeeld een cyclus met periode 12, dan zien we alternerende neatieve en positieve correlaties R_6, R_{12}, R_{18}, R_{24}, ...

Voorbeelden van ACF's zijn gegeven in de figuren 7D, 8D, 11, 14 en 15 uit de hoofstukken 6 en 7.

3. KALMANFILTER

Structurele Tijdreeksmodellen worden geschat met het zogenaamde Kalmanfilter. Dit filter werd in de zestiger jaren ontwikkeld door R.E. Kalman op het gebied van de regeltechniek. Daarna is het op vele terreinen toegepast. Het filter is zo populair geworden door het recursieve karakter van het filter en de mooie statistische eigenschappen van de schatters voor onbekende parameters. Toepassing van het Kalmanfilter vereist wel dat een specifiek gekozen model in de zogenaamde *toestandsvorm* geschreven wordt.

De toestandsvorm wordt in detail in Appendix A beschreven (§ A.2), terwijl de wijze waarop Structurele Tijdreeksmodellen in deze vorm geschreven kunnen worden, uiteengezet wordt in § A.3 uit deze Appendix. De wiskundige formulering van het Kalman filter zelf wordt gegeven in § A.4. In dit hoofdstuk worden enkele aspecten van het Kalmanfilter kwalitatief beschreven.

3.1 Filteren, smoothen, voorspellen

Het Kalmanfilter werkt recursief. Dat wil zeggen dat vanaf een zeker tijdstip t een beste voorspelling wordt gemaakt voor de waarneming y_{t+1} op tijdstip t+1. Door vergelijking van de voorspelling met de werkelijk gemeten waarde van y_t stelt het filter zich in meer of mindere mate bij. Zo ontstaan voor alle waarnemingen y_t, t= 1,...,N, één-staps-voorspelfouten (ook wel aangeduid met de term *innovaties*). Het Kalmanfilter genereert modelschattingen zodanig dat de som van gekwadrateerde een-staps-voorspelfouten minimaal is. Dit recursieve proces heet filteren.

Naast schattingen voor de componenten uit het structurele model geeft het Kalmanfilter ook betrouwbaarheidsintervallen voor de trend, de cyclus en de weegfactoren uit het regressiegedeelte. De 1-σ grenzen staan voor 68%-betrouwbaarheidsintervallen, terwijl de 2-σ-grenzen staan voor 95%-betrouwbaarheidsintervallen (mits de ruis normaal-verdeeld is).

In ‘off line’ situaties, dat wil zeggen dat alle metingen voor de analyse aanwezig zijn, kunnen betere modelschattingen voor een waarneming y_t verkregen worden door het smoothen of ‘gladstrijken’ van alle waarnemingen. Dit betekent dat een voorspelling voor y_t niet alleen gebaseerd is op alle waarnemingen y_i, i = 1,...,t, maar ook op de daaropvolgende waarnemingen y_i, i = t+1,...,N. In de meeste praktische situaties zullen de modelschattingen gesmoothed worden.

Het Kalmanfilter kan ook voorspellingen genereren met betrouwbaarheidsintervallen, dus schattingen genereren voor y_t, t = N+1,N+2,...,N+L. Als er verklarende variabelen in het model aanwezig zijn, ontstaan er twee situaties: (i) de waarden van de verklarende variabelen zijn aanwezig over de voorspelperiode en kunnen dus gebruikt worden om betere voorspellingen te genereren of (ii) deze waarden zijn afwezig. Voor beide situaties kunnen met TrendSpotter schattingen gemaakt worden.

3.2 Optimaliseren van ruisvarianties

De 'flexibiliteit' van de afzonderlijke componenten wordt bepaald door de waarde van de ruisvarianties die bij die component behoren. Met flexibiliteit bedoelen we hier de overgang van model (2) naar model (3). Deze component-gerelateerde ruisvarianties kunnen in TrendSpotter met de hand gekozen worden, maar kunnen ook door het pakket geoptimaliseerd worden. In dit geval worden *maximum-likelihood-waarden* geschat.

Wel kan de rekentijd van het pakket bij optimalisatie aanzienlijk oplopen. Een Regressiemodel met 10 verklarende variabelen heeft 10 onbekende ruisvarianties en optimalisatie kan een uur of langer vergen.

Het Kalmanfilter begint te itereren vanaf t = 1. Echter, startwaarden op tijdstip t = 0 moeten dan aanwezig zijn. Aangezien schattingen voor deze waarden meestal niet voorhanden zijn, moet het filter zichzelf inregelen. De tijd hiertoe wordt de *inregeltijd* genoemd en moet aan TrendSpotter opgegeven worden.

De optimalisatie van ruisvarianties vindt plaats vanaf het moment dat het filter ingeregeld is. De inregelperiode levert namelijk geen betrouwbare schattingen. Een visuele indruk van de inregeltijd voor een specifieke toepassing kan verkregen worden door in TrendSpotter de 'filter-optie' te kiezen in plaats van direct de 'smooth-optie' toe te passen. Uit ervaring blijkt dat voor de meeste toepassingen een inregeltijd van 20 tijdstappen voldoende is. Dit betekent dus dat voor een goede schatting van de begincondities circa 20 metingen nodig zijn.

3.3 Ontbrekende waarnemingen

De situatie kan zich voordoen dat sommige metingen onbetrouwbaar of zelfs geheel afwezig zijn. In dit geval kan TrendSpotter deze waarnemingen negeren en zal zelf interpoleren tussen omringende meetwaarden. Op deze wijze wordt het schattingsproces niet verstoord. Er zijn twee manieren om ontbrekende waarden op te geven:
- een code voor elke ontbrekende of onbetrouwbare waarneming;
- het opgeven van één of meer periodes waarvoor geen waarnemingen beschikbaar zijn, of waarvoor de metingen zeer onbetrouwbaar zijn.

Rudolf E. Kalman bracht in 1961 een schok te weeg in de regeltechniek en systeemtheorie met zijn artikel 'A new approach to linear filtering and prediction problems'. Zijn benadering droeg al zeer snel zijn naam, het Kalmanfilter, en is sindsdien op zeer veel terreinen toegepast. Heel algemeen gezegd wordt zijn methode gebruikt om een optimale 'mix' te vinden van informatie uit het wiskundige model en meetresultaten. Het filter is recursief, dat wil zeggen dat op basis van het gekozen model en alle metingen tot nu toe een beste voorspelling wordt gedaan voor de volgende tijdstap. Daarna wordt deze voorspelling vergeleken met nieuwe metingen en de 'toestand' van het model wordt bijgesteld. Dit bijregelen is sterker naarmate de voorspelfout groter is.

De software kan niet corrigeren voor ontbrekende waarden in *verklarende variabelen*. Als deze toch aanwezig zijn, dan worden tijdstippen waarop één of meer verklarende variabelen ontbreken, weggelaten uit de analyse. Er ontstaan dus 'gaten' in het databestand. Dit kan hinderlijk zijn als de reeks ook cycli bevat. In dat geval moeten de gaten in de predictors vooraf 'gevuld' worden. Als bijvoorbeeld een reeks een weekcyclus bevat, en één of meer x'en hebben ontbrekende waarden op een specifieke maandag, dan wordt die maandag uit het bestand weggelaten en heeft die week daarmee slechts 6 dagen. De schatter voor de weekcyclus wordt nu sub-optimaal.

Recentelijk is een optie toegevoegd om automatisch geïnterpoleerde waarden te substitueren in de oorspronkelijke datafile. Dit kan van belang zijn als er een groot aantal metingen ontbreekt. De overige kolommen in de datafile blijven onveranderd.

3.4 Modelvalidatie

De vooronderstellingen die aan *goed-gedefinieerde* Structurele Tijdreeksmodellen ten grondslag liggen, zijn dat de ruisprocessen uit het model *witte* ruisprocessen vormen. Dit betekent dat elk ruisproces bestaat uit ongecorreleerde waarden met constante variantie in de tijd (homoscedasticiteit). Verder is een zeer prettige (maar niet noodzakelijke voorwaarde) dat de ruisprocessen normaal-verdeeld zijn. Verschillende tests op de eigenschappen van de ruisprocessen uit het model worden uitgevoerd op de zogenaamde 'gestandardiseerde innovaties' (zie vergelijking (A.31)). Deze tests worden niet binnen TrendSpotter uitgevoerd maar binnen S-PLUS middels een speciaal script (zie Visser (2002) voor details).

Voor een uitgebreidere set van tests verwijzen we naar Harvey (1989, pagina's 256-258). Harvey geeft ook een aantal tests op de residuen waarmee **misspecificatie van het gebruikte model** kan worden getest (pagina's 258 – 260). Deze tests zijn momenteel nog niet geïmplementeerd binnen S-PLUS als onderdeel van TrendSpotter.

De 'goodness of fit' van een tijdreeksmodel geeft aan hoe goed het model de data kan beschrijven en wordt meestal getest met de 'prediction error variance' die, als de tijdreeks niet te kort is, gelijk is aan de som van de gekwadrateerde innovaties (= een-stap-vooruit-voorspellingen).

In §2.1 noemden we het probleem van de identificeerbaarheid van een Structureel Tijdreeksmodel. Identificatie-problemen treden vooral in twee gevallen op:
1. verschillende predictors $x_{i,t}$ zijn onderling sterk gecorreleerd;
2. we schatten een trend μ_t, terwijl een (of meer) predictor(s) $x_{i,t}$ ook een duidelijke trend bevat(ten).

Een voorbeeld van het eerste probleem treedt op bij het schatten van een meteo-correctie op luchtverontreinigende stoffen. Als verklarende variabelen worden vaak gebruikt de (i) daggemiddelde temperatuur, (ii) de maximum temperatuur van diezelfde dag, (iii) idem de minimum temperatuur, (iv) de globale straling op die dag en (v) de relatieve vochtigheid. Als twee of meer van deze variabelen als regressors worden meegenomen in een te schatten model, dan zijn de weegfactoren $\alpha_{i,t}$ niet meer maatgevend voor de echte invloed van de betreffende x-variabelen. Immers deze x-en zijn onderlang sterk tot zeer sterk gecorreleerd.

Een voorbeeld van het tweede probleem is de analyse van ozonmetingen. Als we een meteocorrectie voor dagwaarden van ozon (= y_t) willen toepassen met de luchttemperatuur als

verklarende variabele (= x_t), dan ontstaat het probleem dat ozonconcentraties over een groot aantal jaren een neerwaardse trend vertonen, terwijl de luchttemperatuur stijgt door een record aantal warme jaren tussen 1987 en 2001. Het te schatten model 'weet' nu niet hoe de trend in yt toe te kennen: aan een aparte trend μ_t of via een weegfactor voor x_t.

Probleem 1) kunnen we signaleren door de correlatiematrix van alle predictors $x_{i,t}$ te bekijken (wordt berekend door TrendSpotter). Als twee of meer x'en hooggecorreleerd zijn, dan kunnen we er voor kiezen om een of meer predictors uit de analyse weg te laten. Een echte oplossing voor het probleem, dat ook wel aangeduid wordt met de term 'heteroscedasticiteit', is er echter niet.

Evenzo kan de covariantiematrix P bestudeerd worden (wordt gegeven in TrendSpotter voor het einde van de reeks). Grote negatieve covarianties duiden op identificatieproblemen voor de betreffende x-vairabelen.

Wat betreft het tweede probleem kan voor elke $x_{i,t}$ *vooraf* bekeken worden of er een trend in de tijd aanwezig is. Men kan er voor kiezen om deze trend *vooraf* uit $x_{i,t}$ te filteren. Deze aanpak is bijvoorbeeld gekozen in Visser en Molenaar (1992 , figuren 3 en 4).

Tot slot merken we op dat een veel toegepaste validatie-methode, cross-validatie genaamd, niet standaard in de TrendSpotter-software is opgenomen. Bij cross-validatie wordt uit de meetreeks een deel van de metingen weggelaten. Men kan bijvoorbeeld *at random* 20% van de metingen weglaten. Daarna wordt het gewenste model geschat op de resterende 80% van de dataset. Het model genereert voorspellingen voor de weggelaten 20% en de voorspellingen kunnen vergeleken worden met de werkelijke data. Door de kwadratische afwijkingen te sommeren heeft men een indruk van de voorspelkwaliteit van het model.

Hoewel deze methode niet standaard geïmplementeerd is, kan hij wel 'handmatig' uitgevoerd worden. Het verdient echter aanbeveling deze aanpak standaard op te nemen in een nieuwe *release* van TrendSpotter.

4. WERKEN MET TrendSpotter

In dit hoofdstuk geven we een korte beschrijving van het gebruik van de TrendSpotter-software. Een gedetailleerde beschrijving wordt gegeven in Visser (2003). We gaan er hierna gemakshalve vanuit dat alle relevante files voor het runnen van de software, staan in de *basis-directory* **c:/TrendSpotter/run**. In deze directory staan de executable en DLL-file voor het runnen van Trendspotter, en een executable voor het opsporen van rekenfouten in de software (een debug-versie).

4.1 Invoer en uitvoer

Het starten van van de software geschiedt door het dubbelclicken van de executable 'TrendSpotter'. Trendspotter vraag nu naar een file **par.inp** waarin de (pad)namen staan van de input- en outputfiles. Default verwacht Trendspotter de file **par.inp** in dezelfde directory als waarin de executable staat. De file **par.inp** ziet er als volgt uit:

PARAMETERS

INPUT DATA : data.km
INPUT OPTIONS : optie.km
LOGFILE : logfile
OUPUT PLOTTING DATA : writeall.out
OUTPUT COMPUTATIONS : uitvoer

Als er een padnaam ontbreekt voor de genoemde vijf files, dan wordt automatisch het basispad **c:/TrendSpotter/run/** verondersteld. Als één of meer file-namen ontbreken, dan worden bovenstaande *default* file-namen gebruikt.

Specificaties van het model staan in bovengenoemde voorbeeld in de optiefile **optie.km.** Data staan in **data.km**. De file **optie.km** bevat alle opties voor het te schatten model. Een beschrijving van deze file wordt gegeven in Visser (2003). De file **data.km** staat in ASCII-format en bevat de data in kolommen (fixed format is verplicht). De kolommen zijn steeds rechts uitgelijnd en de tijd loopt altijd verticaal naar beneden, in equidistante stappen. De eerste kolom hoeft overigens niet de tijd te zijn. Het mag ook een oplopende index zijn. Wel moet er altijd een kolom met een of andere tijdsaanduiding of index aanwezig zijn in de datafile.

Uitvoer wordt op drie niveaus aangemaakt door TrendSpotter. In de eerste plaats wordt de rekenuitvoer gezet in de file achter **'OUTPUT COMPUTATIONS'**. In deze file staan alle relevante tabellen en statistische kengetallen. Merk op dat de oude uitvoer van een vorige run overschreven wordt bij het starten van TrendSpotter, tenzij deze filenaam aangepast wordt!

In de tweede plaats wordt een file aangemaakt welke betrekking heeft op het maken van grafieken binnen S-PLUS. De data die S-PLUS nodig heeft, staan in kolomvorm in de ASCII file met *default*-naam **writeall.out**. In de derde plaats wordt een file genaamd **logfile** aangemaakt. Hierin staan meldingen van FORTRAN omtrent de vorderingen van TrendSpotter. Bekijken van deze file kan nut hebben als het programma voortijdig stopt.

4.2 Plotten van grafieken en statistiek

Het maken van grafieken op basis van de rekenuitvoer kan in principte met elk plotpakket. Alle relevante componenten als trend, trend-plus-cyclus, weegfactor voor verklarende variabelen, elk met een standaarddeviatie, staan in de file genoemd achter **'OUTPUT PLOTTING DATA'** (*default*: **writeall.out**). Omdat het handig is om statistische tests uit te voeren op de gestandaardiseerde residuen van het model, is gekozen voor S-PLUS. S-PLUS is zowel zeer geschikt voor grafische presentatie van de resultaten als ook voor statistische analyses van de schattingsresultaten. Voor een algemene inleiding over S-PLUS binnen het RIVM zie Dekkers (2001), en voor statistische methoden in milieu-onderzoek met S-PLUS zie Millard and Neerschal (2001).

Binnen S-PLUS is een plot- en analyse-routine geschreven in de vorm van een script, genaamd **PlotTrendSpotter**. Dit script zoekt zelf uit wat de componenten zijn van het geschatte tijdreeksmodel en maakt de relevante grafieken. De grafieken staan in het scherm '**Kalman**'. Dit scherm is op de voorgrond te krijgen door de knop '*Window*' aan te clicken op de menubalk van S-PLUS. In een tweede scherm komen de grafieken die horen bij tests op de oorspronkelijke data, zoals de *Range-Mean plot*. Ook worden grafieken getoond die berekend zijn op de residuen van het model (meer precies: de gestandaardiseerde innovaties). Dit zijn tests op normaliteit, cycli en de afhankelijkheid van opeenvolgende residuen.

De vormgeving van de grafieken kan binnen S-PLUS eenvoudig 'opgepoetst' worden al naar gelang de publicatiedoeleinden (zie handboeken S-PLUS, de helpfunctie, of *trial and error*). Het opstarten van het *script* geschiedt door het openen van het script '**PlotTrendSpotter**' (kies File → Open), en door op de toetsenbord-knop **F10** (= start script) te drukken. Een deel van het script kan uitgevoerd worden door het 'blokken' van de gewenste regels, en vervolgens op **F10** te drukken.

Als nevenproduct maakt het script een *dataframe* **writeall** op basis van de file **writeall.out**. Dit dataframe kan gebruikt worden voor additionele statistiek. Het kan bijvoorbeeld handig zijn op de voorspel-*performance* van het Kalmanfilter te evalueren met het MAD-criterium (gemiddelde van de absolute voorspelfouten). Additionele tests op de residuen kunnen uitgevoerd worden op de variabele **stinnov** (dit is de vector writeall$stinnov). Standaard worden een drietal grafische tests uitgevoerd op de residuen: een test of de residuen een normale verdeling volgen, een test of opeenvolgende residuen statistisch onafhankelijk zijn en een test of er cycli aanwezig zijn in de residuen.

Verder is S-PLUS handig om data aan te maken (bijvoorbeeld simulatievoorbeelden), of aan te passen (bijvoorbeeld maandwaarden maken op basis van dagwaarden van een variabele y_t). Het wegschrijven van het dataframe uit S-PLUS naar een ASCII-file gaat als volgt. Als het *dataframe* **PM10concentraties** heet en in onze directory **c:/TrendSpotter/run** de naam **PM10.dat** moet krijgen, dan typen we in het *command window* van S-PLUS:

```
sink (file= "c:/TrendSpotter/run/PM10.dat")
PM10concentraties
sink( )
```

5. PSEUDOREEKSEN

5.1 Het maken van pseudoreeksen

We geven in deze paragraaf een viertal voorbeelden van de analyse van tijdreeksen aan de hand van gesimuleerde 'pseudoreeksen'. Een pseudo-tijdreeks is een reeks die we samenstellen uit zelfgekozen componenten: bijvoorbeeld een trend met buigpunten, een cyclus en de invloed van een predictor, waarbij de weging die de predictor heeft in het model niet constant in de tijd hoeft te zijn. Tenslotte genereren we zelf een witte ruisproces. Hoe groter de variantie van de ruis, hoe moeilijker het zal zijn voor het model om de juiste trend, cyclus en weegfactor te reconstrueren. Hierna illustreren we een reeks van modellen aan de hand van twee pseudoreeksen.

We stellen als volgt een tweetal tijdreeksen, **Pseudo1** en **Pseudo2**, samen. Beide reeksen zijn opgebouwd uit een trend (deel van een parabool), een jaarcyclus en de invloed van één verklarende variabele. Zie **figuur 1**.

De modellen voor Pseudo1 en Pseudo2 zijn additief:

$$Pseudo1_t = trend_t + cyclus_t + \alpha_1 * x_t + ruis_t \qquad (6a)$$

$$Pseudo2_t = trend_t + cyclus_t + \alpha_{2,t} * x_t + ruis_t \qquad (6b)$$

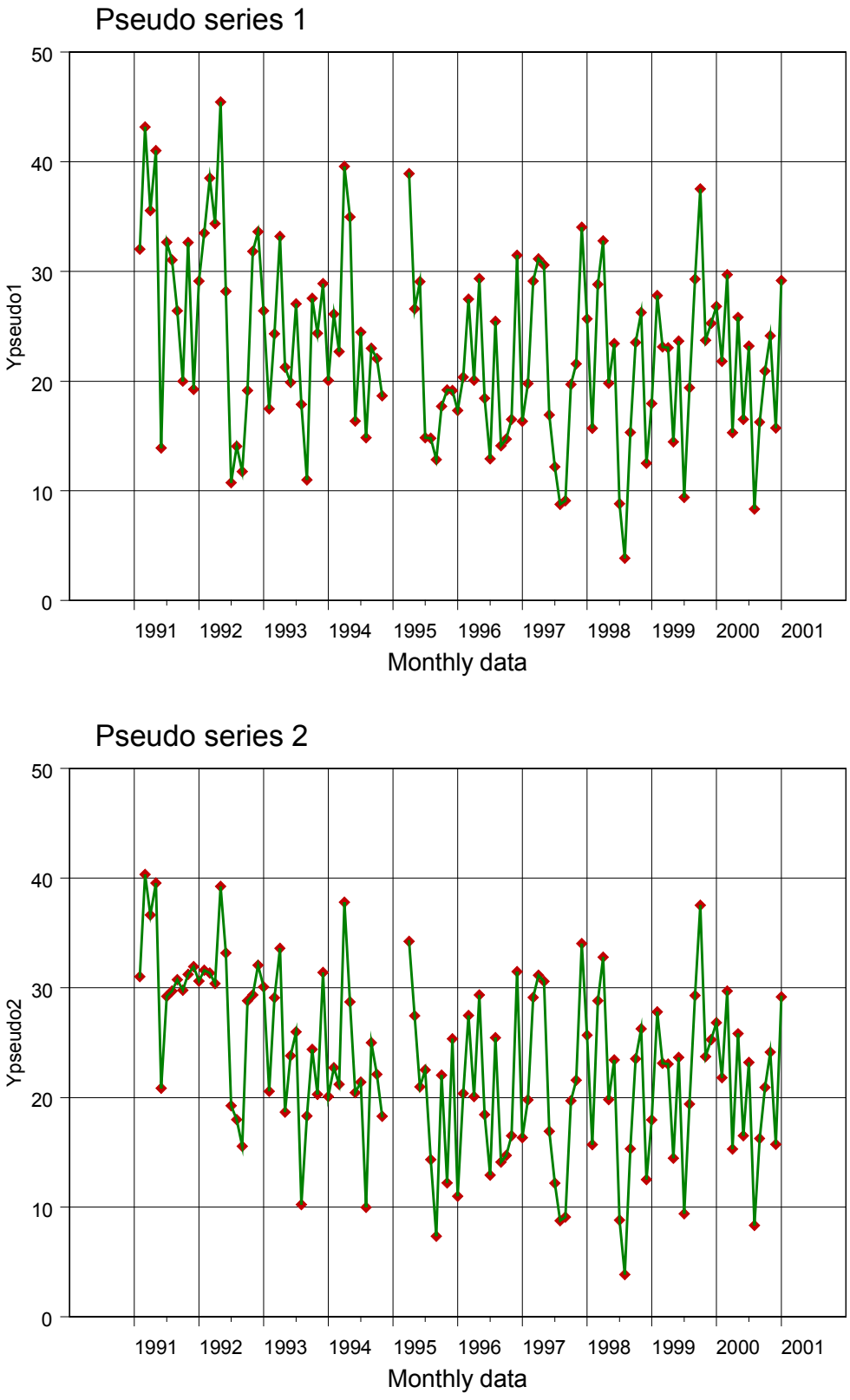
De pseudoreeksen zijn zo geconstrueerd dat het patroon lijkt op een luchtverontreinigende component als fijn stof, met een concentratie uitgedrukt in $\mu g/m^3$.

De tijdstap is maand en beslaat de jaren 1991 tot en met 2000. Hiermee loopt de tijdvariabele t van 1991.083 (= januari 1991) tot en met 2001.000 (= december 2000).

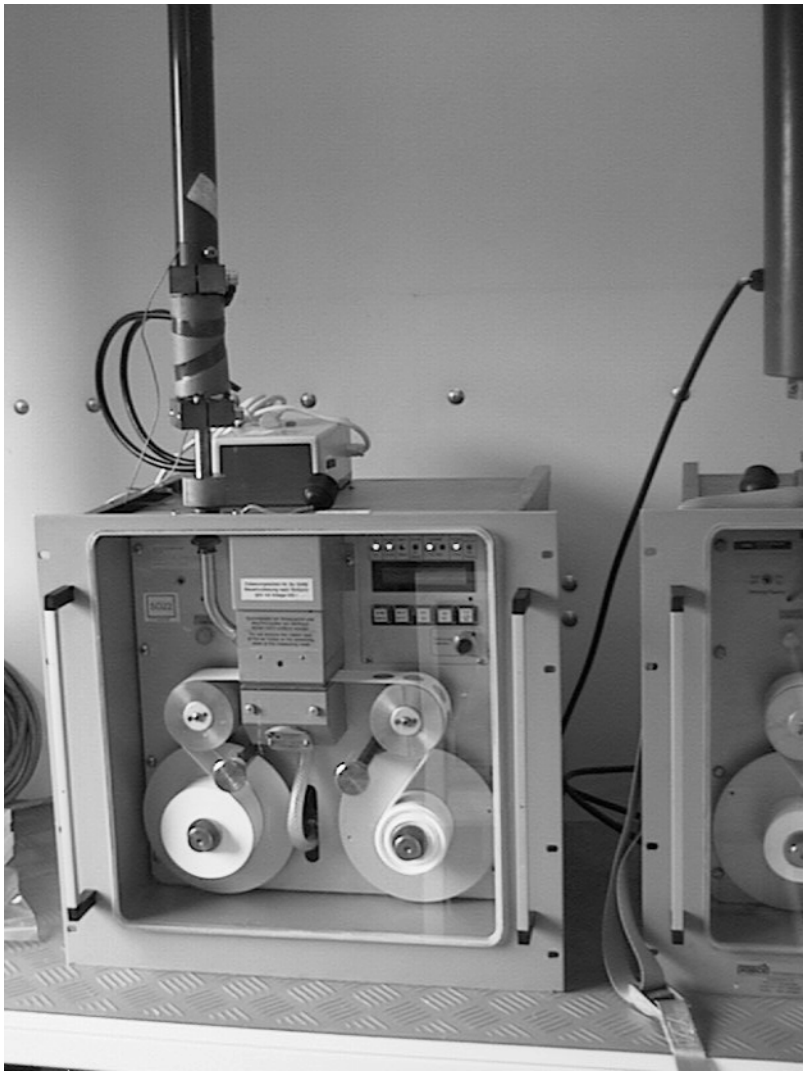
De vergelijking van de trend is

$$Trend_t = (t - 1997.0)^2 / 3.0 + 20.0 \ ,$$

en heeft een minimumwaarde van 20.0 in december 1996.



Figuur 1 *Gesimuleerde reeksen Pseudo1 (boven) en Pseudo2 (onder). Vanaf 1996 zijn beide reeksen identiek.*



De gesimuleerde reeksen Pseudo1 en Pseudo2 zijn zo geconstrueerd dat ze qua patroon en range lijken op fijn-stof-concentraties, zoals bijvoorbeeld gemeten met bovenstaande RIVM β-stof-monitor. *Foto: RIVM-LLO.*

De cyclus heeft voor de maanden januari tot en met december de waarden

0, 3, 6, 3, 0, –5, -8, -5, 0, 1, 5 en 0.

De vorm van deze cyclus houden we identiek voor alle jaren. De weegfactor α_1 heeft de constante waarde 5, en de tijdvariabele weegfactor $\alpha_{2,t}$ is stapvormig: 0.0 van januari 1991 tot en met december 1995, en daarna 5.0. Vanaf 1996 zijn daarmee de reeksen Pseudo1 en Pseudo2 identiek.

De reeks x_t bestaat uit normaalverdeelde witte ruis met gemiddelde 0.0 en standaarddeviatie 1.0. Het ruisproces 'ruis$_t$' is normaal-verdeeld met gemiddelde 0.0 en een standaarddeviatie van 5.0. De bijdragen van ruis tot de reeks Pseudo1$_t$ ligt daarmee in de orde van de invloed van $\alpha_1 * x_t$.

5.2 Data

De data zijn weergegeven in **tabel B.1** in Appendix B. De waarden van de individuele componenten zijn te vinden in de file **SimPlusKalfi** die te *downloaden* is van de installatie-CD van TrendSpotter.

Hierna zullen we laten zien hoe we:

- een DD-trend kunnen schatten voor de reeks Pseudo1 (§5.3);
- een DD-trend plus jaarcyclus kunnen schatten voor de reeks Pseudo1 (§5.4);
- een DD-trend plus invloed van een verklarende variabele met constante weegfactor kunnen schatten voor de reeks Pseudo1 (§5.5);
- een DD-trend plus invloed van een verklarende variabele met tijdvariabele weegfactor kunnen schatten voor de reeks Pseudo2 (§5.5);
- een DD-trend plus invloed van een verklarende variabele met constante weegfactor plus jaarcyclus kunnen schatten voor de reeks Pseudo1 (§5.6).

5.3 Trends

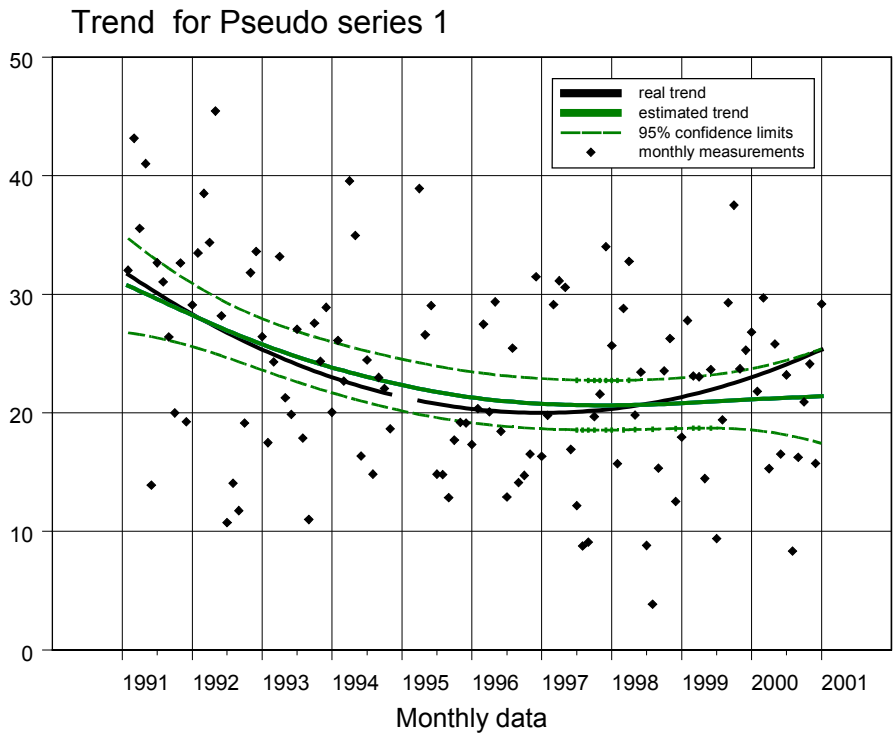
Als eerste stap willen we een trend schatten voor de reeks y_t = Pseudo1$_t$ (data in figuur 1A). Een methode om te komen tot een juiste keuze van het trendmodel, is gegeven in Appendix A.3.1. Analyse van de autocorrelatiefunctie (ACF) op de tweemaal gedifferentieerde reeks y_t geeft de correlaties R_1= -0.62 (±0.20) en R_2 = 0.18 (±0.20). Aangezien de verhouding tussen R_1 en R_2 bij benadering -4 is, kiezen we voor het DD-trendmodel. Voor een uitleg zie het einde van §A.3.1.

De optiefile voor het schatten van een DD-trend voor de data uit tabel B.1, is gegeven in **tabel B.2** in Appendix B. De datafile bevat ontbrekende waarden met de code –1 (laatste twee maanden van 1994 en eerste twee maanden van 1995). Deze code is in de optiefile opgenomen in stap 9.

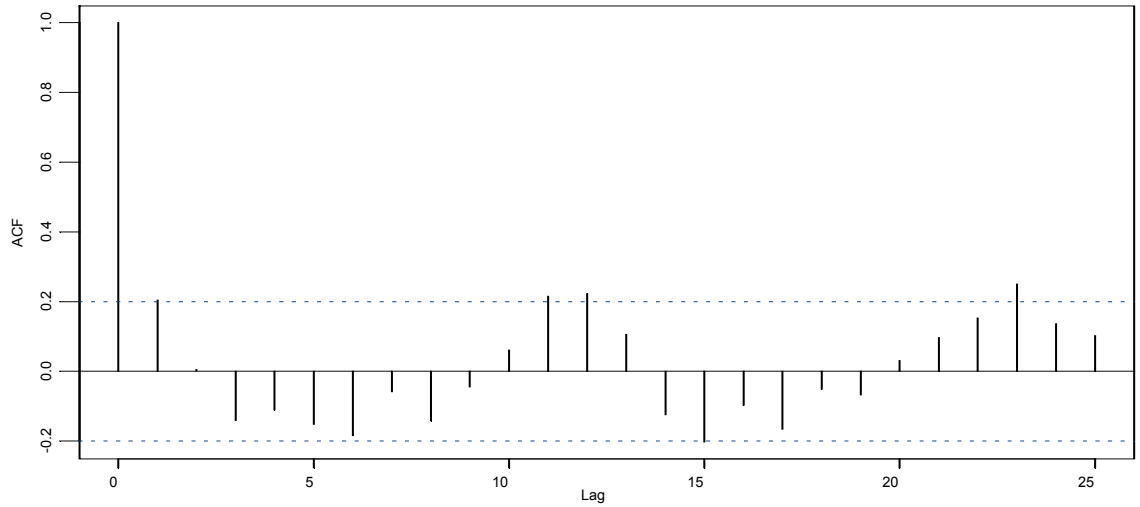
De schattingsresultaten zijn weergegeven in **figuur 2A**. Ook is de werkelijke trend (zwarte curve) geplot in de figuur. De werkelijke trend blijkt geheel tussen de 95%-betrouwbaarheidsintervallen te liggen

De grootste verschillen tussen de DD-trend, de groene curve, en de werkelijk trend treedt op in het jaar 2000. Dit is niet verwonderlijk gezien het hoge aantal maandwaarde onder de 20.0. Opmerkelijk is de mooie interpolatie van de ontbrekende vier maanden rond 1994/1995.

Uit de autocorrelatiefunctie voor de gestandaardiseerde innovaties van het geschatte model (**figuur 2B**) zien we dat R_1 niet helemaal nul is (opeenvolgende innovaties zijn dus niet geheel ongecorreleerd). Verder zien we significante correlaties voor data die 12 en 24 tijdstappen verschillen (R_{12} en R_{24}). Hieruit leiden we af dat we een cyclus met periode 12 moeten toevoegen aan het model (om het periodieke signaal uit de innovaties weg te krijgen), en wellicht nog additionele verklarende variabele (om de innovaties onafhankelijk te krijgen).



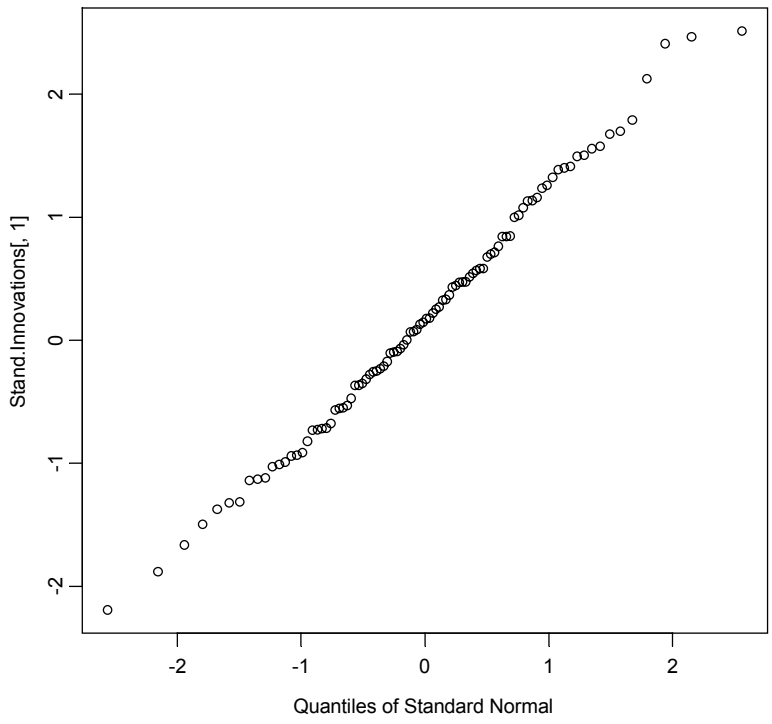
Figuur 2A *Schatting van het DD-trendmodel met 95%-betrouwbaarheidsintervallen (groene curves). De werkelijke trend is weergegeven door de zwarte curve in de figuur.*
De interpretatie van de betrouwbaarheidsintervallen is niet dat de *werkelijke* curve met 95% kans tussen deze boven- en ondergrenzen ligt, maar dat *per tijdstap* (hier één maand) de werkelijke trendwaarde met 95% kans tussen de bijbehorende onder- en bovengrens ligt.



Figuur 2B *Autocorrelatiefunctie voor lags 0 tot en met 25, berekend op de gestandaardiseerde innovaties.*
De stippellijnen geven een 95% betrouwbaarheidsinterval voor de correlaties. Correlaties binnen deze stippellijnen beschouwen we als niet-significant.

Uit **figuur 2C** blijkt dat de gestandaardiseerde innovaties normaal verdeeld zijn. De punten liggen nagenoeg op een rechte lijn.

In de volgende paragrafen zullen we zien of de trendschatting verbetert wanneer we additionele informatie toevoegen, zoals een jaarcyclus met periode 12.



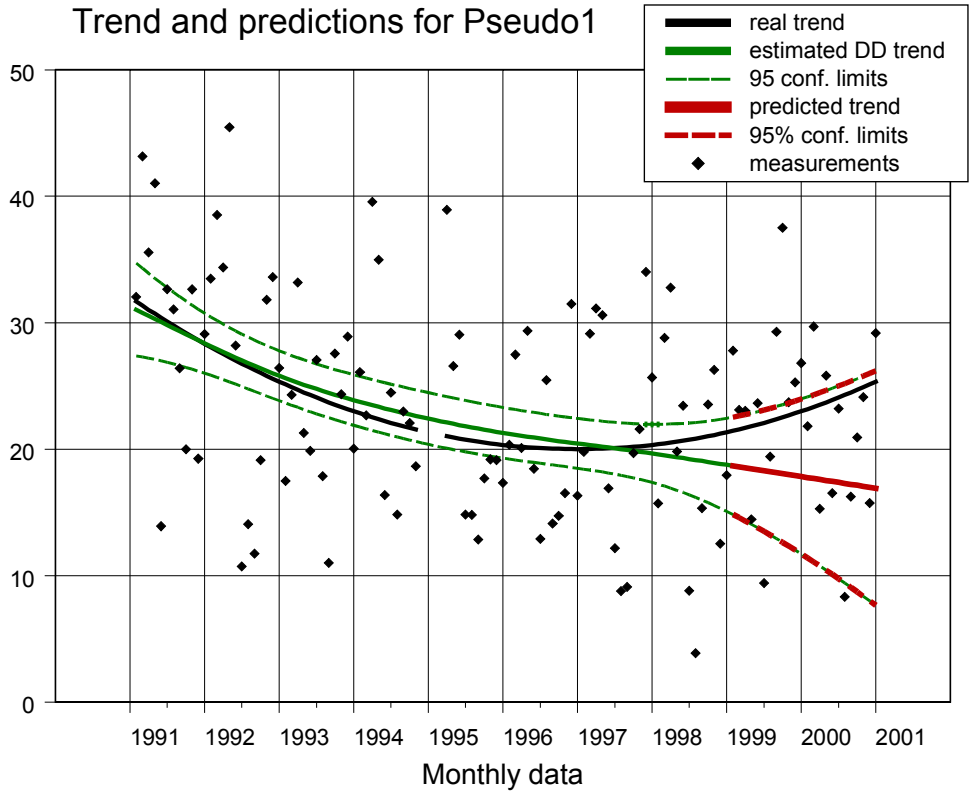
Figuur 2C *Normaliteitsplot voor de gestandaardiseerde innovaties.*
De punten liggen bij benadering op een rechte lijn, zodat we concluderen dat de innovaties normaalverdeeld zijn. Bovendien liggen de punten op een rechte die door de punten (–2,-2) en (2,2) gaat, zodat de innovaties ook standaardnormaal verdeeld zijn (gemiddelde 0.0 en standaarddeviatie 1.0).

We kunnen ook *voorspellingen* genereren met TrendSpotter. Als voorbeeld schatten we opnieuw een DD-trendmodel, maar nu niet op de maanddata 1991 tot en met 2000 van de reeks Pseudo1 (N= 120), maar op de data 1991 tot en met 1998 (N= 96). Vervolgens voorspellen we de trend voor de jaren 1999 en 2000. Het geschatte model is dus op geen enkele manier beïnvloed door de data in deze laatste twee jaar.

Het schattingsresultaat voor de trend is gegeven in **figuur 2D**. De figuur laat zien dat
- de voorspelde trend een extrapolatie is van de trend in de laatste jaren (1996, 1997, 1998);
- de betrouwbaarheidsintervallen wijder worden naarmate de voorspellingen verder in de toekomst liggen;
- de werkelijke trend voor alle maanden in de period 1999 - 2000 binnen de voorspelde 95%-betrouwbaarheidsintervallen ligt.

We merken op dat de betrouwbaarheidsintervallen bedoeld zijn om een statistische toets uit te voeren *per tijdstap*. Ze betekenen dus niet dat de werkelijke trend *voor alle meetpunten* met 95% zekerheid tussen de stippellijnen zal liggen! Overigens is dat hier wel het geval (zowel in het meetgebied als in het voorspelgebied).

Verder maken we de kanttekening dat de betrouwbaarheidsintervallen in figuur 2D **niet** gelden voor een voorspelling van y_t, met t een tijdstip in de toekomst. De onbetrouwbaarheid in zo'n voorspelling is groter dan de onzekerheid in de trend alleen.



Figuur 2D *DD-trendmodel voor Pseudo1, geschat over de periode 1991 tot en met 1998 (groene lijnen). Voorspellingen voor de trend beslaan de jaren 1999 en 2000 (rode lijnen).*

N.B.: de betrouwbaarheidsintervallen voor componenten als trend of cyclus kunnen met TrendSpotter geschat worden door kunstmatig y-waarden toe te voegen en deze als zeer onbetrouwbaar te definiëren. Voorspellingen van y_t worden door TrendSpotter gegenereerd middels een speciale voorspel-optie.

5.4 Cyclus

We voegen aan het model uit §5.3 het schatten van een jaarcyclus toe, met periodelengte 12. De optiefile is gegeven in **tabel B.3** in Appendix B.

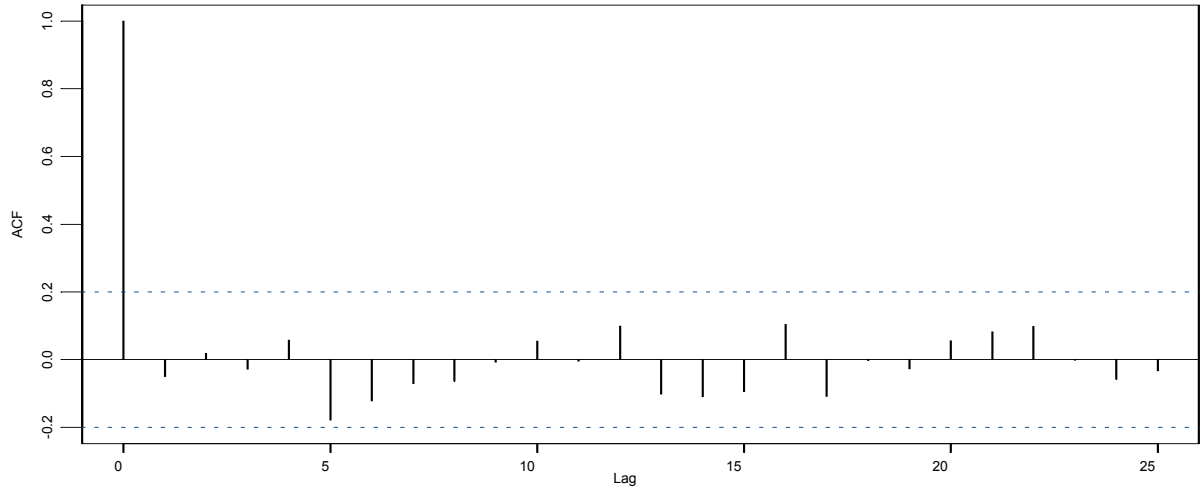
De schattingsresultaten zijn gegeven in figuur 3. De autocorrelatiefunctie op de gestandaardiseerde innovaties is gegeven in **figuur 3A**. Vergelijking van de figuren 2B en 3A laat zien dat de correlaties R_1, R_{12} en R_{24} nu niet meer statistisch significant zijn.

Figuur 3B geeft de schattingsresultaten:
- de geschatte DD-trend met 95%-betrouwbaarheidsintervallen (groene curves);
- de echte trend (zwarte curve);
- de geschatte trend-plus-jaarcyclus (rode curve);
- de werkelijke trend-plus-jaarcyclus (blauwe curve);
- de gesimuleerde metingen (♦).

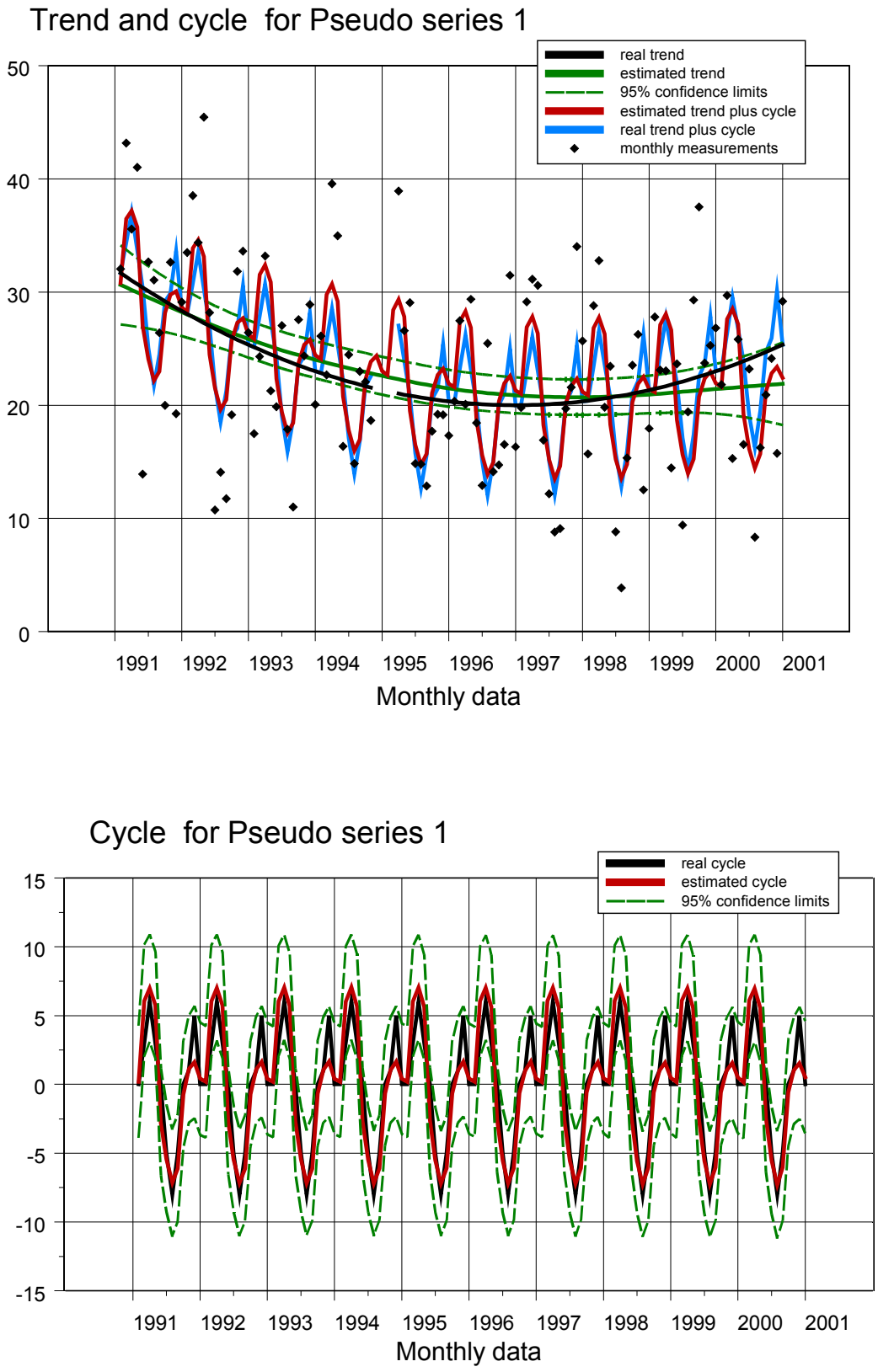
De schattingsresultaten van de jaarcyclus γ_t is gegeven in de onderste grafiek van figuur 3B. De rode curve is de geschatte cyclus en de zwarte curve de echte cyclus. We merken op dat de som van de maandwaarden in een jaar altijd sommeert tot 0.0 (vergelijking (2c)).

De fit van de cyclus is goed te noemen. De grootste afwijking valt aan het eind van ieder jaar (november). Toch ligt deze novemberwaarde nog binnen de 95%-betrouwbaarheidsgrenzen.

De bovenste grafiek van figuur 3B laat zien dat de trendschatting niet verbeterd is ten opzichte van die in figuur 2A (afwijking in het jaar 2000 blijft gelijk). In de onderste grafiek zien we dat de maand november systematisch *onderschat* wordt. Maar de *werkelijke* novemberwaarde ligt nog wel precies op de geschatte bovenste 2-σ-grens.



Figuur 3A *Autocorrelatiefunctie voor lags 0 tot en met 25, berekend op de gestandaardiseerde innovaties.*
De stippellijnen geven een 95% betrouwbaarheidsinterval voor de correlaties. Geen van de correlaties is significant van 0.0 te onderscheiden.

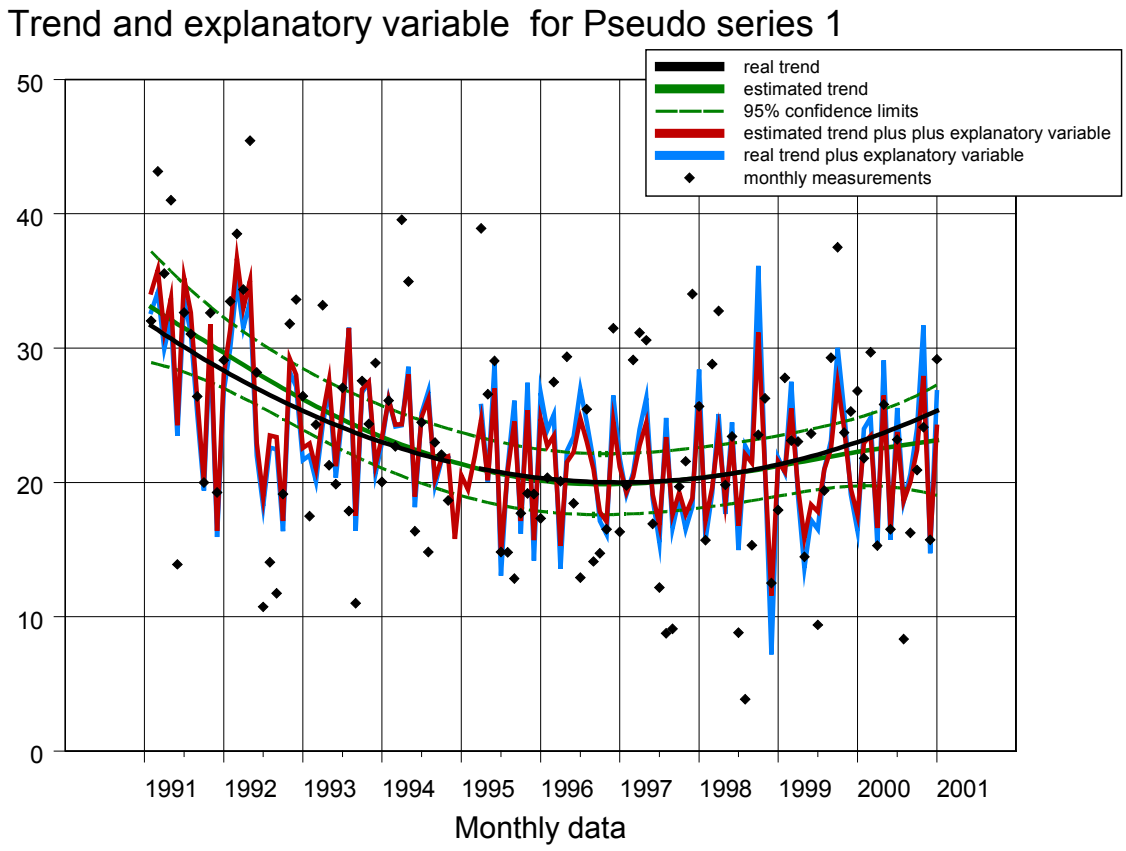


Figuur 3B *Schattingsresultaten voor DD-trend en jaarcylus op de reeks Pseudo1. De schatting van de cyclus γ_t is apart weergegeven in de onderste grafiek.*
De groene stippellijnen geven 95%-betrouwbaarheidsintervallen voor de trend (bovenste grafiek) en cyclus (onderste grafiek).

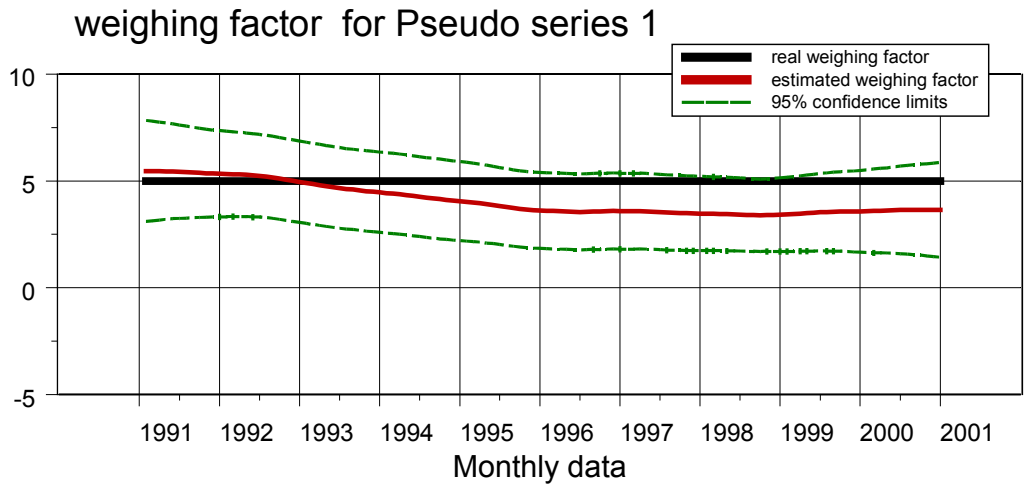
5.5 Verklarende variabelen

We schatten nu niet een trend-plus-cyclus, maar een DD-trend-plus-de-invloed-van-een-verklarende-variabele. De optiefile is weergegeven in **tabel B.4** uit Appendix B.

De schattingsresultaten zijn gegeven in **figuur 4A**. De grafiek laat een goede fit zien (verschil rode en blauwe curve is klein in verhouding tot de aanwezige ruis in de data). De figuur laat ook zien dat de echte trend μ_t (de zwarte curve) overal binnen de geschatte 95%-betrouwbaarheidsintervallen ligt (de groene stippellijnen). Dit geldt ook voor de weegfactor α_t in **figuur 4B**. Een goed resultaat!



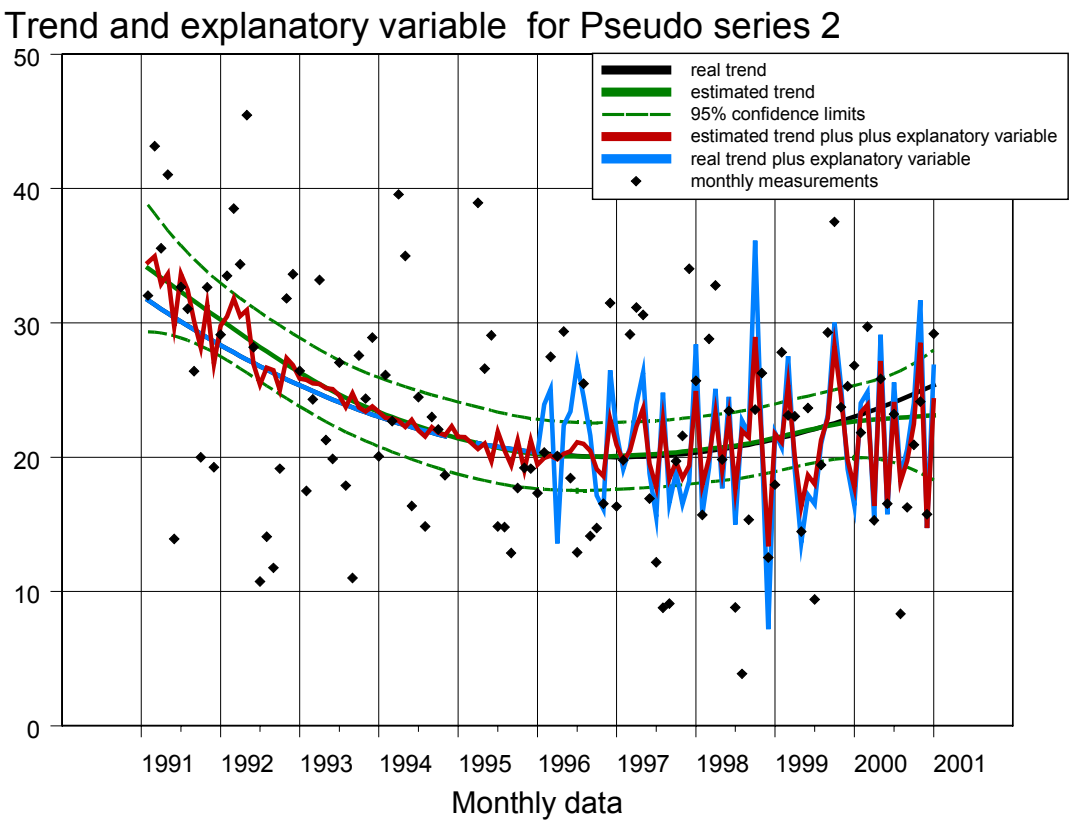
Figuur 4A *DD-trend en invloed van een verklarende variabele x_t voor Pseudo1.*
De grafiek geeft de geschatte trend (groene lijn), de werkelijke trend (zwarte curve), de trend-plus-invloed van verklarende variabele (ofwel $\mu_t + \alpha_t * x_t$), en de **werkelijke** curve in blauw.

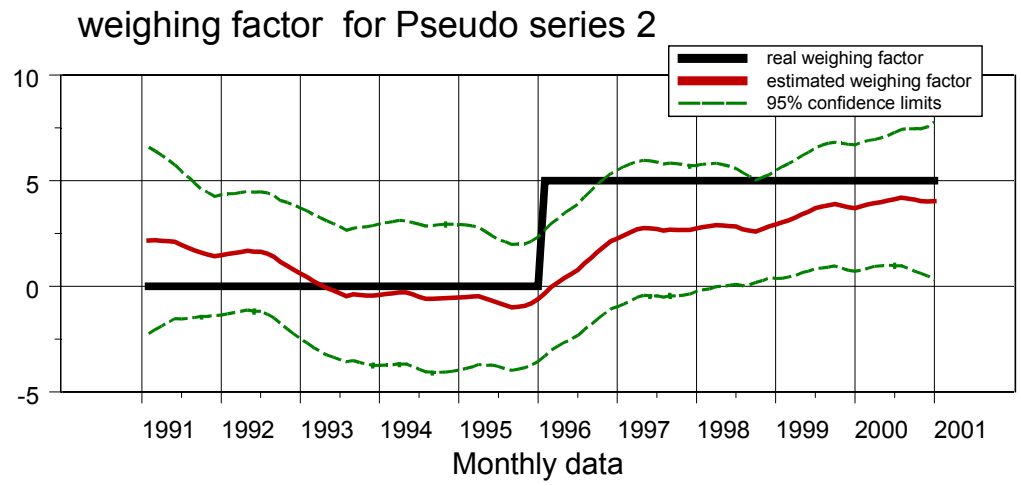


Figuur 4B Verloop van de weegfactor α_t. De werkelijke curve is weergegeven in zwart.

In **figuur 5** zijn de schattingsresultaten voor de reeks Pseudo2 gegeven. Het enige verschil tussen Pseudo1 en Pseudo2 is dat over de jaren 1991 tot en met 1995 de invloed van x_t nul is ($\alpha_{1,t} = 5.0$ en $\alpha_{2,t} = 0.0$). Ná 1995 zijn de reeksen Pseudo1 en Pseudo2 identiek. Voor de jaren 1991 tot en met 1995 geldt dus dat de trend plus invloed van x_t samenvalt met de trend zelf (blauw en zwarte curve in bovenste grafiek van figuur 5).

Het schattingsresultaat voor de weegfactor laat zien dat de stapfunctie met een vertraging van enkele jaren goed geschat wordt. Essentieel is hier dat we het model *geen informatie* hebben gegeven over de structurele veranderingen vanaf 1996.





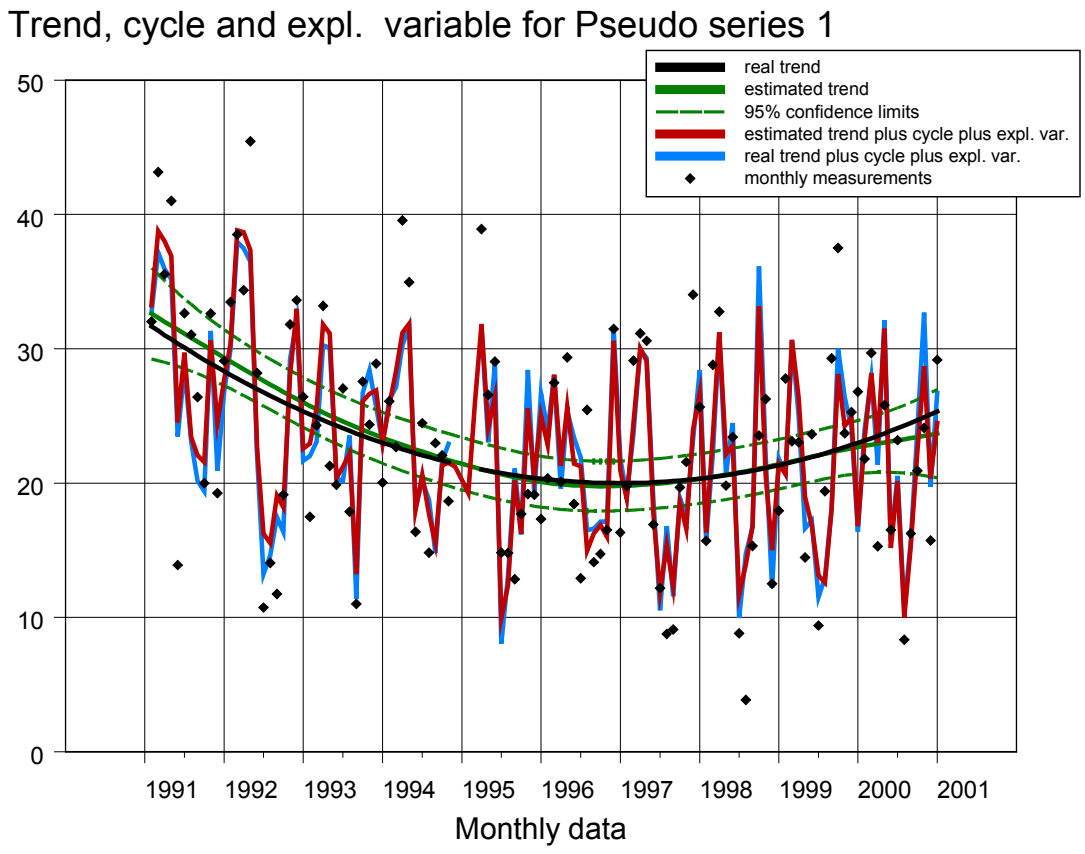
Figuur 5 *DD-trend en invloed van een verklarende variabele x_t voor Pseudo2.*
De bovenste grafiek geeft de geschatte trend (groene lijn), de werkelijke trend (zwarte curve), de trend-plus-invloed-van-verklarende-variabele, en de werkelijke curve in blauw. De onderste grafiek geeft de weegfactor α_t. De werkelijke curve is weergegeven in zwart.

5.6 Trend, cyclus en verklarende variabele

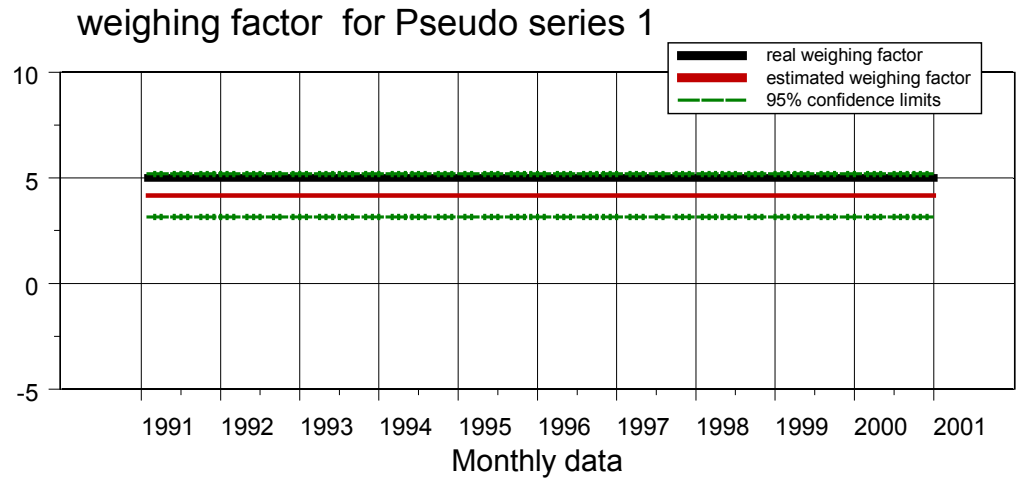
Tenslotte schatten we het complete model voor de reeks Pseudo1: DD-trend, jaarcyclus en de invloed van een verklarende variabele, ofwel het model $y_t = \mu_t + \gamma_t + \alpha_t * x_t + \varepsilon_t$. De optiefile is gegeven in **tabel B.5** in Appendix B. De schattingsresultaten zijn weergegeven in de **figuren 6A, B en C.**

Aan de hand van figuur 6 kunnen de volgende opmerkingen gemaakt worden:

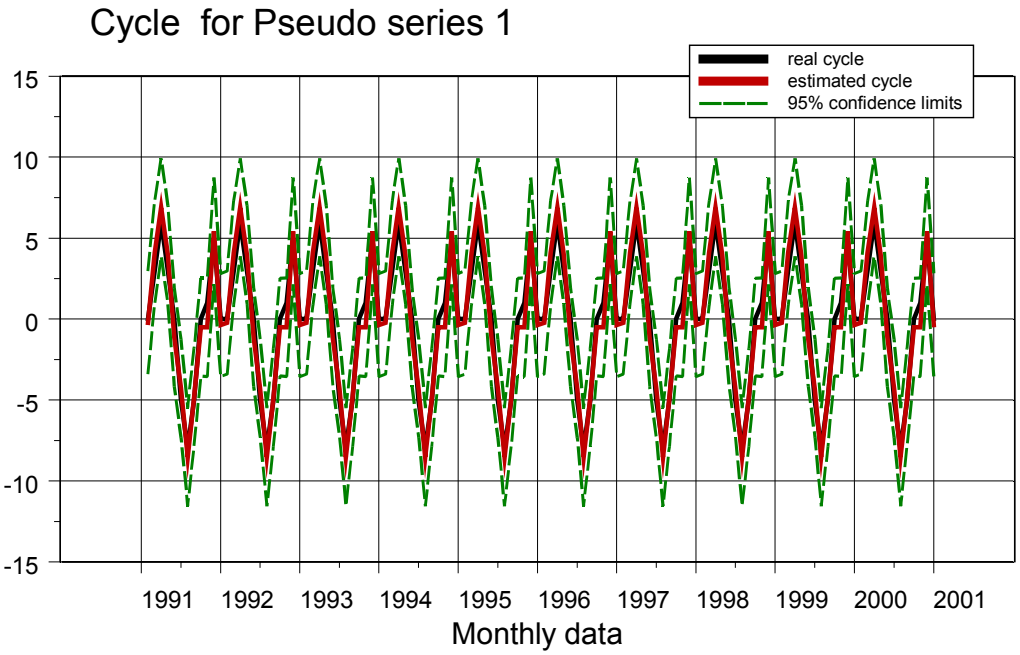
- de trend wordt beter geschat dan in de voorgaande modellen;
- ondanks het ruisproces ε_t in de data wordt het model $\mu_t + \gamma_t + \alpha * x_t$ nauwkeurig teruggeschat;
- de weegfactor α wordt iets aan de lage kant geschat. Maar de werkelijke waarde ($\alpha = 5.0$) ligt nog net binnen de 95%-betrouwbaarheidsintervallen;
- de cyclus wordt nagenoeg perfect geschat (de zwarte curve in figuur 6C is bijna onzichtbaar).



Figuur 6A *DD-trend, jaarcylus en de invloed van een verklarende variabele.*
De grafiek toont het geschatte totale model (rode curve) en het werkelijke model $\mu_t + \gamma_t + \alpha * x_t$ (blauwe curve). De aangeboden reeks y_t is weergegeven met het symbool ♦.



Figuur 6B Geschatte en werkelijke weegfactor α.



Figuur 6C De geschatte cyclus (rode curve) met de werkelijke cyclus (zwarte curve) en 95%-betrouwbaarheidsintervallen (groene curves).

Analyse van gesimuleerde pseudoreeksen zoals getoond in figuur 1, heeft een aantal voordelen. In de eerste plaats kunnen software-fouten opgespoord worden. Immers het antwoord is van te voren bekend. In de tweede plaats kunnen we de grenzen opzoeken van wat mogelijk is met Structurele Tijdreeksmodellen. Hoeveel ruis kunnen we toevoegen aan het deterministische signaal zodat de schattingen toch realistisch blijven? En geven de onzekerheidsbanden inderdaad een goede indruk van de werkelijke onzekerheid (die we a-priori gekozen hebben)? In alle voorbeelden blijkt TrendSpotter goed tot zeer goed te scoren.
Foto: H. Visser

6. KLIMAATVERANDERING IN NEDERLAND

De gemiddelde temperatuur is de laatste tien jaar zowel in Nederland als gemiddeld over de hele wereld hoger dan zeg 100 jaar geleden. De stijging van de wereldgemiddelde temperatuur is door het IPCC in het Third Assessment Report (2001) met grote zekerheid mede toegeschreven aan de toenemende concentratie broeikasgassen:
'There is new and stronger evidence that most of the warming observed over the last 50 years is attributable to human activities.'

Van Oldenborgh en Komen (2001) hebben laten zien dat de opwarming in Nederland aan de mondiale opwarming kan worden gerelateerd. Daarbij is gevonden dat er een correlatie bestaat tussen Nederlandse en mondiale temperaturen. Dit geldt zowel voor *jaargemiddeldes* als ook voor de *afzonderlijke seizoenen*. Volgens hun model is ongeveer de helft van de Nederlandse opwarming in de negentiger jaren gerelateerd aan de mondiale opwarming. De andere helft moet als een 'toevallige uitschieter' worden gezien.

Hier zullen we laten zien hoe de temperatuurverandering er in Nederland uitziet, en of er ook een verschuiving in de jaarcyclus is ontstaan. Dit laatste is met name van belang voor veranderingen in de lengte van het groeiseizoen (en ook van het recreatieseizoen). Naast temperatuur kijken we ook naar neerslag over de afgelopen eeuw.

Voor de analyse maken we gebruik van maand- en jaargemiddelde temperaturen en neerslagsommen voor De Bilt vanaf 1901. Deze reeksen zijn wel verder in de tijd terug te reconstrueren, maar er moeten dan verschillende meetlocaties 'aan elkaar geplakt' worden. Om inhomogeniteiten te vermijden, hebben we daarom de reeksen ingeperkt tot de periode vanaf januari 1901. De maandgemiddelde reeksen voor temperatuur en neerslag zijn rechtstreeks te *downloaden* van de KNMI-site www.knmi.nl. Eerdere analyses van deze (en andere) klimaatreeksen met Structurele Tijdreeksmodellen zijn gegeven door Visser en Molenaar (1995).

We merken op dat de hier gegeven analyses van langjarige temperatuur- en neerslagreeksen niet gericht zijn op de detectie van verandering *in extreme weersituaties*, zoals hittegolven of extreme neerslaghoeveelheden in enkele dagen. Deze fenomenen zijn weggemiddeld in de maand- en jaargemiddelde data die we hier gebruiken.

6.1 Data

Op de Installatie-CD van TrendSpotter zijn drie files geplaatst: 'Temperatuur.dat', 'Neerslag.dat' en 'MeteoDebiltMonthly.dat'. De file **Temperatuur.dat** heeft format (3X,F5.0,2F5.1,F4.1,9F5.1,F10.6). De file bevat in deze volgorde: jaar, temperatuur januari tot en met temperatuur december, en jaargemiddelde temperatuur in ˚C. De periode is 1901 tot en met 2002. De file **Neerslag.dat** heeft het format (3X,F5.0,12F6.1,F8.1), en is analoog opgebouwd aan de file **Temperatuur.dat**. De file **MeteoDeBiltMonthly.dat** heeft format (4X,F9.3,2F6.1). De file bevat in deze volgorde: oplopende tijdvariabele, temperatuur per maand in ˚C en neerslagsom per maand in mm. De periode is januari 1901 tot en met maart 2002. Daarmee geldt N= 1214.

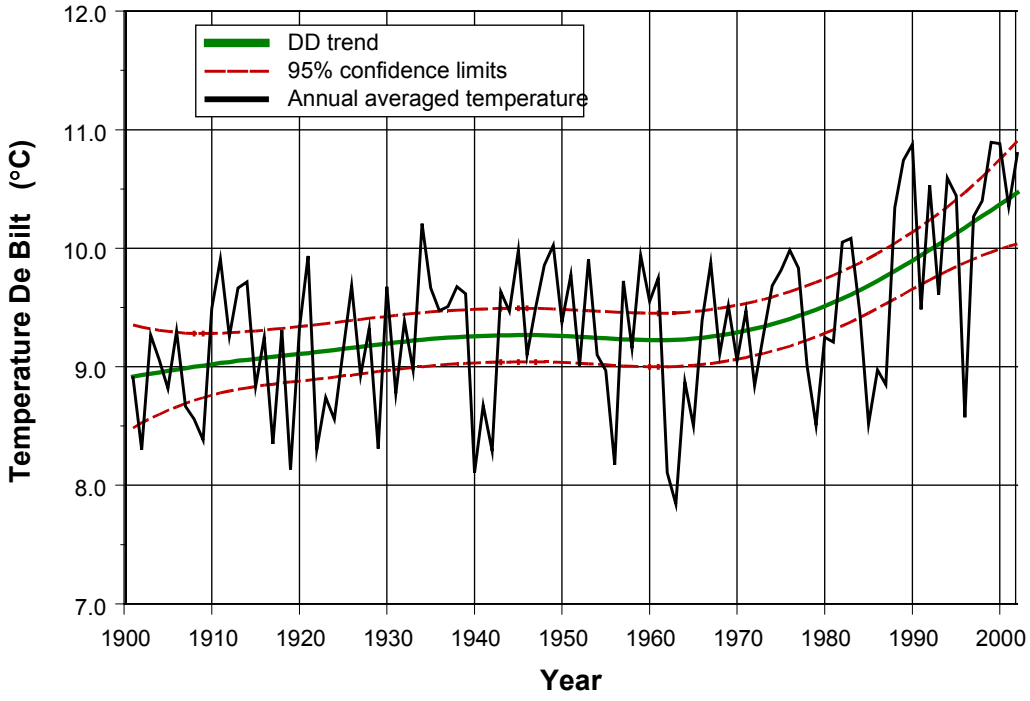
6.2 Temperatuur 1901 – 2002

6.2.1 Trend

In **figuur 7A** is het DD-trendmodel toegepast op jaargemiddelde temperaturen van De Bilt. De groene curve geeft de geschatte trend. De rode curves geven per jaar een 95%-betrouwbaarheidsinterval. Dit interval betekent **niet** dat met 95% zekerheid de **werkelijke** trendmatige opwarming binnen de onderste en bovenste rode stippellijnen ligt. De intervallen moeten geïnterpreteerd worden als intervallen voor individuele jaren.

Een voorbeeld. De betrouwbaarheidsgrenzen kunnen worden gebruikt om te toetsen of de trendwaarde in 2002 significant hoger is dan de grenswaarde van bijvoorbeeld 10.0 °C. Omdat uit de grafiek blijkt dat 10.0 °C net buiten de rode stippellijnen ligt, kunnen we concluderen dat de trendwaarde voor 2002 (10.47 ± 0.44 °C) significant van 10.0 °C te onderscheiden is.

De temperatuurtrend kan als volgt worden samengevat. De trend start in 1901 met 8.92 ± 0.44 ºC, en stijgt naar een constant plateau over de period 1940-1960 van 9.26 ± 0.22 ºC. Na 1960 vertoont de trend een sterke stijging naar 10.47 ± 0.44 ºC in het eindjaar 2002.



Figuur 7A Verloop van jaargemiddelde temperaturen voor De Bilt met daarbij de langjarige trend en 95%-betrouwbaarheidsintervallen.

Merk op dat de betrouwbaarheidsintervallen in het midden van de reeks veel kleiner zijn dan aan de eindpunten van de reeks (0.22 respectievelijk 0.44 ºC). Dit verschijnsel hangt samen met het *smoothen* van data (zie §A.4.2). Hierbij wordt voor elk punt in de tijd alle *omringende* data gebruikt om een zo goed mogelijke schatting voor dat punt te geven. Maar aan het begin en einde van de reeks zijn er veel minder omringende punten beschikbaar dan in het midden van de reeks. Vandaar dus dat de schattingen aan het begin en einde van de reeks veel onzekerder zijn.

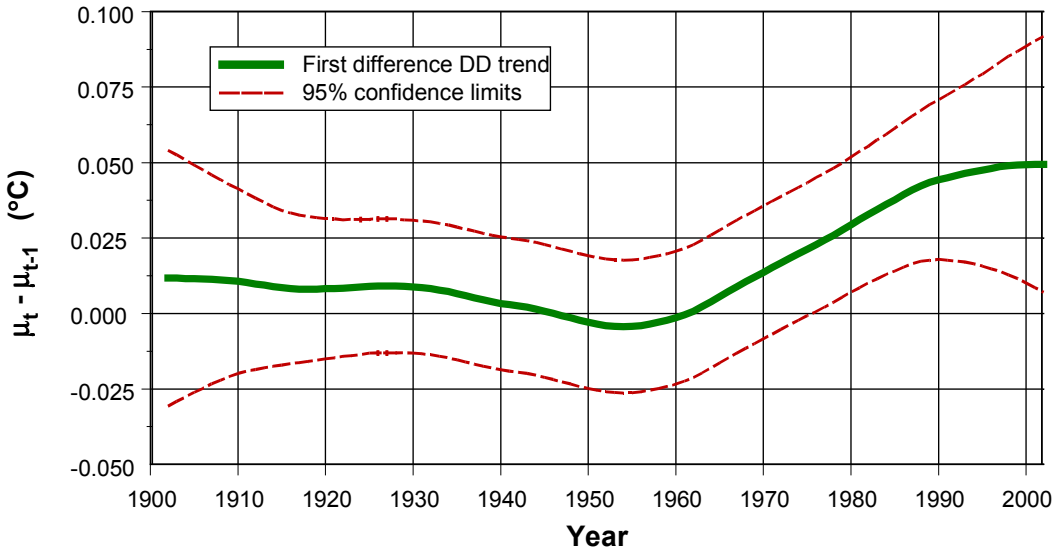
Over de periode 1901 – 1987 was de jaargemiddelde temperatuur in De Bilt het hoogst in 1934 (10.2 ˚C). De foto geeft een impressie van de warmte in 1934. Maar vanaf 1988 volgen ***12*** *jaar met temperaturen hoger dan 10.2 ˚C. Het jaar 2002 is net geen record met 10.8 ºC. De hoogste jaren zijn 1990, 1999 en 2000, elk met een jaargemiddelde van 10.9 ºC. Door de grote concentratie van jaren met hoge temperaturen aan het einde van de reeks onstaat in de negentiger jaren een sterke trendmatige stijging van de jaargemiddelde temperatuur.*
Foto: Archief Spaarnestad.

6.2.2 Is de temperatuurstijging statistisch significant?

Figuur 7B geeft een uitvergroting van de trend door te kijken naar de eerste differentie van de trend μ_t. Dit zijn de opeenvolgende verschillen van de trend (ofwel $\Delta\mu_t = \mu_t - \mu_{t-1}$). De figuur geeft ook de bijbehorende 95%-onzekerheidsintervallen. Waarden boven 0.0 geven stijging aan, waarden onder 0.0 een daling.

De grafiek laat zien dat de jaar-op-jaar-toename vanaf 1975 significant is. De toename in 2002 bedraagt 0.05 ± 0.04 ˚C/jaar. Zou deze toename in de komende een eeuw constant blijven, dat zou dit een toename van **5 ˚C in 100 jaar** betekenen!

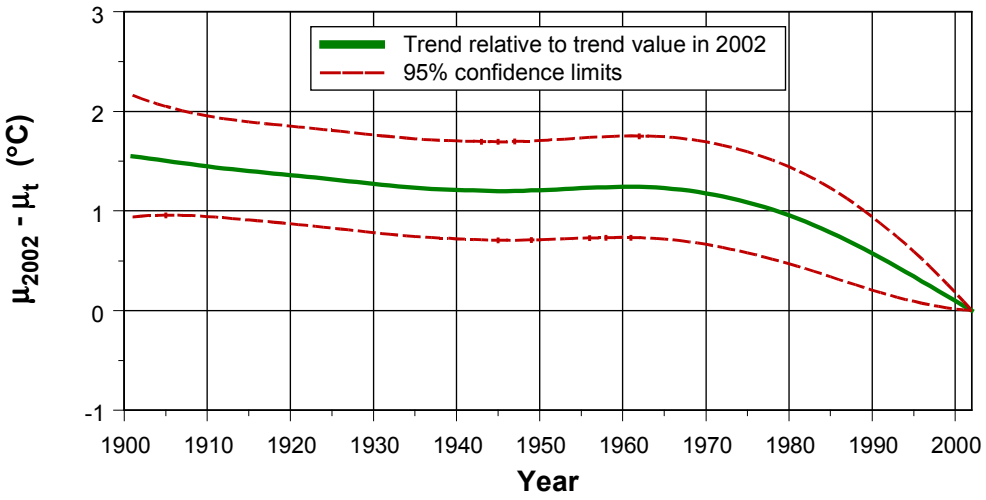
De betrouwbaarheidsintervallen in figuur 7B gelden voor de differentie van de trend. Maar als deze toename niet significant is, dan wil dat niet zeggen dat verschillen tussen jaren die ver van elkaar liggen, bijvoorbeeld 10 of 100 jaar (dus $\mu_t - \mu_{t-10}$ of $\mu_t - \mu_{t-100}$), niet significant zouden kunnen zijn. Betrouwbaarheidsintervallen voor grotere lags dan 1 kunnen ook geschat worden met het Kalmanfilter. Zie Visser (1994, pagina's 124-125), en Visser en Molenaar (1995, figuur 2).



Figuur 7B *Eerste differentie $\mu_t - \mu_{t-1}$ van de trend met 95%-betrouwbaarheidsintervallen. Positieve waarden duiden op een stijgende trend, negatieve waarden op een dalende trend.*

Voor de trend uit figuur 7A hebben we het verschil berekend tussen de laatst bekende waarde van de trend (de waarde uit 2002) en de waarde van de trend in een specifiek jaar, ofwel het verschil $\Delta_{2002}\mu_t \equiv \mu_{2002} - \mu_t$. Vanwege deze definitie is de verschilwaarde en de onzekerheid daarin altijd nul in het eindjaar. De functie $\Delta_{2002}\mu_t$ is zo gekozen omdat men over het algemeen de huidige situatie (weerspiegeld door de laatst bekende meting) wil vergelijken met het verleden. Vervolgens wil men een uitspraak doen over sinds wanneer de huidige situatie statistisch significant afwijkt van de historische gegevens.

Het resultaat van de verschillen $\Delta_{2002}\mu_t \equiv \mu_{2002} - \mu_t$ voor de temperatuurtrend uit figuur 7A is gegeven in **figuur 7C**. De figuur laat zien wat al te verwachten was uit de resultaten van figuur 7B. De trendwaarde in 2002 is significant groter dan alle jaargemiddelde trendwaarden in de 101 jaar ervoor! Het verschil $\mu_{2002} - \mu_{1901}$ bedraagt 1.55 ± 0.62 ºC.



Figuur 7C De differentie $\mu_{2002} - \mu_t$ met 95%-betrouwbaarheidsintervallen.

6.2.3 Persistentie en cycli

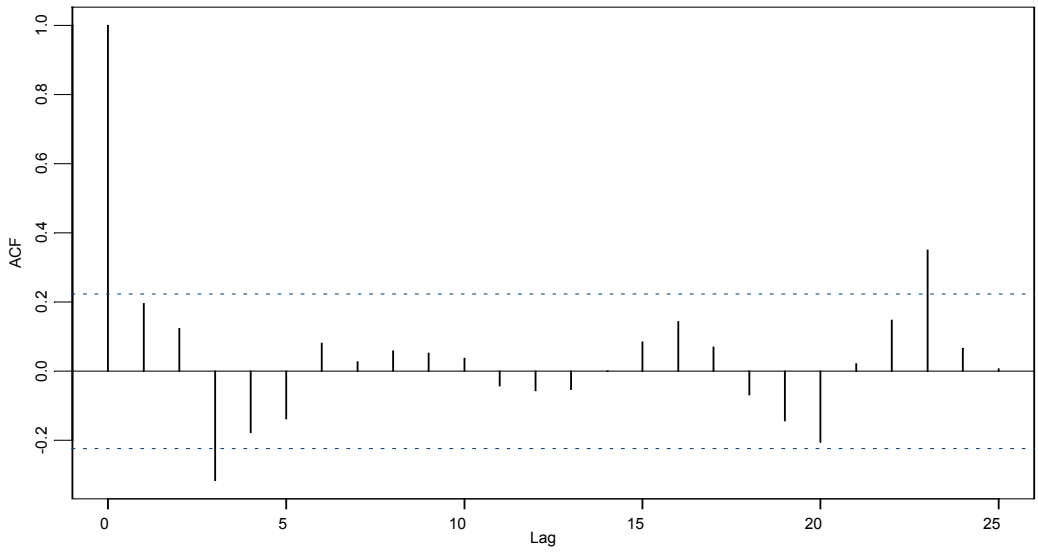
Om te onderzoek of het geschatte trendmodel voldoet aan de vooronderstellingen van het Structurele Tijdreeksmodel in combinatie met het Kalmanfilter, is het van belang om de *residureeks*, dat is het verschil tussen de metingen en het model, te onderzoeken. Een belangrijke controle op deze residuen is de autocorrelatiefunctie (ACF). Het doel van deze functie is uitgelegd in §2.6.

Figuur 7D toont de autocorrelatiefunctie op de residuen van het model uit figuur 7A. De figuur laat zien dat er nog geringe autocorrelaties bestaan tussen opeenvolgende jaren (R_1 en R_2). Echter deze persistentie valt binnen de 2-σ-betrouwbaarheidsintervallen (+0.22 en –0.22).

Een andere opmerkelijke correlatie is die voor jaren die 23 jaar verschillen (R_{23} uit de figuur). Deze significante correlatie suggereert een cyclisch verband tussen jaargemiddelde temperaturen met een periode van 23 jaar. Dit kan een verwijzing zijn naar de magnetische zonnevlekcyclus (de zogenaamde Hale-cyclus), die een periode heeft van gemiddeld 22 jaar, het dubbele van de zonnevlekcyclus. Dit getal is een *gemiddelde*, want de zonnevlekcyclus varieert tussen de 8 en 15 jaar. Daarmee varieert de magnetische cyclus tussen de 16 en 30 jaar.

Of R_{23} inderdaad gekoppeld kan worden aan de magnetische cyclus van de zon, zullen we hier niet verder uitdiepen. Er bestaat een zeer uitgebreide literatuur over dit thema. Zie bijvoorbeeld Scholten (1992) en IPCC (2001, §6.1 en §12.2.3).

Tenslotte is de negatieve correlatie R_3 ook significant. De fysische betekenis is onduidelijk.

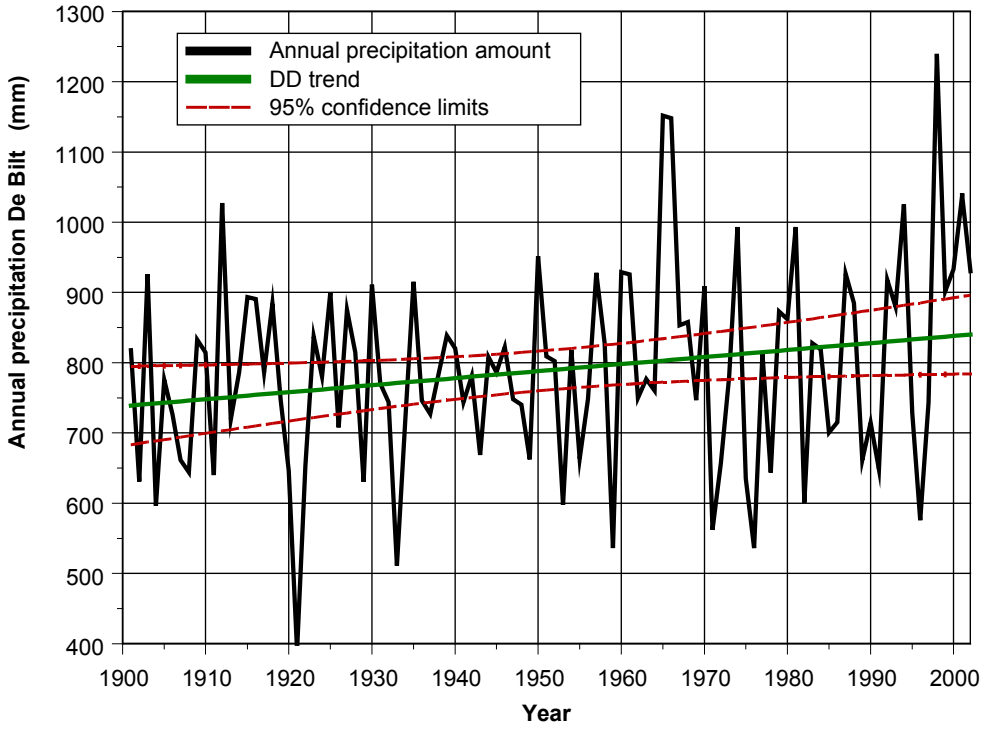


Figuur 7D Autocorrelatiefunctie van de residuen, behorend bij het model uit figuur 7A.

6.3 Neerslag 1901 – 2002

6.3.1 Trend

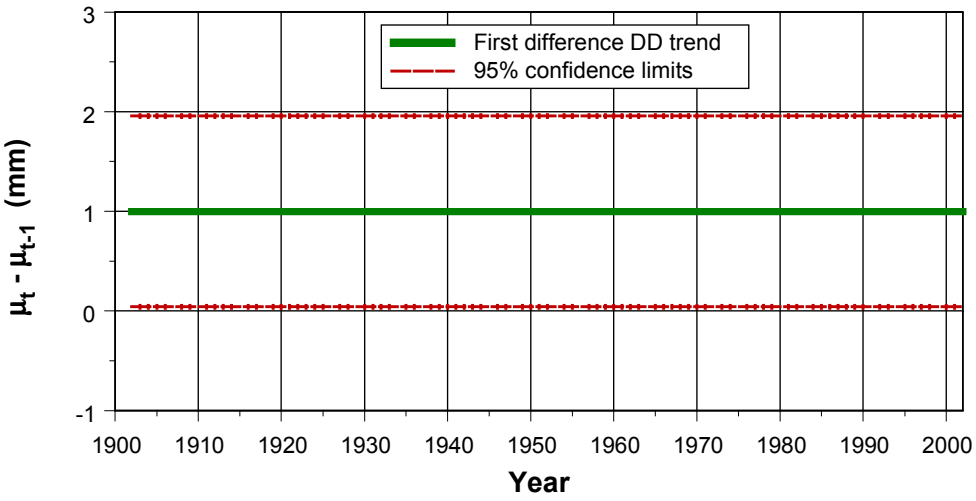
Figuur 8A geeft de trendschatting voor neerslag in De Bilt. Het toegepaste trendmodel is gelijk aan die voor temperatuur in figuur 7A. De trend in neerslag is nagenoeg constant. De figuur laat zien dat de trend in 1901 739 ± 56 mm bedraagt en oploopt naar 840 ± 56 mm in het jaar 2002. De toename over de hele periode is dus 101 mm, ofwel een toename met 13% in een eeuw tijd.



Figuur 8A Verloop van de jaarsom van neerslag voor De Bilt met daarbij de langjarige trend en 95%-betrouwbaarheidsintervallen.

6.3.2 Is de neerslagstijging statistisch significant?

Figuur 8B geeft een uitvergroting van de trend uit figuur 8A. De figuur toont, analoog aan figuur 7B, de eerste differentie van de trend, met bijbehorende betrouwbaarheidsintervallen. De figuur laat zien dat de jaarlijkse toename nagenoeg constant is over de hele reeks: 1.00 ± 0.96 mm/jaar (2-σ-grenzen). Op basis van de betrouwbaarheidsintervallen zijn alle jaarlijkse verschillen (net) statistisch significant.

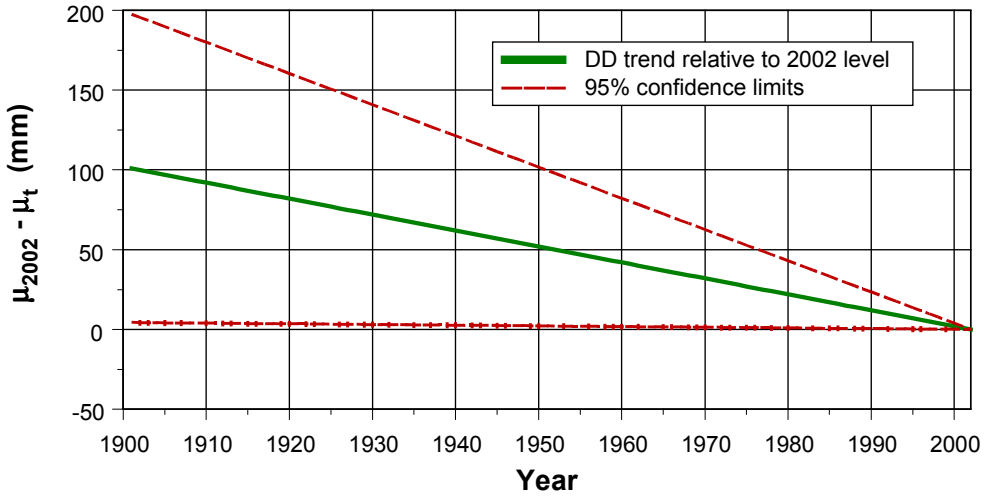


Figuur 8B *Eerste differentie μ_t - μ_{t-1} van de trend met 95%-betrouwbaarheids-intervallen.*

Figuur 8C laat zien, analoog aan figuur 7C, dat verschillen tussen het eindjaar 2002 en eerdere jaren steeds (net) significant zijn. Het verschil $\mu_{2002} - \mu_{1901}$ bedraagt 101 ± 96 mm.

De resultaten uit de figuren 8B en 8C laten zich als volgt eenvoudig verklaren. De geschatte trend heeft een lineair karakter en kan ook geschat worden met het Single Regression Model $y_t = a + b*t + \varepsilon_t$, met a de intercept, b de helling, t de tijd in jaren en ε_t een ruiscomponent. Omdat nu $\mu_t = a + b*t$, geldt dat de differentie $\mu_t - \mu_{t-1}$ gelijk is aan de helling b. Om de significantie van de trend te testen, hoeven we alleen de onzekerheid in de helling b te kennen. Op dezelfde wijze vinden we voor de differentie $\mu_{2002} - \mu_t$ de waarde b * (2002 – t). Ook hier bepaalt de fout in de helling b *voor alle jaren* het wel of niet significant zijn van de differentie.

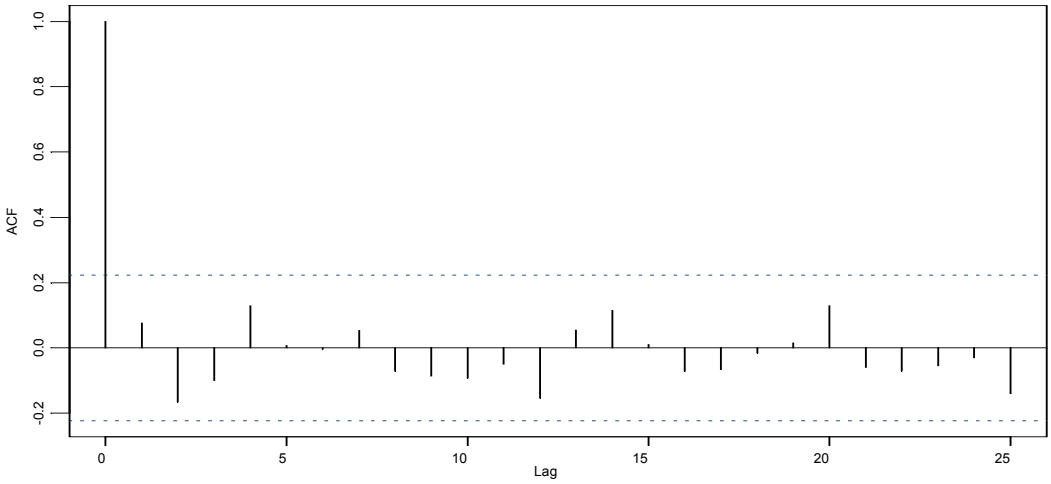
De conclusie hier is dat als de trend lineair is, de betrouwbaarheden voor de differenties $\mu_t - \mu_{t-1}$ en $\mu_{2002} - \mu_t$ geheel bepaald worden door de betrouwbaarheidsgrenzen van de helling b. Maar als de trend één of meer buigpunten vertoont, zoals bijvoorbeeld bij de temperatuurreeks in §6.2, dan geven de figuren 7B en 7C noodzakelijke onzekerheids-informatie.



Figuur 8C De verschil $\mu_{2002} - \mu_t$ met 95%-betrouwbaarheidsintervallen.

6.3.3 Persistentie en cycli

Figuur 8D geeft de autocorrelatiefunctie voor de residuen van het trendmodel. De figuur laat zien dat er geen significante correlaties zijn. Er is dus geen verdere modellering nodig zoals dat voor de temperatuurreeks wèl het geval was.



Figuur 8D Autocorrelatiefunctie op de residureeks behorend bij het model uit figuur 8A.

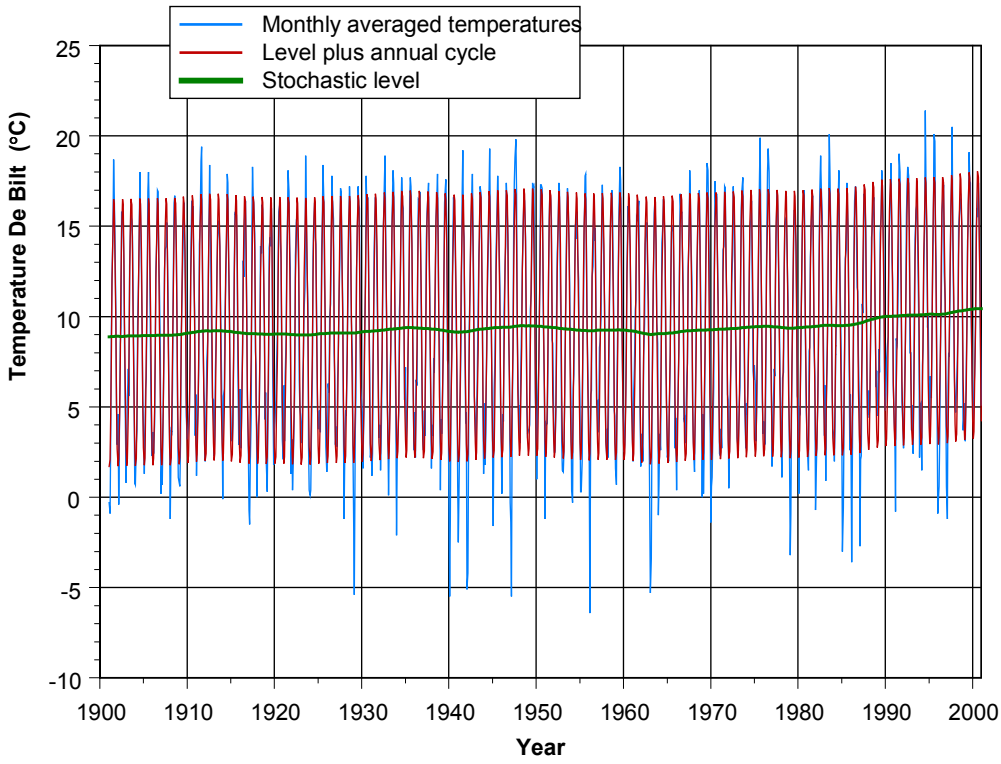
Neerslag in De Bilt vertoont een monotone toename van precies 1.0 mm per jaar. De haast schoksgwijze veranderingen zoals die bij de temperatuurreeks zichtbaar zijn sinds 1970, zijn niet terug te vinden in de neerslagreeks. Foto: H. Visser.

6.4 Jaarcyclus 1901 - 2002

6.4.1 Trend en jaarcyclus

Nu we de trends in neerslag en temperatuur onderzocht hebben, is het interessant om te kijken of er ook binnen het jaar veranderingen zijn opgetreden. Dit kunnen we testen door tegelijk met de trend een jaarcyclus te schatten en te kijken of de vorm van deze jaarcylus constant blijft over 100 jaar. Zo zou het kunnen zijn dat, naast een langjarige trend, de winters minder koud zijn geworden en tegelijkertijd de zomers minder warm.

Om deze hypothese te testen, kiezen we voor het schatten van een Stochastisch Level model voor trend in combinatie met een cyclus op basis van *maandwaarden*. De cycluslengte is daarmee dus 12. De schattingsresultaten voor temperatuur zijn gegeven in **figuur 9**. De variatie van de metingen rond de geschatte trend (dus de variantie van het verschil y_t - μ_t) neemt af met 91% door toevoeging van de cyclus. Geen verrassend resultaat want de jaargang van temperatuur is zeer sterk, zoals iedereen uit ervaring weet.



Figuur 9 *Trend- en cyclus-schatting voor maandwaarden van temperatuur over de periode januari 1901 tot en met maart 2002.*
Het Stochastic-Level-trendmodel is weer-gegeven door de groene curve, en de som van trend en jaarcyclus door de rode curve. De maandgemiddelde metingen voor station De Bilt zijn weergegeven in blauw.

Verder blijkt uit het geschatte model dat de vorm van de jaarcyclus over alle 102 jaren constant blijft (in jargon: de maximum-likelihood-schatting voor de ruiscomponent in γ_t is nul). Ofwel, er zijn in de afgelopen 102 jaar *geen significante veranderingen* opgetreden in de *vorm* van de jaargang van temperatuur.

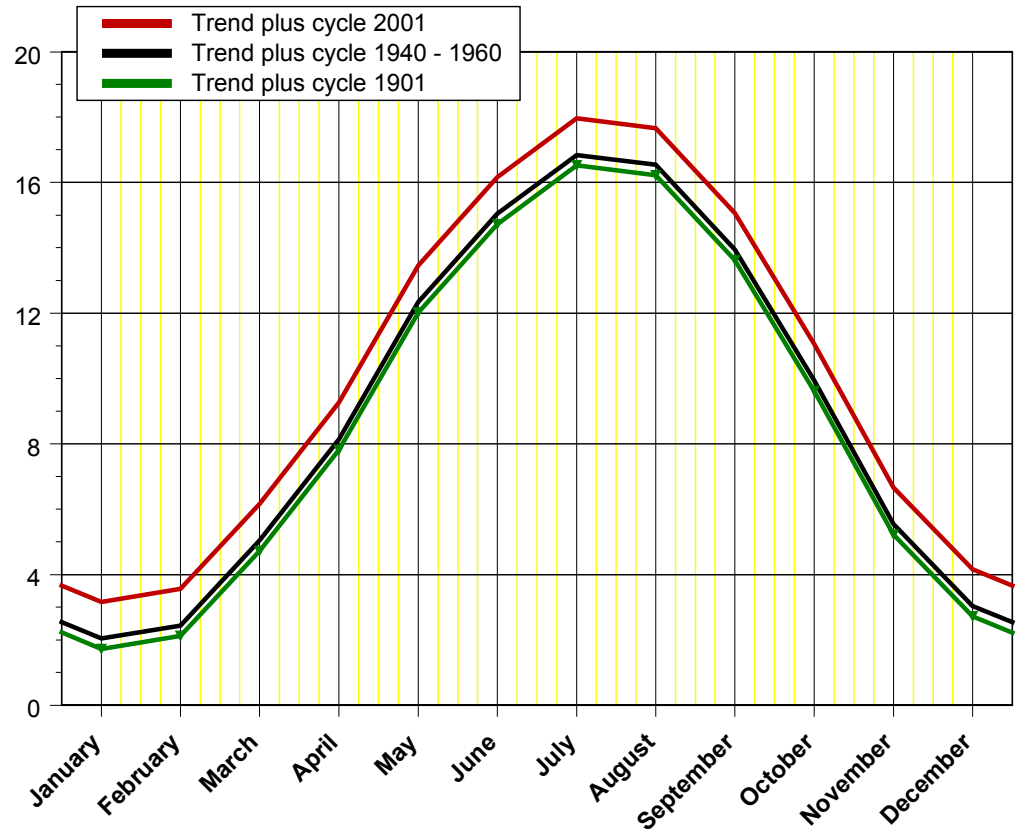
De *eeuw-gemiddelde jaargang* ziet er als volgt uit voor de maanden januari tot en met december:

-7.2, -6.8, -4.2, -1.1, 3.1, 5.8, 7.6, 7.3, 4.7, 0.7, -3.7 en –6.2 ºC.

Overeenkomstig vergelijking (2c) is de som van deze 12 temperaturen gelijk aan 0.0 ºC. De standdaardeviatie voor deze schattingen is voor elke maand hetzelfde: σ_γ = 0.16 ºC. Overigens bedraagt de standaarddeviatie $\sigma_{\mu+\gamma}$ (de rode curve in figuur 9) 0.24 ˚C aan het begin en einde van de reeks, en 0.19 ˚C in het middendeel van de reeks.

6.4.2 Verschuivingen in het groeiseizoen

Figuur 10 geeft voor drie periodes een uitvergroting van de trend plus jaarcyclus uit figuur 9. De groene lijn geeft het gemiddelde verloop voor het jaar 1901, de zwarte lijn voor het gemiddelde verloop in de periode 1940-1960, en de rode curve het gemiddelde verloop voor het jaar 2001. Uit de figuur kan afgelezen worden in hoeverre er verschuivingen zijn opgetreden in de lengte van het groeiseizoen.



Figuur 10 *Verloop van de jaarcyclus van maandgemiddelde temperaturen voor station De Bilt.*
De rode curve geeft het gemiddelde verloop voor het jaar 2001, de zwarte curve het gemiddelde verloop voor de periode 1940-1960 (trend in temperaturen stabiliseert in deze 20 jaar), en de groene curve het gemiddelde verloop voor het jaar 1901.

Voor planten is het overschrijden van 8 ºC een belangrijke grens voor het uitlopen in de lente. In de herfst is deze temperatuur van belang voor bladval. Er zijn uiteraard ook andere factoren van belang zoals neerslag en bodemtemperatuur (die veel trager reageert dan variaties in de luchttemperatuur). Maar om het eenvoudig te houden, kiezen we als definitie voor 'groeiseizoen' de periode dat de luchttemperatuur, gemeten op 150 cm hoogte, boven de 8 ºC ligt.

Als we in figuur 10 de horizontale lijn bij 8 ºC volgen, dan blijkt uit lineaire interpolatie dat in 1901 de 8 ºC -grens gemiddeld gepasseerd wordt op 16 april [1-σ betrouwbaarheidsgrenzen: 14 tot 17 april]. In 2001 is deze gemiddelde datum verschoven naar 2 april [1-σ-betrouwbaarheidsgrenzen: 31 maart tot 4 april]. Hiermee is de lente dus in 101 jaar met **2 weken vervroegd**.

Evenzo wordt in de herfst in 1901 de 8 ºC -grens gemiddeld gepasseerd op 27 oktober [1-σ betrouwbaarheidsgrenzen: 26 tot 29 oktober]. In 2001 is deze gemiddelde datum verschoven naar 5 november [1-σ betrouwbaarheidsgrenzen: 4 tot 7 november]. Hiermee valt de 8 ºC -grens in 2001 dus **ruim een week later** dan 100 jaar daarvoor. Het totale groeiseizoen, zoals boven gedefinieerd, is hiermee in 100 jaar tijd **met 23 dagen verlengd**.

We hebben ook gekeken naar de jaarcyclus in *neerslag*. Maar deze jaarcyclus verklaart slechts 9% van de variaties rond de langjarige trend in neerslag (bij temperatuur was dit 91%!). Eventuele veranderingen in de jaargang hebben daarmee slechts een heel klein effect op bijvoorbeeld plantengroei. De schattingsresultaten zijn daarom hier niet opgenomen.

Resultaten uit deze paragraaf zijn gebruikt in hoofdstuk 4 van de Milieubalans 2002. Overigens zij opgemerkt dat figuur 4.1.1 uit de Milieubalans niet ontleend is aan deze studie. Deze figuur gebruikt een ***5 ºC*** **–grens** als definitie voor de lengte van het groeiseizoen, en komt tot een verlenging van het groeiseizoen met 30 dagen. Het verschil in definitie kan dus enige verwarring geven. Resultaten over verlenging van het groeiseizoen zijn ook gerapportteerd door Oppewal (2002).

6.4.3 Persistentie

De bovenste grafiek in **Figuur 11** toont de autocorrelatiefunctie voor de residuen van het geschatte trend-cyclus-model uit figuur 9. De grafiek laat zien dat er nog enige correlatie bestaat tussen opeenvolgende maanden (R_1 is significant van nul te onderscheiden). Het is interessant om deze persistentie in de residuen te schatten. Immers als deze persistentie hoog zou zijn, dan is het mogelijk om vanuit de temperatuur van een willekeurige maand de temperatuur van de volgende maand te voorspellen!

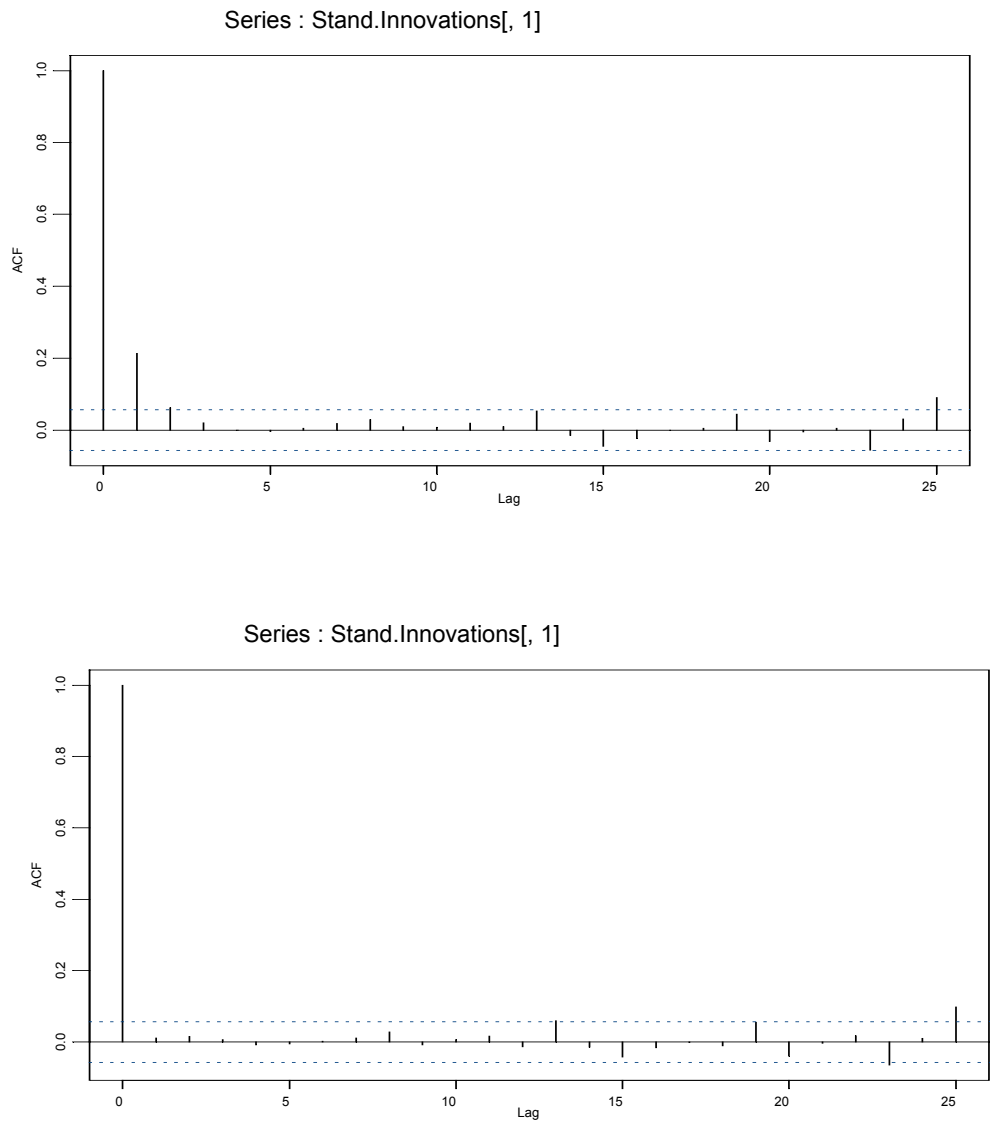
Voor het residu r_t hebben we een eenvoudig regressief model geschat: $r_t = \alpha_t\, r_{t-1} + \xi_t$. Het model geeft een relatie tussen openvolgende maandwaarden (r_t en r_{t-1}). Als de weegfactor α_t nul is dan bestaat er geen enkele relatie (= persistentie) tussen opeenvolgende maanden. Als α_t gelijk aan één is, dan ontstaat er een volledige koppeling tussen opeenvolgende maanden.

Na schatting vinden we voor de weegfactor α_t de constante waarde 0.21 ± 0.03. De weegfactor is dus statistisch significant. Echter, de verklaarde variantie door toevoegen van de verklarende variabele r_{t-1} is zeer gering, slechts 4%.

Wel is het zo dat de waarde van R_1 voor de gestandaardiseerde innovatiereeks van dit model verdwenen is. Zie onderste grafiek van figuur 11.

Het hier gevonden resultaat betekent voor de praktijk dat er weliswaar persistentie bestaat tussen opeenvolgende maanden, maar dat die te gering is om van praktische waarden te zijn voor weersvoorspellingen.

Voor referenties over persistentie tussen opeenvolgende maanden verwijzen we naar Visser (1995) en de daar vermelde literatuur.



Figuur 11 *Autocorrelatiefuncties op de residureeks voor het trend-cyclus-model uit figuur 9 (boven) en voor het autoregressieve model op de residuen van dat model (onder).*

De trendwaarde van de jaargemiddelde temperatuur in De Bilt is in 2002 gestegen naar 10.5 ± 0.4 °C. Daarmee is deze trendwaarde voor het eerst in 102 jaar significant groter geworden dan 10.0 °C. Aan het begin van de reeks, in 1901, bedraagt de trendwaarde 8.9 ± 0.4 °C. Foto: H. Visser.

7. PM_{10}-CONCENTRATIES IN EEN STAD

Binnen de luchtverontreinigingsproblematiek speelt fijn stof een belangrijke rol. Wereldwijd tonen statistische modellen positieve associaties tussen stofconcentraties enerzijds en gezondheid van mensen anderzijds. Voor een recent overzicht over deze problematiek zie Brunekreef en Holgate (2002), en Hoek et al. (2002).

Een belangrijke beleidsvraag op dit moment is of fijn-stof-concentraties in Nederland een dalende tendens vertonen en zo ja, of deze tendens toe te kennen is aan dalende emissies van fijn stof. Hierna zal fijn stof aangeduid worden met de afkorting PM_{10} ofwel *Particulate Matter* met een aërodynamische diameter kleiner dan 10 µm. Voor meer informatie over PM_{10} verwijzen we naar Visser, Buringh en Breugel (2001), en Buringh en Opperhuizen (2002a,b).

Om de vraag naar aanwezige trends te beantwoorden, is het noodzakelijk om de invloed van meteorologische condities in de verschillende jaren uit de concentraties te filteren. Niet alleen is het zo dat meteorologische condities van jaar op jaar sterk kunnen variëren zoals blijkt uit de figuren 7 en 8, maar ook dat er systematische veranderingen ingezet zijn sinds de jaren 70 van de vorige eeuw (figuur 7).

Door Dekkers en Noordijk (1997) is een methode ontwikkeld om de invloed van meteorologie uit concentratie-meetreeksen te filteren. De methode is gebaseerd op Regressieboom-Analyse en is in staat om *sterke niet-lineare verbanden* tussen concentraties enerzijds en meteorologie anderzijds op te sporen. Deze methode is later verfijnd door Visser en Noordijk (2002). Structurele tijdreeksmodellen gaan uit van zuiver additieve of zuiver multiplicatieve modellen (modellen (1a) en (1b)). Daarentegen kunnen niet-lineaire fenomenen gedetecteerd worden door de weegfactoren in het regressiemodel (2a) tijdsafhankelijk te maken.

Hierna laten we als voorbeeld een analyse zien van een fijn-stof-reeks voor één station, namelijk Eindhoven. We laten zien hoe de relatie tussen PM_{10} (y_t) en meteorologie geschat kan worden met Structurele tijdreeksanalyse. Hierbij gaan we uit van drie meteorologische variabelen $x_{1,t}$ $x_{2,t}$ en $x_{3,t}$ die via Regressieboom-analyse zijn opgespoord. Het gaat om temperatuur, neerslag en windsnelheid. Visser en Noordijk vonden verder dat maandwaarden als tijdstap veel betere resultaten geeft dat modellering op basis van dagwaarden. Daarom hebben we hierna ook maandwaarden gebruikt. Zie verder Visser en Noordijk (2002).

7.1 Data

Maandgemiddelde PM_{10}-data voor station Eindhoven (stationscode 236) zijn te vinden in de file **PM10Eindhoven.dat** die te vinden is op de Installatie-CD van TrendSpotter. De file heeft het format (6X,F9.3,F9.5,F12.7,F11.6,F9.6). De kolommen hebben de volgende betekenis: tijd in maanden, PM_{10}-concentratie in $\mu g/m^3$, temperatuur in ˚C, neerslag in mm, en windsnelheid in m/sec.

De ligging van het meetstation is weergegeven in **figuur 12**.

Figuur 12 *Station Eindhoven-Genovevalaan. De behuizing van de PM_{10}-monitor is aangeven door de groene pijl.* Foto: RIVM-LLO.

7.2 Invloed meteorologie

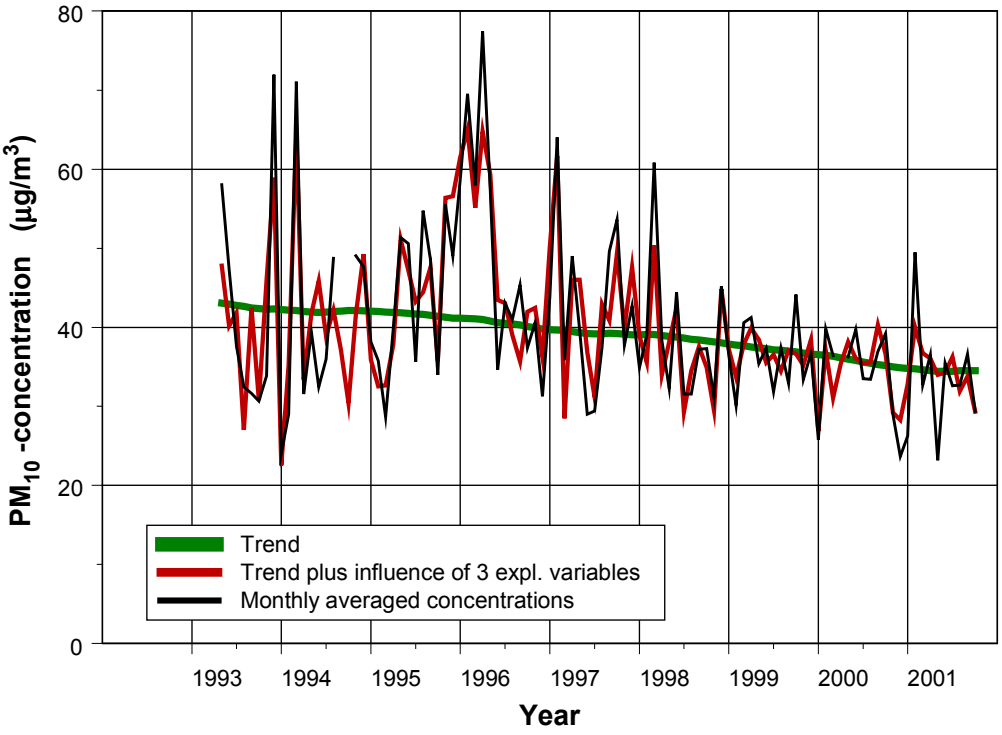
Voor de analyse hebben we het volgende model gekozen:

$$y_t = \mu_t + \alpha_{1,t} * x_{1,t} + \alpha_{2,t} * x_{2,t} + \alpha_{3,t} * x_{3,t} + \varepsilon_t \qquad (7)$$

Identificatie van het trendmodel uit model (7) voeren we uit door het bekijken van de autocorrelatiefuncties op de reeksen y_t, $\Delta y_t = y_t - y_{t-1}$, en $\Delta^2 y_t = y_t - 2y_{t-1} + y_t$ (§A3.1). We vinden dat het Stochastic-Level-model het beste voldoet.

Elk van de variabelen $x_{1,t}$, $x_{2,t}$ en $x_{3,t}$ uit model (7) is vooraf geschaald naar gemiddelde 0.0 en standaarddeviatie 1.0 (standaardisatie). Deze transformatie vooraf heeft het voordeel dat de groottes van de weegfactoren $\alpha_{1,t}$, $\alpha_{2,t}$ en $\alpha_{3,t}$ nu onderling vergeleken kunnen worden.

De schattingsresultaten zijn gegeven in **figuur 13A**. De figuur laat zien dat er een goede overeenkomst bestaat tussen de metingen (zwarte curve) en het model (rode curve). Het model verklaart 72% van de variatie van de concentraties rond de geschatte trend (groene curve). Ter vergelijk, via Regressieboom-Analyse wordt 65% verklaard van de variatie rond een constant niveau.

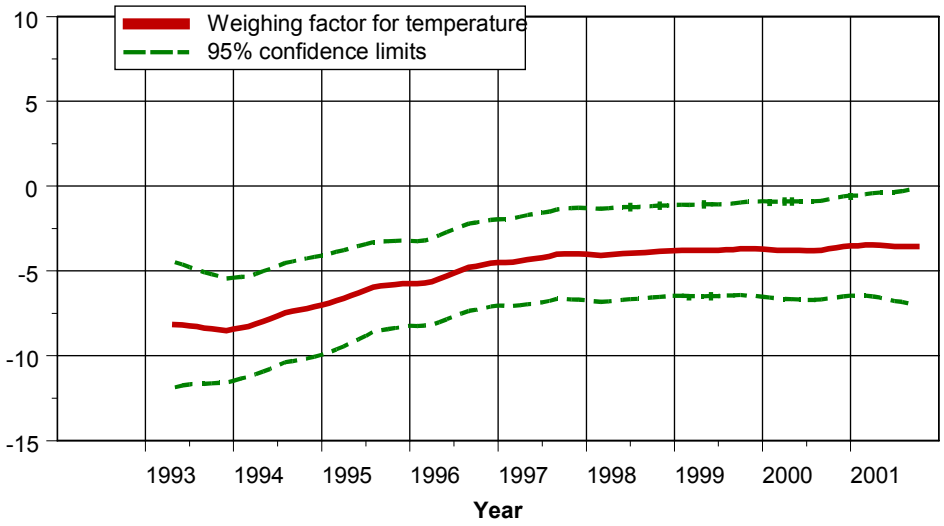


Figuur 13A *Schattingsresultaten voor maandgemiddelde PM_{10}-concentraties (rode curve). De metingen zijn voor station Eindhoven (de zwarte curve).*
De groene curve geeft de geschatte trend via het Stochastische niveaumodel. Het totale model bestaat uit het stochastisch niveau in combinatie met de invloed van drie verklarende variabelen: temperatuur, neerslag en windsnelheid. De tijdsafhankelijke weegfactoren $\alpha_{1,t}$, $\alpha_{2,t}$ en $\alpha_{3,t}$ zijn op de volgende pagina afgebeeld. Elk van de variabelen $x_{1,t}$, $x_{2,t}$ en $x_{3,t}$ is vooraf geschaald naar gemiddelde 0.0 en standaarddeviatie 1.0 (standaardisatie).

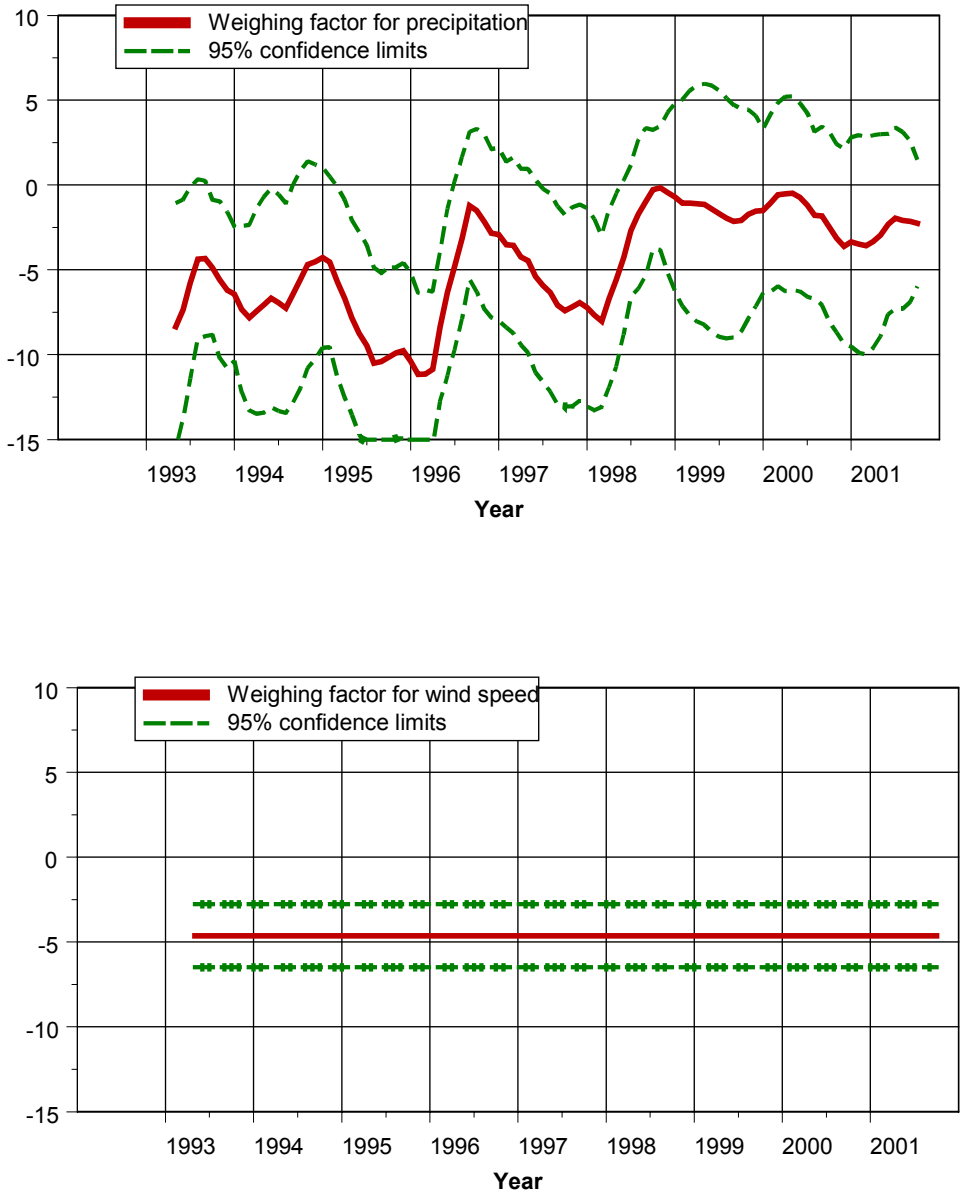
De weegfactoren zijn weergegeven in **figuur 13B**. De variabelen temperatuur en neerslag vertonen een sterk tijdsafhankelijk gedrag. Zulk gedrag kan twee oorzaken hebben. In de eerste plaats kan de relatie tussen y_t en de bijbehorende x_t structureel wijzigen in de tijd. Een voorbeeld van zo'n proces is de groei van bomen (y_t is jaarringdikte per jaar) en neerslag in het groeiseizoen, terwijl vanaf een bepaald jaar de grondwaterstand gaat dalen. Aanvankelijk heeft neerslag geen invloed op de groei omdat de wortels bij het grondwater kunnen. Bij voortschrijdende daling van het grondwaterniveau zal de boom echter steeds meer afhankelijk worden van neerslag voor zijn vochtvoorziening. Hierdoor zal er een significante positieve relatie ontstaan tussen jaarlijkse houtaanwas en neerslaghoeveelheid in het groeiseizoen (Visser, 1994).

De tweede oorzaak welke hier veel waarschijnlijker is, is dat er een structureel niet-lineaire relatie bestaat tussen y_t en een specifieke $x_{i,t}$. Als bijvoorbeeld de werkelijke relatie kwadratisch is, dus y_t is lineair gekoppeld aan $x_{i,t}^2$, maar we koppelen *in ons model* y_t aan $\alpha_{i,t} * x_{i,t}$, dan zal de weegfactor $\alpha_{i,t}$ tijdvariërend gedrag vertonen.

Zo blijken de maanden voor neerslag met de meest negatieve waarden in $\alpha_{2,t}$ overeen te komen met extreem droge en koude maanden (eind 1995 tot begin 1996, en eind 1997 tot begin 1998). Wat er in deze maanden gebeurt, is dat door langdurige droogte en wind uit oostelijke richtingen de niet-begroeide landerijen (het is winter) zover opdrogen dat opwaaiend stof een factor van belang wordt. Dit stof kan uit Nederland zelf afkomstig zijn maar ook vanuit Duitsland worden aangevoerd. Hiermee is ook het niet-lineaire karakter tussen PM_{10} enerzijds en neerslag en in mindere mate temperatuur anderzijds verklaard.



Figuur 13B Weegfactoren voor temperatuur ($\alpha_{1,t}$), voor neerslag ($\alpha_{2,t}$), en voor windsnelheid ($\alpha_{3,t}$).

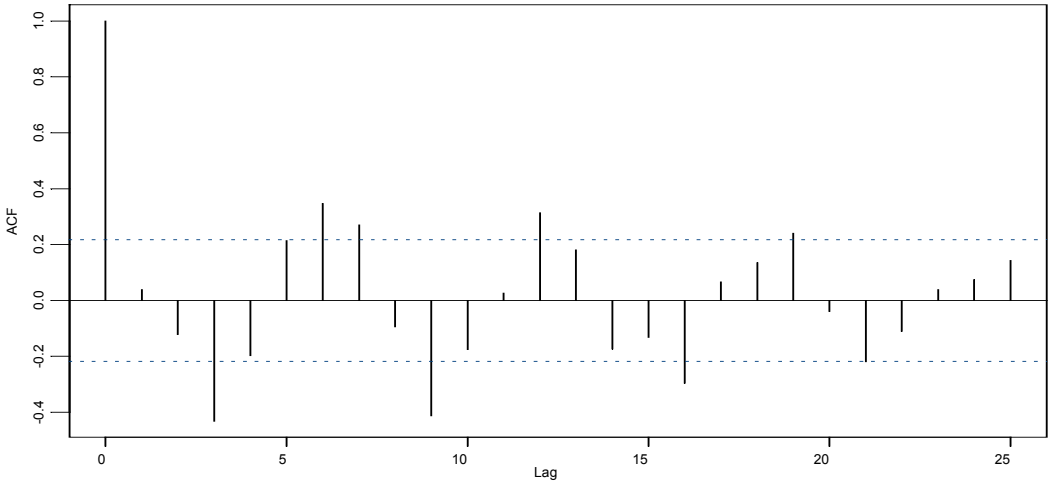


Figuur 13B Vervolg

De relatie met windsnelheid ($\alpha_{3,t}$) is constant en negatief. Dit verband is goed verklaarbaar. Bij lineair-toenemende windsnelheden neemt ook het volume lucht lineair toe waarin stofemissies gemengd worden. De concentratie van PM_{10} zal dan dus lineair afnemen.

7.3 Persistentie en cyclus

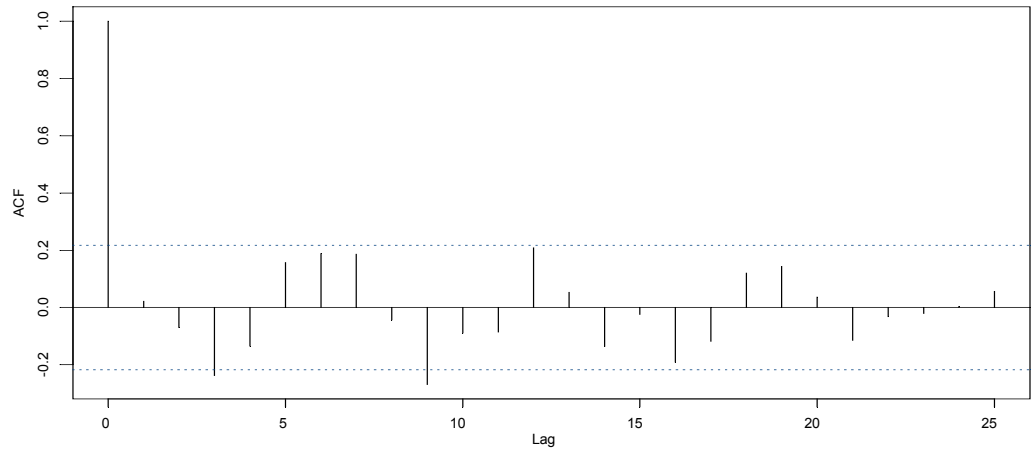
De autocorrelatiefunctie op de residureeks van model (7) is gegeven in **figuur 14** (cf. §2.6). De figuur laat zien dat er geen persistentie meer aanwezig is in de maandelijkse innovaties (R_1 en R_2 zijn niet significant). Wel is er nog een duidelijk cyclisch signaal aanwezig in de residuen. De periode van de cyclus is precies *een half jaar* en dempt geleidelijk uit. We zouden nu het model kunnen herschatten met het toevoegen van een cyclus-component.



Figuur 14 Autocorrelatiefunctie voor de residureeks van model (7).

Om zo'n herschatting uit te voeren, voegen we aan model (7) een cyclus toe met periodelengte 6. Na schatting vinden we inderdaad een significante, maar numeriek geringe half-jaar-cyclus. Schatten we opnieuw een autocorrelatiefunctie op de residuen van het uitgebreide model, dan is de cylus nagenoeg verdwenen (**figuur 15**).

Een verklaring voor de half-jaar-cyclus is niet helemaal duidelijk. Mogelijk is er een half-jaar-cyclus aanwezig in de verkeersdichtgheid in de straat (tussen zomervakantie en kerstvakantie ligt precies een half jaar). Helaas beschikken we niet over de verkeersdichtheden over het jaar om deze hypothese te verifiëren.



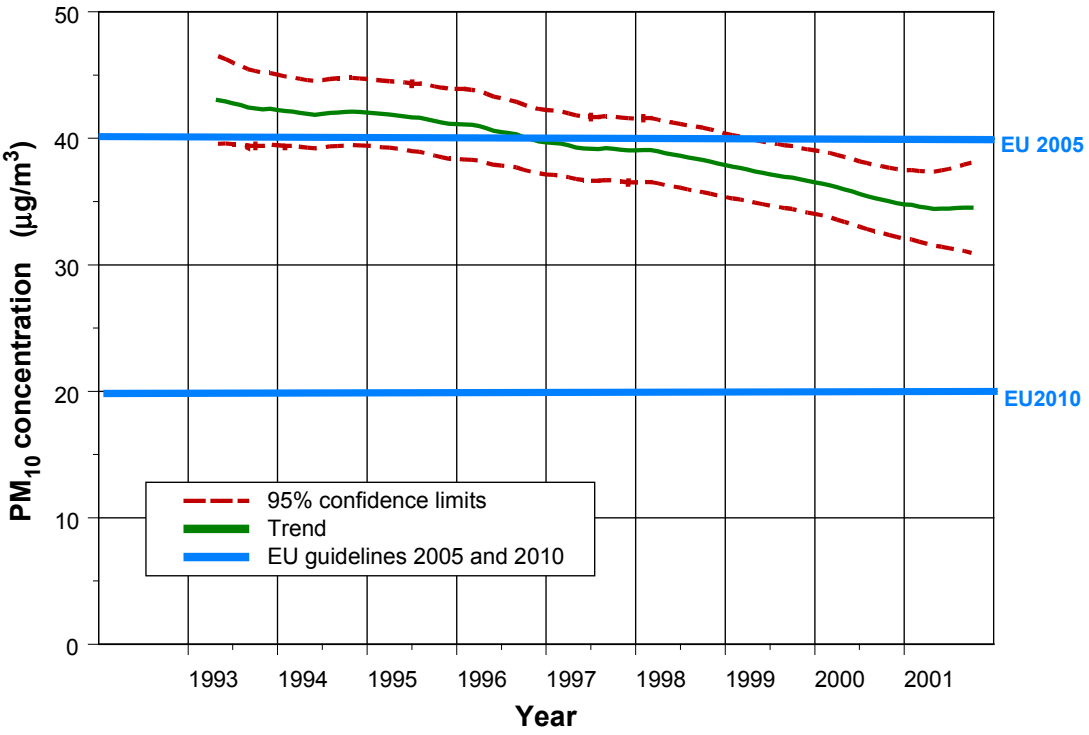
Figuur 15 Autocorrelatiefunctie voor de gestandaardiseerde innovaties van model (7), maar na toevoeging van een half-jaar-cyclus aan dit model.

7.4 Is de concentratiedaling statistisch significant?

Figuur 16A geeft de trend uit figuur 13A weer met 95%-betrouwbaarheidsintervallen. Bovendien is in de figuur de EU-norm voor jaargemiddelden weergegeven, namelijk 40 µg/m^3. Deze norm moet in 2005 gehaald zijn. Het ligt zelfs in de bedoeling om de jaargemiddelde norm van 40 naar 20 µg/m^3 te verlagen. De laatstgenoemde waarde moet in het jaaar 2010 gehaald zijn. Beide grenswaarden gelden voor alle denkbare locaties, dus ook voor het station in Eindhoven.

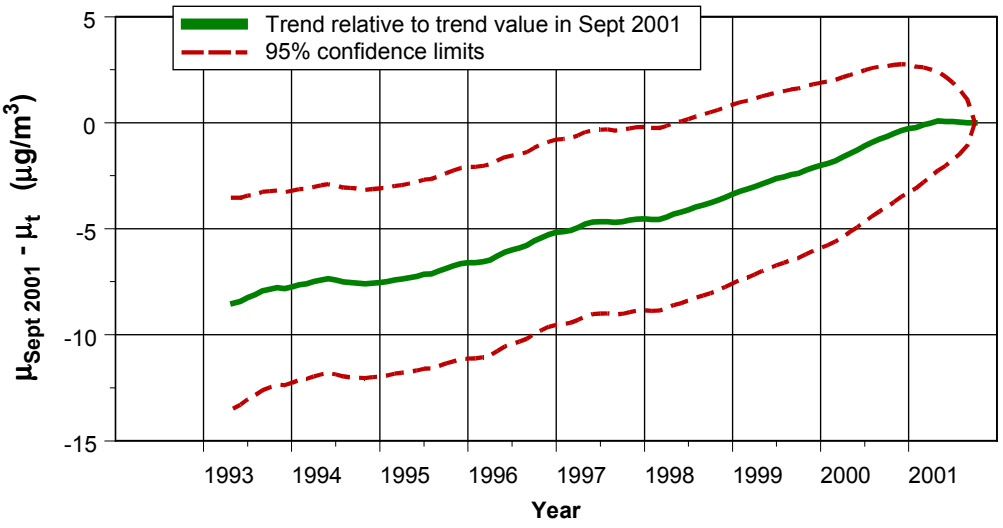
We zien aan figuur 16A dat de trendwaard in april 1993 *boven* de EU-norm-2005 ligt en op de grens van significantie ($\alpha = 0.05$, tweezijdig). Maar in september 2001 ligt de trendwaarde ruim *onder* deze norm. Verder is het duidelijk dat de EU-norm-2010 niet snel gehaald zal worden op dit station!

Bovenstaande conclusie blijkt niet alleen op te gaan voor dit straatstation, maar ook voor regionale PM_{10}-stations. Dit blijkt uit figuur 8 van Buringh en Opperhuizen (2002a). De meteo-gecorrigeerde trendwaarde in 1992 bedraagt 44 µg/m^3, gemiddeld over 9 regionale stations. In 2001 bedraagt de gemiddelde waarde 32 µg/m^3.



Figuur 16A Trend in PM_{10}-concentraties (groen) met 95%-betrouwbaarheidsintervallen (rood). De EU-richtlijnen voor jaargemiddelde concentraties voor de jaren 2005 en 2010 zijn weergegeven in blauw.

In **figuur 16B** is het verschil $\mu_{sept\ 2001} - \mu_t$ gegeven (gelijkenis met 'iets anders' berust geheel op toeval). De figuur laat zien dat de laatste trendwaarde significant groter is dan alle trendwaarden vóór 1998 (bovenste betrouwbaar-heidsgrens wordt negatief). Over de periode 1998 – 2001 is de daling niet significant.
Het verschil $\mu_{sept\ 2001} - \mu_{april\ 1993}$ bedraagt $8.5 \pm 5.0\ \mu g/m^3$.



Figuur 16B De differentie $\mu_{sept\ 2001} - \mu_t$ met 95%-betrouwbaarheidsintervallen.

Conclusie is dat

- *over de hele periode gezien de trendmatige daling statistisch significant is. De daling bedraagt 8.5 $\mu g/m^3$;*
- *de gemiddelde concentratie in 1993 boven de EU-richtlijn voor 2005 ligt, op de grens van significantie (waarde is 43 $\mu g/m^3$ en richtlijn is 40 $\mu g/m^3$). In 2001 is de gemiddelde concentratie statistisch significant gedaald tot ruim onder de richtlijn (waarde is 34.5 $\mu g/m^3$).*

Fijn-stof-concentraties zijn in 9 jaar tijd gedaald van boven de jaargemiddeld EU-richtlijn voor het jaar 2005 naar ruim onder deze richtlijn (40 µg/m^3). Maar de EU-richtlijn voor het jaar 2010 (20 µg/m^3) lijkt vooralsnog onhaalbaar. Foto: H. Visser.

Literatuur

Anderson, B.D.O., Moore, J.B., 1979. Optimal Filtering. Prentice-Hall, Inc. Englewood Cliffs, New Jersey.

Box, G.E.P., Jenkins, G.M., 1976. Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco.

Brakel, J. van den, Visser, H., 1996. The influence of environmental conditions on tree-ring series of Norway Spruce for different canopy and vitality classes. Forest Science 42 (2), pp. 206-219.

Brunekreef, B., Holgate, S.T., 2002. Air pollution and health. The Lancet, 360, pp. 1233-1242.

Buringh, E., Opperhuizen, A. (editors), 2002a. On health risks of ambient PM in the Netherlands. Executive Summary. RIVM-TNO-ECN-Univ.Utrecht report. Printed RIVM Bilthoven.

Buringh, E., Opperhuizen, A.(editors), 2002b. On health risks of ambient PM in the Netherlands. RIVM report, in press.

Chatfield, 1989. The analysis of time series, an introduction. Chapman and Hall, London.

De Jong, P., Mackinnon, M.J., 1988. Covariances for smoothed estimates in state space models. Biometrika, 75(3), pp. 601-602.

Dekkers, A.L.M., 2001. S-PLUS voor het RIVM. Krachtig statistisch software gereedschap. RIVM-rapport 422 516 001.

Dekkers, A.L.M., Noordijk, H., 1997. Correctie van atmosferische concentraties voor meteorologische omstandheden. RIVM rapport 722101 024.

Durbin, J., Koopman, S.J., 2001. Time series analysis by state space methods. Oxford Statistical Science Series.

Harvey, A.C., 1984. A unified view of statistical forecasting procedures, Journal of Forecasting, 3, pp. 245-275.

Harvey, A.C., 1989. Forecasting, structural time series models and the Kalman filter. Cambridge University Press, Cambridge.

Harvey, A.C., 1993. Time series models. Harvester Wheatsheaf, New York.

Hoek, G., Brunekreef, B, Goldbohm, S., Fischer, P., Brandt, P.A. van den, 2002. Association between mortality and indicators of traffic-related air pollution in the Netherlands: a cohort study. The Lancet, 360, pp. 1203-1209.

IPCC, 2001. Climate change 2001. The scientific basis. Cambridge University Press.

Jaarsveld, J.A. van, 1995. Modelling the long-term behaviour of pollutants on various scales. PhD. Thesis, University of Utrecht.

Kitagawa, G., 1981. A nonstationary time series model and its fitting by a recursive filter. J. Time Series Anal., Vol. 2(2), pp. 103-116.

Koopman, S.J., maart 2000. Met het Kalman filter vooruit. Oratie bij de VU, Amsterdam.

Koopman, S.J., Harvey, A.C., Doornik, J.A., Shephard, N., 1999. STAMP 6.0 Structural Time series Analyser, Modeller and Predictor. Published by Timberlake Consultants Ltd.

Milieubalans 2002. Het Nederlandse milieu verklaard. Uitgave RIVM Bilthoven, Kluwer.

Millard, S.P., Neerchal, N.K., 2001. Environmental Statistics with S-PLUS. CRC press Boca Raton, Florida.

Montgomery, D.C., Peck, E.A., 1982. Introduction to Linear Regression Analysis. John Wiley & Sons, Inc. New York.

Oldenborgh, G.J., Komen, G., 2001. Hangt het warme weer de laatste tijd in Nederland samen met het versterkte broeikaseffect? Meteorologica, oktober nummer.

Oppewal, J.R., 2002. Groeiseizoen in eeuw tijd 3 weken langer. Boerderij 88(4), pp. 13-14.

Scholten, R.D.A., 1992. Zonnevlekken en klimaatveranderingen: is de relatie schijn of werkelijkheid? KEMA-rapport 92252-MLU 92-4011.

Van der Wal, J.T., Janssen, L.H.J.M., 1996. Description and analysis of ambient fine particle concentrations in the Netherlands. RIVM report no. 723301 007.

Van der Wal, J.T., Janssen, L.H.J.M., 2000. Analysis of spatial and temporal variations of PM_{10} concentrations in the Netherlands using Kalman filtering. Atm. Environment 34, pp. 3675-3687.

Visser, H., 1994. Regression models with time-varying parameters: applications in the environmental sciences. Proefschrift Universiteit van Amsterdam.

Visser, H., 1995. Structural time series models and the Kalman filter: applications in climatology. Proceedings of the Sixth International Meeting on Statistical Climatology, Galway, Ierland, pp. 7-10.

Visser, H., 2003. Detection of environmental change. Trendspotter, a software package for estimating Structural Time Series models. RIVM report, in preparation.

Visser, H., Buringh, E., Breugel, P.B., 2001. Composition and origin of airborne particulate matter in the Netherlands. RIVM report 650010 029.

Visser, H., Habets, M., Leene, R., 1990. KALFIMAC, a software package to analyse time series with trend, cycli and explanatory variables. Release 3.1. KEMA report 50385-MOF 89-3282.

Visser, H., De Koningh, M.C.J., 2000. Associations between human health and air pollution: a statistical minefield. KEMA report 50030065-KPS 00-3048.

Visser, H., Molenaar, R.J., 1990. Estimating trends in tree-ring data. Forest Science, Vol 36 (1), pp. 87-100.

Visser, H., Molenaar, R.J., 1992. Estimating trends and stochastic response functions in dendroecology with an application to fir decline. Forest Science, Vol 38 (2), pp. 221-234.

Visser, H., Molenaar, R.J., 1995. Trend estimation and regression analysis in climatological time series: an application of structural time-series models and the Kalman filter. Journal of Climate, 8(5), pp. 969-979.

Visser, H., Noordijk. H., 2002. Correcting air pollution time series for meteorological variability. With application to regional PM_{10} concentrations. RIVM report 722601007.

Visser, H., Rőmer, F.G., 1999. PM_{10} concentrations in the Netherlands: measurements and interpretation. KEMA report 99530102-KPS/SEN 99-2039.

Visser, H., Rőmer, F.G., 2000. Ozone concentrations in the Netherlands: measurements and interpretation. KEMA report 50030058-KPS/SEN 00-3044.

Visser, H., Vermeulen, A.T., Janssen, L.H.J.M., Draaijers, G.P.J., Hollander, J.C.Th., Roemer, M.G.M., Vosbeek, M.E.J.P., Van der Wal, J.T., 1999. Methane concentrations in the Netherlands: measurements and interpretation. KEMA/ECN/RIVM/TNO report 64720-KST/ENR 99-2006.

Appendix A Theorie

A.1 Inleiding

Bij het modelleren van tijdreeksen kan onderscheid worden gemaakt tussen structurele modellen en niet-structurele modellen. Tot de laatste groep behoren bijvoorbeeld Box-Jenkins ARIMA modellen (Box en Jenkins, 1976). Het structurele tijdreeksmodel is gebaseerd op een additieve decompositie van de meting y_t in de componenten trend, seizoensinvloed, verklarende variabelen en ruis (Harvey, 1989). Dit geschiedt zodanig dat deze afzonderlijke componenten direct interpreteerbaar zijn:

$$y_t = trend + seizoen + verkl.var. + ruis \qquad \text{(A.1)}$$

De afzonderlijke componenten uit (A.1) zijn niet direct observeerbaar. Zij worden geschat uit een te definiëren model en de metingen y_t. Bij niet-structurele modellen zoals ARIMA modellen, is een additieve decompositie in principe niet mogelijk of in ieder geval erg lastig.

Het Structurele Tijdreeksmodel gaat uit van de additieve decompositie (A.1). Indien deze modelformulatie niet verantwoord is, dan kan worden geprobeerd om hier met behulp van een geschikte transformatie naar toe te werken. Een multiplicatief model kan bijvoorbeeld via een logaritmische transformatie additief worden gemaakt:

$$y_t = trend \;\; x \;\; seizoen \;\; x \;\; verkl.var. \;\; x \;\; ruis$$

$$\Leftrightarrow$$

$$\log(y_t) = \log(trend) + \log(seizoen) + \log(verkl.var.) + \log(ruis) \qquad \text{(A.2)}$$

Bij het modelleren van tijdreeksen kan het een beperking zijn dat de parameters voor trend, seizoensinvloed en verklarende variabelen over het hele tijdsinterval constant moeten worden gekozen. Harvey (1984, 1989) heeft daarom een Structureel Tijdreeksmodel voorgesteld waarbij deze parameters geleidelijk mogen veranderen. Om analyse hiervan eenvoudig mogelijk te maken, moet het structureel model in de zogenaamde *toestandsruimte-vorm* worden gebracht. Een model in toestandsruimte-vorm kan namelijk met het Kalmanfilter geschat worden. Het Kalmanfilter levert schattingen voor variabelen en paramaters uit het oorspronkelijke model met mooie statistische eigenschappen (waarover meer hierna).

Het toestandsruimte-model wordt in §A.2 beschreven, terwijl de Structurele Tijdreeksmodellen die in TrendSpotter worden gebruikt, in §A.3 gegeven worden. Alle Structurele modellen worden geschat met het Kalmanfilter. Dit filter wordt in detail in §A.4 beschreven. Het zij hier vermeld dat de ARIMA-trendmodellen in deze bijlage niet wiskundig beschreven worden. Hiervoor wordt verwezen naar Visser en Molenaar (1995).

A.2 Het toestandsruimtemodel

Gegeven zijn de N observaties: y_N, y_{N-1}, ... ,y_3, y_2, y_1. De observaties worden verondersteld te zijn opgebouwd uit een signaal dat verstoord wordt door een ruisterm:

$$y_t = signaal + ruis \qquad \text{(A.3)}$$

Het signaal is ten gevolge van de verstoringen niet direct waarneembaar maar kan met behulp van een stochastisch filter zo optimaal mogelijk worden geschat. Daartoe moet eerst een toestandsruimte-model worden geformuleerd dat vervolgens met het discreet Kalmanfilter kan worden geanalyseerd.

Het Kalmanfilter maakt het mogelijk om (onder bepaalde veronderstellingen):
- optimale schattingen te geven van het niet observeerbaar signaal door middel van filteren of smoothen;
- optimale voorspellingen te geven van toekomstige observaties;
- het optimaal schatten van ontbrekende observaties.

Het toestandsruimte-model is gebaseerd op twee aannames, uitgedrukt in twee vergelijkingen:
- Het signaal kan worden uitgedrukt in een lineaire combinatie van een aantal 'toestandsvariabelen' α_t. Dit levert de observatievergelijking of meetvergelijking:

$$y_t = z_t' \alpha_t + \xi_t \qquad \text{(A.4)}$$

  In deze vergelijking zijn y_t en ξ_t scalair, en zt en ξt zijn vectoren met lengte k.
- De toestandsvariabelen veranderen in de tijd. Dit wordt mogelijk gemaakt door elke toestandsvariabele met een eigen stochastisch proces te modelleren. Dit levert de systeemvergelijking:

$$\alpha_t = T_t \alpha_{t-1} + \eta_t \qquad \text{(A.5)}$$

  In deze vergelijking zijn α_t en η_t vectoren ter lengte k, en T is een k*k matrix.

De systeemvergelijking geeft aan hoe de toestandsvariabelen van tijdstip t-1 overgaan naar tijdstip t. De matrix T wordt opgesteld op grond van theoretische overwegingen over het gedrag van het systeem.

De modelveronderstellingen die aan het toestandsruimte model ten grondslag liggen om met het Kalmanfilter te kunnen worden geanalyseerd, luiden:

$$\begin{aligned}
&\mathrm{cov}(\xi_{t1}, \xi_{t2}) = \delta_{t1,t2}\,\sigma^2 h_t && (\forall t_1, \forall t_2) && \text{(A.6.a)}\\
&\mathrm{cov}(\eta_{t1}, \eta_{t2}) = \delta_{t1,t2}\,\sigma^2 Q_t && (\forall t_1, \forall t_2) && \text{(A.6.b)}\\
&\mathrm{cov}(\eta_{t1}, \xi_{t2}) = 0 && (\forall t_1, \forall t_2) && \text{(A.6.c)}\\
&\mathrm{cov}(\eta_t, \alpha_0) = 0 && (\forall t) && \text{(A.6.d)}\\
&\mathrm{cov}(\xi_t, \alpha_0) = 0 && (\forall t) && \text{(A.6.e)}\\
&E(\eta_t) = 0 && (\forall t) && \text{(A.6.f)}\\
&E(\xi_t) = 0 && (\forall t) && \text{(A.6.g)}
\end{aligned}$$

Niet noodzakelijke modelveronderstellingen zijn normaliteitsaannames voor de ruistermen:

$$\begin{aligned}
&\xi_t \approx N(0,\sigma^2 h_t) && \text{(A.6.h)}\\
&\eta_t \approx N(0,\sigma^2 Q_t) && \text{(A.6.i)}\\
&\alpha_0 \text{ normaal verdeeld} && \text{(A.6.j)}
\end{aligned}$$

In deze aannames staat ‘cov’ voor covariantie, ‘E’ voor verwachtingswaarde, ‘$\delta_{t1,t2}$’ voor het Kronecker symbool (zie hieronder), ‘$\forall$’ staat voor ‘voor alle’ en ‘$\approx N(0,\sigma^2 h_t)$’ voor ‘is normaalverdeeld met verwachting 0 en variantie $\sigma^2 h_t$.

De variabelen in (A.4) en (A.5) hebben de volgende betekenis:

y_t : observatie (scalar)
α_t : toestandsvector (k*1)
z_t : vector (k*1)
T_t : systeemmatrix (k*k)
ξ_t : storingsterm van de meetvergelijking (scalar, meetruis)
η_t : storingstermen van de systeemvergelijking (k*1, systeemruis)
$\sigma^2 Q_t$: covariantiematrix van η_t (k*k)
$\sigma^2 h_t$: variantie van ξ_t (scalar)
$\delta_{t1,t2}$: Kronecker symbool ($\delta_{t1,t2} = 1$ als $t_1 = t_2$ en $\delta_{t1,t2} = 0$ als $t_1 \neq t_2$)

Het bovenstaande toestandsruimte-model is een *univariaat model*, dat wil zeggen dat y_t een scalar is. Het toestandsruimte-model kan ook multivariaat worden geformuleerd waarbij yt een vector (p*1) is met p waarnemingen, ξ_t een vector (p*1) is met p ruistermen, h_t een covariantie matrix (p*p) is en z_t een matrix (p*k), (Harvey, 1989, H.3.1, en Koopman, et al. 1999).

In het meest algemene geval zijn T_t, Q_t, z_t en h_t tijdafhankelijk. Indien al deze variabelen constant zijn, dat wil zeggen *tijdonafhankelijk*, dan spreekt men van een tijdinvariant model. Zie Harvey (1989, §3.3) voor een overzicht van speciale eigenschappen van tijdinvariante modellen.

In TrendSpotter worden de volgende veronderstellingen gemaakt:
- de matrices T_t en Q_t zijn tijdonafhankelijk en bekend vóór dat het toestandsruimte-model met behulp van het Kalmanfilter wordt geschat. Indien Q niet bekend is, dan kunnen door optimalisatie van de likelihood-functie de elementen van Q bepaald worden;
- Q is een diagonaal-matrix ($Q_{ij} = \delta_{ij}Q_{ii}$) met op de hoofddiagonaal de varianties van de ruisvector η_t. Dat wil zeggen dat alle componenten van de systeemruis onderling ongecorreleerd zijn;
- z_t en h_t zijn bekende tijdfuncties. Er geldt dat $h_t = 1$, tenzij de observatie y_t als ontbrekend wordt beschouwd.

Modelveronderstelling (A.6.h), (A.6.i) en (A.6.j) zijn niet noodzakelijk voor de geldigheid van het Kalmanfilter. Indien (A.6.h), (A.6.i) en (A.6.j) gelden, dan hebben de schatters van de toestandsvariabelen die het Kalmanfilter levert, mooiere eigenschappen (Harvey, 1984 in hoofdstuk 1).

Omdat σ^2 apart bepaald wordt uit de modelschattingen, is het duidelijker om σ^2 expliciet te scheiden van de variantie van ξ_t en η_t. Deze notatie is geïntroduceerd door Harvey (1984).

A.3 Structurele tijdreeksmodellen

In deze paragraaf geven we de wiskundige formulering van de Structurele Tijdreeksmodellen uit hoofdstuk 2. We geven steeds de formulering in de toestandvorm (A.4) en (A.5).

A.3.1 Trendmodellen

Local Linear trendmodel

Het Local Linear (LL) trendmodel wordt in de literatuur ook wel aangeduid als het 'linear growth model' (Chatfield 1989, §10.1). Voor een beschrijving zie Harvey (1989, §2.3.2). Het is als volgt gedefinieerd:

Meetvergelijking:

$$y_t = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} \mu_{1,t} \\ \mu_{2,t} \end{pmatrix} + \xi_t \Leftrightarrow y_t = z_t' \alpha_t + \xi_t$$

$$\xi_t \approx NID(0, h_t \sigma^2) \qquad \text{(A.7)}$$

Systeemvergelijking:

$\mu_{1,t}$: nivoparameter

$\mu_{2,t}$: trend- of hellingparameter

$$\begin{pmatrix} \mu_{1,t} \\ \mu_{2,t} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_{1,t-1} \\ \mu_{2,t-1} \end{pmatrix} + \begin{pmatrix} \lambda_t \\ \zeta_t \end{pmatrix} \quad \Leftrightarrow \quad \alpha_t = T\,\alpha_{t-1} + \eta_t$$

$$\eta_t \approx NID(0, \sigma^2 Q) \qquad Q = \begin{pmatrix} \sigma_\lambda^2 & 0 \\ 0 & \sigma_\zeta^2 \end{pmatrix} \qquad \text{(A.8)}$$

Indien $\sigma_\lambda^2 = \sigma_\zeta^2 = 0$, ontstaat het globaal lineair trendmodel (deterministisch trendmodel):

$$y_t = \mu_{1,0} + \mu_{2,0} * t + \xi_t \qquad \text{(A.9)}$$

Vergelijking (A.9) volgt eenvoudig uit de matrix-eigenschap

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}^t = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}$$

Doubly Differenced trendmodel

Het 'Doubly Differenced' (DD) trendmodel ontstaat door het dubbel discreet differentiëren van de trend μ_t, ofwel $\Delta^2\mu_t = \mu_t - 2\mu_{t-1} + \mu_{t-2} = \lambda_t$, met λ_t een ruisproces. Voor een beschrijving van het DD-trendmodel zie Kitagawa (1981), en Visser en Molenaar (1990). Het model is als volgt gedefinieerd:

Meetvergelijking:

$$y_t = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} \mu_{1,t} \\ \mu_{2,t} \end{pmatrix} + \xi_t \qquad \Leftrightarrow \qquad y_t = z_t' \alpha_t + \xi_t$$

$$\xi_t \approx NID(0, h_t \sigma^2) \tag{A.10}$$

Systeemvergelijking:

$$\begin{pmatrix} \mu_{1,t} \\ \mu_{2,t} \end{pmatrix} = \begin{pmatrix} 2 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \mu_{1,t-1} \\ \mu_{2,t-1} \end{pmatrix} + \begin{pmatrix} \lambda_t \\ 0 \end{pmatrix} \qquad \Leftrightarrow \qquad \alpha_t = T\,\alpha_{t-1} + \eta_t$$

$$\eta_t \approx NID(0, \sigma^2 Q) \qquad Q = \begin{pmatrix} \sigma_\lambda^2 & 0 \\ 0 & 0 \end{pmatrix} \tag{A.11}$$

Indien $\sigma_\lambda^2 = 0$, dan ontstaat het globaal lineair trendmodel:

$$\begin{aligned} y_t = \; & \mu_{1,0} * (t+1) + \mu_{2,0} * (-t) + \xi_t = \\ & \mu_{1,0} + (\mu_{1,0} - \mu_{2,0}) * t + \xi_t \end{aligned} \tag{A.12}$$

Vergelijking (A.12) volgt eenvoudig uit de matrix-eigenschap

$$\begin{pmatrix} 2 & -1 \\ 1 & 0 \end{pmatrix}^t = \begin{pmatrix} t+1 & -t \\ t & -t+1 \end{pmatrix}$$

Stochastic Level niveaumodel

Het Stochastic Level (SL) niveaumodel wordt in de literatuur ook wel aangeduid met 'steady model', 'local level model' of 'random walk plus noise model' (Harvey 1989, §2.1.1). Voor een uitvoerige beschrijving zie Durbin en Koopman (2001, hoofdstuk 2).

Het stochastische niveau is als volgt gedefinieerd:

Meetvergelijking:

$$y_t = \mu_t + \xi_t \quad \Leftrightarrow \quad y_t = z_t' \alpha_t + \xi_t$$
$$\xi_t \approx NID(0, h_t \sigma^2) \qquad \text{(A.13)}$$

Systeemvergelijking:

$$\mu_t = \mu_{t-1} + \lambda_t \quad \Leftrightarrow \quad \alpha_t = T \alpha_{t-1} + \eta_t$$
$$\eta_t \approx NID(0, \sigma^2 Q) \qquad Q = \sigma_\lambda^2 \qquad \text{(A.14)}$$

De trend wordt dus eenvoudig gemodelleerd met een random walk. Als $\sigma_\lambda^2 = 0$, dan ontstaat een constant niveau:

$$y_t = \mu_0 + \xi_t \qquad \text{(A.15)}$$

Welk trendmodel moet wanneer gekozen worden?

Een keuze tussen bovenbeschreven drie trendmodellen kan gemaakt worden door gebruik te maen van de autocorrelatiefunctie (ACF) van elk model na stationair maken door discreet differentiëren (y_t, Δy_t, en $\Delta^2 y_t$).

Als uit de ACF op y_t blijkt dat de reeks y_t stationair is (autocorrelaties R_1, R_2, R_3, ... doven snel uit), dan is geen trendmodel nodig. Als de ACF op de reeks Δy_t zeer snel stationair is, met $-0.5 \leq R_1 \leq 0.0$, en $R_2 \cong R_3 \cong .... \cong 0.0$, dan kiezen we voor het SL-model.

Als de ACF van Δy_t langzaam uitdooft, dan kijken we naar de reeks $\Delta^2 y_t$. Als voor de ACF op $\Delta^2 y_t$ geldt (i) $R_1/R_2 \cong -4.0$, (ii) $-0.67 \leq R_1 \leq 0.0$, en $0.0 \leq R_2 \leq 0.17$, dan kiezen we het DD-model. Indien er alleen geldt $-0.67 \leq R_1 \leq 0.0$ en $0.0 \leq R_2 \leq 0.17$, dan kiezen we het LL-model. De voorwaarden voor het LL-model zijn minder stringent omdat dit trendmodel *twee* onbekende ruisvarianties heeft, en het DD-model maar *één*.

A.3.2 Periodiciteit

Cyclische invloeden en seizoensinvloeden (periodiciteit) met een vaste periodelengte van s tijdstappen worden gemodelleerd door s cycluscomponenten (γ_i) in het model op te nemen:

$$\sum_{i=0}^{s-1} \gamma_{t-i} = \omega_t \qquad \omega_t \approx NID(0, \sigma^2 \sigma_\omega{}^2) \tag{A.16}$$

Uit (A.16) volgt dat de verwachtingswaarde van de sommatie gelijk aan nul is. Als dit model in toestandsruimte vorm wordt gebracht, dan kan de vorm van de cyclische component door de tijd heen geleidelijk veranderen. Stel dat $\gamma_{i,t} \equiv \gamma_{t-i+1}$, dan wordt de meetvergelijking voor de cyclische component:

$$y_t = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} \gamma_{1,t} \\ \gamma_{2,t} \\ \gamma_{3,t} \\ \\ \\ \\ \gamma_{s-1,t} \end{pmatrix} + \xi_t \quad \Leftrightarrow \quad y_t = z_t' \alpha_t + \xi_t$$

$$\xi_t \approx NID(0, h_t \sigma^2) \tag{A.17}$$

Systeemvergelijking:

(A.18)

$$\begin{pmatrix} \gamma_{1,t} \\ \gamma_{2,t} \\ \gamma_{3,t} \\ . \\ . \\ . \\ \gamma_{s-1,t} \end{pmatrix} = \begin{pmatrix} -1 & -1 & -1 & . & . & . & -1 \\ 1 & 0 & 0 & . & . & . & 0 \\ 0 & 1 & 0 & . & . & . & 0 \\ 0 & 0 & 1 & . & . & . & 0 \\ . & . & . & . & . & . & . \\ 0 & 0 & . & . & 1 & 0 & 0 \\ 0 & 0 & . & . & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \gamma_{1,t-1} \\ \gamma_{2,t-1} \\ \gamma_{3,t-1} \\ . \\ . \\ \gamma_{s-2,t-1} \\ \gamma_{s-1,t-1} \end{pmatrix} + \begin{pmatrix} \omega_t \\ 0 \\ 0 \\ . \\ . \\ 0 \\ 0 \end{pmatrix}$$

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

$$\eta_t \approx NID(0, \sigma^2 Q) \qquad Q = \begin{pmatrix} \sigma_\omega^2 & 0 & 0 & . & . & . & 0 \\ 0 & 0 & 0 & . & . & . & 0 \\ 0 & 0 & 0 & . & . & . & 0 \\ . & . & . & . & . & . & . \\ 0 & 0 & 0 & . & . & . & 0 \end{pmatrix}$$

Als $\sigma_\omega^2 = 0$, dan ontstaat een model met een deterministische cyclus.

A.3.3 Het basisstructureel model

Een algemeen toegepast model is het zogenaamde basisstructureel model (BSM) (Harvey 1989, H.4.1):

$$y_t = \mu_t + \gamma_t + \xi_t \qquad \text{(A.19)}$$

μ_t: trend component
γ_t: cyclische component
ξ_t: ruis component

Het model is structureel in die zin dat elke component (trend, cyclus en ruis) direct interpreteerbaar is. Het BSM-model kan worden opgevat als de som van drie ARIMA-modellen welke niet direct observeerbaar zijn (Harvey, 1994, pag. 255). Deze klasse van modellen worden daarom Unobserved Components Autoregressive Integrated Moving Average (UCARIMA) modellen genoemd. Als voorbeeld wordt een basisstructureel model gegeven dat is opgebouwd uit een ST-model en een seizoenseffect met een lengte van 4 tijdstappen.

Meetvergelijking:

$$y_t = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mu_{1,t} \\ \mu_{2,t} \\ \gamma_{1,t} \\ \gamma_{2,t} \\ \gamma_{3,t} \end{pmatrix} + \xi_t \qquad \Leftrightarrow \qquad y_t = z_t' \alpha_t + \xi_t$$

$$\xi_t \approx NID(0, h_t \sigma^2) \tag{A.20}$$

Systeemvergelijking:

$$\begin{pmatrix} \mu_{1,t} \\ \mu_{2,t} \\ \gamma_{1,t} \\ \gamma_{2,t} \\ \gamma_{3,t} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \mu_{1,t-1} \\ \mu_{2,t-1} \\ \gamma_{1,t-1} \\ \gamma_{2,t-1} \\ \gamma_{3,t-1} \end{pmatrix} + \begin{pmatrix} \lambda_t \\ \zeta_t \\ \omega_t \\ 0 \\ 0 \end{pmatrix} \tag{A.21}$$

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

$$\eta_t \approx NID(0, \sigma^2 Q) \qquad Q = \begin{pmatrix} \sigma_\lambda^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma_\zeta^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_\omega^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

A.3.4 Multiple Regressiemodel

Het Multiple Lineaire Regressiemodel luidt:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \ldots + \beta_m x_{m,t} + \xi_t \qquad \text{(A.22)}$$

β_0 : constante (intercept)
β_i : regressie-coëfficiënt i
$x_{i,t}$: verklarende variabele i op tijdstip t

Door het model in toestandsruimte-vorm te formuleren, wordt het mogelijk om de regressie-coëfficiënten van het regressiemodel tijdsafhankelijk te maken (Harvey, 1989, Hoofdstuk 4).

Meetvergelijking:

$$y_t = \begin{pmatrix} 1 & x_{1,t} & x_{2,t} & . & . & . & x_{m,t} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_{1,t} \\ \beta_{2,t} \\ . \\ . \\ . \\ \beta_{m,t} \end{pmatrix} + \xi_t \quad \Leftrightarrow \quad y_t = z_t' \alpha_t + \xi_t \qquad \text{(A.23)}$$

$$\xi_t \approx NID(0, h_t \sigma^2)$$

Indien elke regressie-coëfficiënt met een random walk proces wordt gemodelleerd, ziet de systeemvergelijking er als volgt uit:

$$\begin{pmatrix} \beta_{0,t} \\ \beta_{1,t} \\ \beta_{2,t} \\ . \\ . \\ . \\ \beta_{m,t} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & . & . & . & 0 \\ 0 & 1 & 0 & . & . & . & 0 \\ 0 & 0 & 1 & . & . & . & 0 \\ . & . & . & . & . & . & . \\ 0 & 0 & 0 & . & . & . & 1 \end{pmatrix} \begin{pmatrix} \beta_{0,t-1} \\ \beta_{1,t-1} \\ \beta_{2,t-1} \\ . \\ . \\ . \\ \beta_{m,t-1} \end{pmatrix} + \begin{pmatrix} 0 \\ \varsigma_{1,t} \\ \varsigma_{2,t} \\ . \\ . \\ . \\ \varsigma_{m,t} \end{pmatrix}$$

$$\Leftrightarrow \quad \alpha_t = T\,\alpha_{t-1} + \eta_t$$

$$\eta_t \approx NID(0, \sigma^2 Q) \qquad Q = \begin{pmatrix} 0 & 0 & 0 & . & . & . & 0 \\ 0 & \sigma_{\varsigma,1}^{\ 2} & 0 & . & . & . & 0 \\ 0 & 0 & \sigma_{\varsigma,2}^{\ 2} & . & . & . & 0 \\ . & . & . & . & . & . & . \\ 0 & 0 & 0 & . & . & . & \sigma_{\varsigma,m}^{\ 2} \end{pmatrix} \tag{A.24}$$

Indien Q = 0, dan komt het schatten van (A.23) en (A.24) overeen met het toepassen van het gewone Multiple Lineaire Regressiemodel (A.22).

A.3.5 Uitbreiding basisstructureel model met regressie

Het basisstructureel model kan nu verder worden uitgebreid door verklarende variabelen in het model op te nemen:

$$y_t = \mu_t + \gamma_t + \sum_{i=1}^{m} \beta_{i,t}\, x_{i,t} + \xi_t \tag{A.25}$$

De parameters voor trend, cyclus en verklarende variabelen kunnen tijdafhankelijk worden gemaakt door (A.25) als een toestandsruimte-model te schrijven. Een voorbeeld is een structureel model opgebouwd uit een LL-trendmodel, een cyclus van 4 tijdstappen en twee verklarende variabelen.

Meetvergelijking:

$$y_t = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & x_{1,t} & x_{2,t} \end{pmatrix} \begin{pmatrix} \mu_{1,t} \\ \mu_{2,t} \\ \gamma_{1,t} \\ \gamma_{2,t} \\ \gamma_{3,t} \\ \beta_{1,t} \\ \beta_{2,t} \end{pmatrix} + \xi_t$$

$$\Leftrightarrow \quad y_t = z_t' \alpha_t + \xi_t \tag{A.26}$$

$$\xi_t \approx NID(0, h_t \sigma^2)$$

Systeemvergelijking:

$$\begin{pmatrix} \mu_{1,t} \\ \mu_{2,t} \\ \gamma_{1,t} \\ \gamma_{2,t} \\ \gamma_{3,t} \\ \beta_{1,t} \\ \beta_{2,t} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_{1,t-1} \\ \mu_{2,t-1} \\ \gamma_{1,t-1} \\ \gamma_{2,t-1} \\ \gamma_{3,t-1} \\ \beta_{1,t-1} \\ \beta_{2,t-1} \end{pmatrix} + \begin{pmatrix} \lambda_t \\ \zeta_t \\ \omega_t \\ 0 \\ 0 \\ \varsigma_{1,t} \\ \varsigma_{2,t} \end{pmatrix}$$

$$\Leftrightarrow \quad \alpha_t = T\,\alpha_{t-1} + \eta_t$$

$$\eta_t \approx NID(0, \sigma^2 Q) \qquad Q = \begin{pmatrix} \sigma_\lambda^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_\zeta^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_\omega^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_{\varsigma,1}^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma_{\varsigma,2}^2 \end{pmatrix} \qquad \text{(A.27)}$$

Indien Q = 0, dan ontstaat een deterministisch model met constante parameters. Het deterministisch model dat ook kan worden geanalyseerd met behulp van het gewone Multiple Lineaire Regressiemodel, is dus te beschouwen als één bijzondere situatie van het algemene structurele model met tijdsafhankelijke parameters.

A.4 Kalmanfilter

In §A.2 is het toestandsruimte-model via vergelijkingen (A.4) en (A.5) beschreven. Het toestandsruimte-model wordt geanalyseerd met behulp van het discreet Kalmanfilter. Het Kalmanfilter levert in een bepaalde zin optimale schattingen voor de (niet direct meetbare) toestandsvector α_t. In (A.6.a) tot en met (A.6.j) zijn de benodigde modelveronderstellingen gegeven. Indien wordt uitgegaan van het Gaussisch model (A.6.h), (A.6.i) en (A.6.j), dan levert het Kalmanfilter de Minimum Mean Square Estimator (MMSE) voor α_t. Indien de veronderstellingen (A.6.h), (A.6.i) en (A.6.j) niet gelden, dan levert het Kalmanfilter de Minimum Mean Square Lineair Estimator (MMSLE) voor α_t. Voor details zie Harvey (1989, Hoofdstuk 3).

We definiëren nu de volgende grootheden. Vector a_t is de MMSE voor α_t bij het Gaussisch model en de MMSLE bij het niet-Gaussisch model, gebaseerd op de observaties: $Y_t = (y_t, y_{t-1}, y_{t-2}, ..., y_3, y_2, y_1)'$. De matrix $\sigma^2 P_t$ is de Mean Square Error (MSE) matrix van a_t (dus de variantie-covariantie-matrix van de voorspelfout $a_t - \alpha_t$). Vector $a_{t|t'}$ is de MMSE voor α_t bij het Gaussisch model en de MMSLE bij het niet-Gaussisch model, gebaseerd op alle observaties tot en met tijdstip t': $Y_{t'} = (y_{t'}, y_{t'-1}, ..., y_2, y_1)'$. Tenslotte is $\sigma^2 P_{t|t'}$ de MSE-matrix van $a_{t|t'}$. Er geldt dus $a_t = a_{t|t}$ en $P_t = P_{t|t}$.

De variantie van $a^i_{t|t'}$ (i-de element van $a_{t|t'}$) is $\sigma^2 P^{i,i}_{t|t'}$ ($P^{i,i}_{t|t'}$ is het i-de diagonaal element van $P_{t|t'}$).

Het discreet Kalmanfilter levert een recursief schema voor de MMSE (bij een Gaussisch model) of een MMSLE (bij een niet-Gaussisch model) $a_{t|t}'$ voor α_t en een Mean Square Estimator matrix $\sigma^2 P_{t|t'}$ voor $a_{t|t'}$.

Er kunnen drie situaties worden onderscheiden:
- $t > t'$: voorspellen;
- $t = t'$: filteren of signaalextractie;
- $t < t'$: smoothen.

In alle drie de situaties levert het Kalmanfilter optimale schattingen voor de toestandsvector α_t.

Het Kalmanfilter veronderstelt dat h_t, Q en T bekend zijn. Indien dit niet het geval is, kunnen deze variabelen met een *maximum-likelihood-schatter* worden geschat (Harvey, 1989, §4.2.2).

A.4.1 Filteren

Het discreet Kalmanfilter bestaat uit voorspellingsvergelijkingen en filtervergelijkingen.

Gegeven a_{t-1} kan een voorspelling van α_t worden gemaakt: $a_{t|t-1}$. De voorspellings-vergelijkingen leveren de optimale schatting voor α_t op basis van Y_{t-1}:

$$a_{t/t-1} = T\, a_{t-1}$$

$$P_{t/t-1} = T\, P_{t-1} T' + Q \qquad \text{(A.28)}$$

Een optimale voorspelling voor y_t op basis van Y_{t-1} wordt verkregen door:

$$y_{t/t-1} = z_t' a_{t/t-1} \tag{A.29}$$

Op het moment dat y_t beschikbaar komt, kan de één-staps-voorspelfout worden berekend:

$$v_t = y_t - z_t' a_{t/t-1} \tag{A.30}$$

Dit wordt ook wel de innovatie genoemd en kan worden opgevat als de nieuwe informatie die door observatie y_t wordt toegevoegd aan de bestaande informatie aanwezig in Y_{t-1}. Er kan worden bewezen dat de innovaties onderling onafhankelijk en normaal verdeeld zijn, met een variantie die voldoet aan:

$$\begin{aligned} v_t &\approx NID(0, \sigma^2 f_t) \\ f_t &= z_t' P_{t/t-1} z_t + h_t \end{aligned} \tag{A.31}$$

Gestandaardiseerde innovaties v_t' worden gegeven door $v_t' = v_t / \sqrt{f_t}$. Deze innovaties zijn normaalverdeeld met gemiddelde nul en standaardeviatie 1.0.

Als y_t bekend is, dan kunnen de schattingen van (A.28) worden geactualiseerd met de filtervergelijkingen:

$$a_t = a_{t/t-1} + \frac{P_{t/t-1} z_t v_t}{f_t} = a_{t/t-1} + K_t v_t$$

$$P_t = P_{t/t-1} - \frac{P_{t/t-1} z_t z_t' P_{t/t-1}}{f_t} = P_{t/t-1} + K_t z_t' P_{t/t-1} \tag{A.32}$$

$$K_t = \frac{P_{t/t-1} z_t}{f_t}$$

K_t wordt de 'gain' genoemd.

Het recursieve schema wordt opgestart met $a_0 = 0$ en $P_0 = \kappa I$. I is de identiteitsmatrix en we nemen voor κ een groot getal. Hiermee wordt aangegeven dat a_0 een slechte schatting is voor α_0. Het filter heeft daarom een zekere inschakeltijd nodig. Gedurende deze inschakelperiode genereert het filter zelf betrouwbare schattingen voor α_t voor de rest van het filterproces. Deze inregeltijd is groter of gelijk aan het aantal elementen in de toestandsvector (Harvey, 1989, H. 3).

Bij onbetrouwbare of ontbrekende observaties kan voor de variabele h een groot getal worden genomen. Hierdoor krijgt de variantie f_t van de innovatie een grote waarde. Daarmee wordt aan het filter doorgegeven dat de betreffende observatie slecht is. De aanpassing van $a_{t|t-1}$ tijdens de actualisatie in (A.32) is in dat geval minimaal waardoor $a_t \approx a_{t|t-1}$. Dit komt neer op een soort interpolatie. De innovatie op tijdstip t wordt uitgesloten door de variantie van de innovatie een hele grote waarde te geven.

A.4.2 Smoothen

In de vorige paragraaf is aangegeven dat het Kalmanfilter de MMSE (bij het Gaussisch model) of de MMSLE (bij het niet-Gaussisch model) levert van α_t , gegeven alle informatie beschikbaar op tijdstip t (vector Y_t). Op tijdstip t wordt een schatting gemaakt van α_{t+1}. Dit levert $a_{t+1|t}$. Op het moment dat observatie y_{t+1} beschikbaar komt, wordt $a_{t+1|t}$ geactualiseerd. Dit levert a_{t+1}. Vervolgens wordt de aandacht geconcentreerd op het schatten van de toestandsvector één tijdstap verder.

Als alle N observaties (vector Y_N) bekend zijn, kunnen veel betere schattingen van de toestandsvectoren worden verkregen door de informatie van alle N observaties te gebruiken. Dit gebeurt door middel van 'smoothen' (gladstrijken). Met behulp van smoothen worden schattingen verkregen voor α_t en y_t door ook gebruik te maken van alle observaties verkregen na tijdstip t. Het doel van smoothen is om betere schattingen van de toestandsvectoren α_1 tot en met α_N te krijgen. De gesmoothde schattingen voor de toestandsvector en de observaties zijn daarom ongevoelig voor de inregeltijd.

Indien Q = 0 en T = I, dan is $a_{t|N}$ en $P_{t|N}$ onafhankelijk van de tijd en komen overeen met Ordinary Least Squares (OLS) schatters van het Multiple Lineaire Regressiemodel. OLS-regressie is dus een speciaal geval van het Kalmanfilter.

Er zijn drie basis algoritmen voor smoothen bekend:
- De 'fixed point smoother'. Deze smoother berekent gesmoothte schattingen van één toestandsvector op een vast tijdstip τ. Dus $a_{\tau|t}$ voor één vast τ op alle tijdstippen $t > \tau$. Het algoritme van de 'fixed point smoother' werkt via een voorwaartse recursie;
- De 'fixed lag smoother'. Deze smoother berekent gesmoothde schattingen voor de toestandsvector met een vaste vertraagde periode. Dus $a_{t-j|t}$ voor j = 1, ..., m (m is de maximale vertragingsperiode). Het algoritme van de 'fixed lag smoother' werkt via een voorwaartse recursie.
- De 'fixed interval smoother. Deze smoother berekent gesmoothde schattingen voor alle toestandsvectoren, gebruik makend van alle informatie uit de vector met observaties Y_N. Het algoritme van de 'fixed interval smoother' werkt via een terugwaartse recursie, die start op tijdstip t = N en eindigt op tijdstip t = 0.

Hier wordt verder alleen ingegaan op de *fixed interval smoother*. Voor een afleiding en een beschrijving van de vergelijkingen van de andere twee smoothers wordt verwezen naar Anderson and Moore (1979, H.7).

De vergelijkingen van de *fixed interval smoother* worden met een achterwaartse recursie berekend, startend met a_N en P_N uit de laatste iteratie van (A.32). De vergelijkingen luiden:

$$a_{t/N} = a_t + P_t^*(a_{t+1/N} - T a_{t/t})$$

$$P_{t/N} = P_t + P_t^*(P_{t+1/N} - P_{t+1/t}) P_t^{*'} \qquad \text{(A.33)}$$

$$P_t^* = P_t T' P_{t+1/t}^{-1}$$

De vergelijkingen van de *fixed interval smoother* worden gestart met de laatste Kalmanfilter schattingen: a_N en P_N. De recursie werkt terug tot tijdstip t = 0.

Een gesmoothde schatting voor y_t wordt verkregen met:

$$y_{t/N} = z_t' a_{t/N} \qquad \text{(A.34)}$$

De gesmoothde innovaties volgen uit:

$$v_{t/N} = y_t - z_t' a_{t/N} \qquad \text{(A.35)}$$

De gesmoothde innovaties zijn nog wel normaal verdeeld, maar zijn niet meer onderling onafhankelijk. Er geldt:

$$v_{t/N} \approx N(0, \sigma^2 f_{t/N})$$

$$f_{t/N} = z_t' P_{t/N} z_t + h_t \qquad \text{(A.36)}$$

De Jong en Mackinnon (1988) geven de volgende vergelijking voor de covariantiematrix $\sigma^2 P_{s,t/N}$ tussen de toestandsvectoren $a_{s/N}$ en $a_{t/N}$:

$$P_{s,t/N} = P_s^* P_{s+1,t/N} \qquad \text{(A.37)}$$

Als randvoorwaarde geldt dat $P_{t,t/N} = P_{t/N}$, voor alle $1 \le t \le N$. De matrix P_s^* is gegeven in (A.33). In de hoofdstukken 6 en 7 wordt gebruik gemaakt van een speciaal geval van de matrix $P_{s,t/N}$, namelijk $P_{t,N/N}$. Zie verder ook §A.6 in Visser (1994).

A.4.3 Voorspellen

Nadat alle N observaties met het Kalmanfilter zijn verwerkt, kunnen voorspellingen van de toestandsvector α_{N+l} na tijdstip N op basis van de observaties Y_N worden verkregen door:

$$a_{N+l/N} = T\, a_{N+l-1/N} = T^l a_N$$
$$P_{N+l/N} = T\, P_{N+l-1/N} T' + Q \qquad \text{(A.38)}$$

Voorspellingen van de waarneming y_{N+l} op basis van de observaties Y_N worden verkregen door:

$$y_{N+l/N} = z'_{N+l}\, a_{N+l/N} = z'_{N+l} T^l a_N \qquad \text{(A.39)}$$

A.4.4 Maximum likelihood

Onbekende variabelen uit het toestandsruimte-model kunnen met een maximum-likelihood-procedure worden geschat. In TrendSetter wordt ervan uitgegaan dat alleen σ^2 en de diagonaal elementen van Q_t onbekend zijn. De matrix T en de vector a_t worden als geheel bekend verondersteld.

Via de voorspelfout-decompositie wordt een log-likelihood-functie (log L) afgeleid (Harvey 1989, Hoofdstuk 3):

$$\log(L) = -\frac{1}{2}\left[(N-d)\log(2\pi\sigma^2) + \sum_{t=d+1}^{N}\left(\log(f_t) + \frac{v_t^2}{\sigma^2 f_t}\right)\right] \qquad \text{(A.40)}$$

waarbij de eerste d observaties buiten beschouwing worden gelaten (d is de inregeltijd van het filter). Uit (A.40) kan een schatting voor σ^2 worden afgeleid door (A.40) te differentiëren naar σ^2 en gelijk te stellen aan nul:

$$\hat{\sigma}^2 = \frac{1}{N-d}\sum_{t=d+1}^{N}\frac{v_t^2}{f_t} \qquad \text{(A.41)}$$

Invullen van (A.41) in (A.40) levert na verwijderen van constante termen en vermenigvuldigen met -2 de zogenaamde geconcentreerde log-likelihood-funktie (log L_c):

$$\log(L_c) = \sum_{t=d+1}^{N}\log(f_t) + (N-d)\log(\hat{\sigma}^2) = \sum_{t=d+1}^{N}\log(\hat{\sigma}^2 f_t) \qquad \text{(A.42)}$$

Maximaliseren van (A.40) naar de onbekende ruisvarianties komt neer op het minimaliseren van (A.42). De maximum- likelihood-schatters van de onbekende parameters, dus de diagonaal elementen van Q, worden bepaald door (A.42) te minimaliseren als functie van deze onbekende parameters met behulp van een numerieke procedure. Uit (A.42) volgt dat de *maximum likelihood* neerkomt op het minimaliseren van de som van de log-varianties van alle innovaties na het inregelen van het filter.

Structurele Tijdreeksmodellen zijn een krachtig instrument voor het analyseren van metingen in de tijd. Componentsgewijs worden trend, cyclus en de invloed van verklarende variabelen geschat. Bovendien is te allen tijde onzekerheidsinformatie beschikbaar. Wel is de theorie zoals in deze appendix beschreven, conceptueel lastig. *Foto: H. Visser.*

Appendix B Inputtabellen TrendSpotter

Tabel B.1 *Datafile met als kolommen een oplopende index 1 tot en met 120, de tijd t (oplopend jaartal plus maandnummer in tienden uitdrukt), de pseudoreeksen Pseudo1 en Pseudo2 en een verklarende variabele x_t.*

1	1991.083	32.023140	31.012030	0.202221970
2	1991.167	43.154073	40.317136	0.567387410
3	1991.250	35.548286	36.632349	-0.216812556
4	1991.333	41.016933	39.546713	0.294043939
5	1991.417	13.905517	20.821592	-1.383215003
6	1991.500	32.658146	29.212907	0.689047778
7	1991.583	31.044355	29.695026	0.269865871
8	1991.667	26.404218	30.738898	-0.866936044
9	1991.750	19.998773	29.790383	-1.958321899
10	1991.833	32.631018	31.201978	0.285808034
11	1991.917	19.249889	31.937647	-2.537551704
12	1992.000	29.116139	30.629237	-0.302619639
13	1992.083	33.482508	31.597916	0.376918505
14	1992.167	38.512464	31.337853	1.434922109
15	1992.250	34.356747	30.376407	0.796067991
16	1992.333	45.453801	39.240448	1.242670740
17	1992.417	28.176602	33.169227	-0.998525160
18	1992.500	10.734899	19.253999	-1.703819939
19	1992.583	14.070388	17.982033	-0.782329015
20	1992.667	11.756126	15.543714	-0.757517576
21	1992.750	19.137152	28.796554	-1.931880411
22	1992.833	31.819779	29.362679	0.491419915
23	1992.917	33.608465	32.040790	0.313535039
24	1993.000	26.421658	30.082810	-0.732230343
25	1993.083	17.474389	20.558016	-0.616725399
26	1993.167	24.299572	29.098873	-0.959860207
27	1993.250	33.185280	33.599740	-0.082892006
28	1993.333	21.268216	18.680757	0.517491962
29	1993.417	19.866244	23.803361	-0.787423402
30	1993.500	27.049184	25.973464	0.215143974
31	1993.583	17.872727	10.235329	1.527479623
32	1993.667	11.001411	18.304469	-1.460611652
33	1993.750	27.561920	24.375639	0.637256054
34	1993.833	24.359072	20.301993	0.811415726
35	1993.917	28.888058	31.393766	-0.501141620
36	1994.000	20.037483	20.088706	-0.010244651
37	1994.083	26.097779	22.716722	0.676211474
38	1994.167	22.676811	21.199775	0.295407219
39	1994.250	39.552622	37.800627	0.350398976
40	1994.333	34.959122	28.718754	1.248073534
41	1994.417	16.362128	20.433644	-0.814303093
42	1994.500	24.468380	21.401999	0.613276042
43	1994.583	14.829861	9.986607	0.968650746
44	1994.667	22.981093	24.997543	-0.403289898
45	1994.750	22.064110	22.104138	-0.008005677
46	1994.833	18.650124	18.281193	0.073786290
47	1994.917	-1.0	-1.0	-1.391393433
48	1995.000	-1.0	-1.0	-0.198971029
49	1995.083	-1.0	-1.0	-0.438466076
50	1995.167	-1.0	-1.0	0.208007590
51	1995.250	38.917586	34.243845	0.934748266
52	1995.333	26.582246	27.453464	-0.174243715
53	1995.417	29.050158	20.972963	1.615439022
54	1995.500	14.823142	22.512200	-1.537811605
55	1995.583	14.799883	14.329483	0.094079978
56	1995.667	12.851098	7.344304	1.101358923
57	1995.750	17.694819	22.038085	-0.868653181
58	1995.833	19.186560	12.211480	1.395015987

59	1995.917	19.132052	25.332823	-1.240154362
60	1996.000	17.329419	10.986227	1.268638342
61	1996.083	20.354470	20.354470	0.726007456
62	1996.167	27.474228	27.474228	0.962199208
63	1996.250	20.081033	20.081033	-1.320494375
64	1996.333	29.362790	29.362790	0.441335266
65	1996.417	18.450713	18.450713	0.656130739
66	1996.500	12.908885	12.908885	1.371664771
67	1996.583	25.458920	25.458920	0.887317851
68	1996.667	14.116466	14.116466	0.314319634
69	1996.750	14.727127	14.727127	-0.576294296
70	1996.833	16.525660	16.525660	-0.774397629
71	1996.917	31.483770	31.483770	1.293484979
72	1997.000	16.316879	16.316879	0.359989047
73	1997.083	19.787994	19.787994	-0.179283316
74	1997.167	29.119771	29.119771	0.137641644
75	1997.250	31.138773	31.138773	0.785544343
76	1997.333	30.578973	30.578973	1.251615130
77	1997.417	16.906850	16.906850	-0.249940829
78	1997.500	12.170318	12.170318	-0.907311340
79	1997.583	8.770194	8.770194	0.936136000
80	1997.667	9.102049	9.102049	-0.715328366
81	1997.750	19.683835	19.683835	-0.259476227
82	1997.833	21.580570	21.580570	-0.744838474
83	1997.917	34.030492	34.030492	-0.414262990
84	1998.000	25.677700	25.677700	1.613806044
85	1998.083	15.702078	15.702078	-0.932480465
86	1998.167	28.806320	28.806320	-0.248727825
87	1998.250	32.772269	32.772269	0.911752363
88	1998.333	19.814740	19.814740	-0.584929345
89	1998.417	23.427025	23.427025	0.761197576
90	1998.500	8.810571	8.810571	-1.154352891
91	1998.583	3.871020	3.871020	0.380736357
92	1998.667	15.336705	15.336705	0.162152433
93	1998.750	23.532753	23.532753	3.019011836
94	1998.833	26.251546	26.251546	-0.024754526
95	1998.917	12.523191	12.523191	-2.806452516
96	1999.000	17.947643	17.947643	0.100051153
97	1999.083	27.797704	27.797704	-0.145905813
98	1999.167	23.101973	23.101973	1.189623941
99	1999.250	23.028875	23.028875	-0.539935092
100	1999.333	14.449453	14.449453	-1.638582899
101	1999.417	23.647351	23.647351	-0.951228305
102	1999.500	9.396191	9.396191	-1.115202135
103	1999.583	19.401072	19.401072	-0.251515321
104	1999.667	29.283666	29.283666	0.152124688
105	1999.750	37.514327	37.514327	1.498987629
106	1999.833	23.721767	23.721767	0.542306615
107	1999.917	25.281136	25.281136	-0.744850420
108	2000.000	26.812408	26.812408	-1.321530443
109	2000.083	21.816908	21.816908	0.167492231
110	2000.167	29.697765	29.697765	0.296791936
111	2000.250	15.294313	15.294313	-1.633119557
112	2000.333	25.815784	25.815784	1.081516093
113	2000.417	16.525926	16.525926	-1.630825294
114	2000.500	23.194981	23.194981	0.293160227
115	2000.583	8.338512	8.338512	-1.137868007
116	2000.667	16.261475	16.261475	-0.780288879
117	2000.750	20.916978	20.916978	-0.133234871
118	2000.833	24.130480	24.130480	1.358982204
119	2000.917	15.732720	15.732720	-2.078667933
120	2001.000	29.171143	29.171143	0.276086218

Tabel B.2 Opbouw optiefile 'optie.km' voor het schatten van een DD-trend. Het aantal metingen is 120.

Stap 1	DD trendschatting
Stap 2	2 0 0
Stap 3	0.0
Stap 4	20
Stap 5	2
Stap 6	0
Stap 7	1
Stap 8	0 0
Stap 9	1 -1.0
Stap 10	0
Stap 12	0
Stap 14	120
Stap 15	0 0 0
Stap 16	'(3X,F9.3,F10.6)'
Stap 17	1 2

Tabel B.3 Opbouw optiefile 'optie.km' voor het schatten van een DD-trend in combinatie met een jaarcyclus. Het aantal metingen is 120.

Stap 1	DD trend plus jaarcyclus
Stap 2	2 12 0
Stap 3	0.0 0.0
Stap 4	20
Stap 5	3
Stap 6	0
Stap 7	1
Stap 8	0 0
Stap 9	1 -1.0
Stap 10	0
Stap 12	0
Stap 14	120
Stap 15	0 0 0
Stap 16	'(3x,F9.3,F10.6)'
Stap 17	1 2

Tabel B.4 *Opbouw optiefile 'optie.km' voor het schatten van een DD-trend plus de invloed van een verklarende variabele met een constante weegfactor. De schatting is voor de reeks Pseudo1. Het aantal metingen is 120.*

Stap 1	DD trend plus invloed xt
Stap 2	2 0 1
Stap 3	0.0 0.0
Stap 4	20
Stap 5	3
Stap 6	0
Stap 7	1
Stap 8	0 0
Stap 9	1 -1.0
Stap 10	0
Stap 12	0
Stap 13	Invloed verklarende variabele
Stap 14	120
Stap 13	0 0 0
Stap 15	'(3X,F9.3,F10.6,10X,F13.9)'
Stap 17	1 2 3

Tabel B.5 *Opbouw optiefile 'optie.km' voor het schatten van een DD-trend plus de invloed van een verklarende variabele met een constante weegfactor plus een jaarcyclus. De schatting is voor de reeks Pseudo1. Het aantal metingen is 120.*

Stap 1	Totale model
Stap 2	2 12 1
Stap 3	0.0 0.0 0.0
Stap 4	20
Stap 5	3
Stap 6	0
Stap 7	1
Stap 8	0 0
Stap 9	1 -1.0
Stap 10	0
Stap 12	0
Stap 13	Invloed verklarende variabele
Stap 14	120
Stap 13	0 0 0
Stap 15	'(3X,F9.3,F10.6,10X,F13.9)'
Stap 17	1 2 3