

Algemene toelichting op de Box-Jenkins-methode

Aan de hand van tijdreeksanalyse volgens de Box-Jenkins-methode [Box and Jenkins, 1976], kan de tijdreeks van een variabele beschreven worden met een discreet lineair dynamisch stochastisch model. In het nu volgende geven we een beknopte uitleg van deze methode. Voor een diepgaande theoretische verhandeling wordt verwezen naar het standaardwerk [Box and Jenkins, 1976]. Voor een praktisch gerichte verhandeling wordt verwezen naar [McLeod, 1983].

Uitleg aan de hand van systeemtheorie

Het principe van de Box-Jenkins benadering kan goed worden toegelicht aan de hand van begrippen uit de systeemtheorie. Een systeem zal daartoe - enigszins intuïtief - worden omschreven als een deel van de werkelijkheid, dat via in- en uitgangssignalen wisselwerkingen vertoont met z'n omgeving. De grondwaterstand in een peilbuis zullen we dan opvatten als het uitgangssignaal (hier: de uitvoerreeks) van een systeem, dat wordt aangedreven door meerdere ingangssignalen (hier: invoerreeksen of ook wel dynamische invloedsfactoren). In het geval van een (freatische) grondwaterstand zullen deze laatste voornamelijk van meteorologische aard zijn. Een dergelijk systeem kan mathematisch worden uitgeschreven als:

$$Y(t) = g(G, X, t) \quad 1$$

met Y de grondwaterstand, $g(\cdot)$ één of andere functie, G een verzameling modelparameters, X de verzameling dynamische invloedsfactoren en t de continue tijd.

Als we de grondwaterstand empirisch willen modelleren verdient het echter aanbeveling deze te beschouwen als de uitvoerreeks van een discreet lineair dynamisch stochastisch systeem:

- discreet: omdat relevante gegevens over de grondwaterstand en de dynamische invloedsfactoren meestal slechts beschikbaar zijn in de vorm van discrete tijdreeksen;
- lineair: omdat dit de minste theoretische en praktische problemen oplevert, bovendien is gebleken dat veel systemen redelijk beschreven kunnen worden met lineaire modellen;
- dynamisch: omdat de tijd een essentiële rol speelt in het gedrag van de uitvoerreeks;
- stochastisch: omdat zo'n systeem nooit exact (deterministisch) kan worden beschreven, enerzijds vanwege de benadering die is ingevoerd door de discretisatie en linearisatie, anderzijds vanwege het feit dat het onmogelijk is alle dynamische invloedsfactoren bij de beschrijving mee te nemen.

Een discreet lineair dynamisch stochastisch systeem kan mathematisch worden uitgeschreven als:

$$Y_t = f(F, X', t) + N_t \quad 2$$

met Y de grondwaterstand, $f(\cdot)$ een lineaire functie, F een verzameling modelparameters, X' een deelverzameling van de dynamische invloedsfactoren, N de ruis en t de discrete tijdsindex. Omdat de algemene vorm van de Box-Jenkins modellen ook kan worden uitgeschreven als bovenstaande formule, is de Box-Jenkins-methode bij uitstek geschikt om zo'n systeem langs empirische weg te beschrijven. En omdat de methode er op is gericht het stochastische element terug te brengen tot een verschijnsel dat minimaal is en dat bekende waarschijnlijkheidswetten volgt, kunnen er uitspraken omtrent het systeem worden gedaan met minimale en kwantificeerbare onzekerheden.

Statistisch modelleren van een tijdreeks

De klassieke statistische methodologie is zelden geschikt voor een statistische modellering van tijdreeksen, omdat deze gebaseerd is op onafhankelijke waarnemingen, terwijl de waarnemingen in een tijdreeks vaak niet onafhankelijk van elkaar zijn. De belangrijkste bijdragen van Box

en Jenkins tot de oplossing van dit probleem bestonden uit de formulering van een familie van statistische modellen om tijdreeksen te beschrijven en uit aanwijzingen tot een verantwoorde modelbouw [Box and Jenkins, 1976]. Voor de modellering van een tijdreeks ontwikkelden ze twee benaderingen, namelijk univariate modellering aan de hand van een Arima-model en transfer-ruismodellering aan de hand van een transfer-ruismodel. Beide zullen in het nu volgende kort en voornamelijk intuïtief worden omschreven.

Univariaat modelleren

Bij de univariate modellering wordt uitsluitend uitgegaan van de te modelleren tijdreeks, die wordt beschouwd als de uitvoerreeks van een discreet lineair dynamisch stochastisch systeem dat wordt aangedreven door een stochastische invoerreeks. Deze laatste is onbekend en ontstaat eigenlijk pas in de modelleerfase als het modelresidu. De modellering is er dan ook op gericht dit terug te brengen tot een verschijnsel dat bekende waarschijnlijkheidswetten volgt ("witte ruis") en een minimale variantie heeft. De variantie mag in dit verband als een soort maat voor de onzekerheid van het model worden beschouwd.

Witte ruis kan worden opgevat als een volledig toevallig signaal, bestaande uit een opeenvolging van onafhankelijke trekkingen uit een normale kansverdeling, met gemiddelde nul en variantie σ_a^2 .

Het zal duidelijk zijn dat de univariate benadering nauwelijks inzicht levert in het beschouwde systeem. Het wordt wel veel toegepast om voorspellingen van een bepaalde variabele te genereren. Deze zijn dan uitsluitend gebaseerd op een statistische analyse van het gedrag van deze variabele in het verleden. Deze toepassing heeft dan ook vooral opgang gemaakt in de economie, met zijn gecompliceerde en minder inzichtelijke dynamische systemen.

Transfer-ruismodelleren

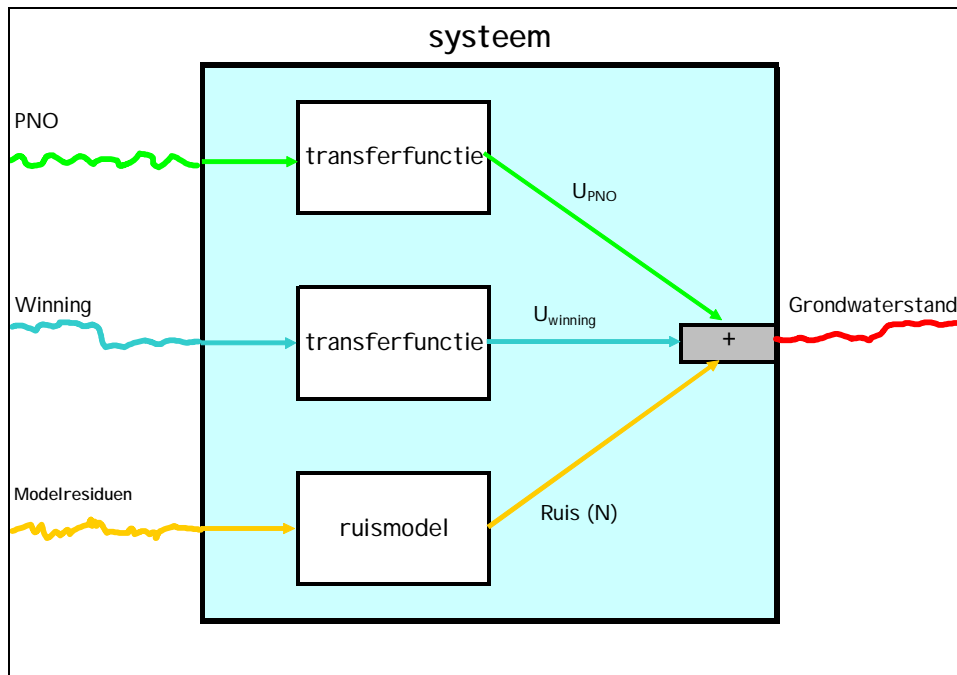
Bij de transfer-ruismodellering wordt de te modelleren tijdreeks beschouwd als de uitvoerreeks van een discreet lineair dynamisch stochastisch systeem dat wordt aangedreven door meerdere invoerreeksen, waaronder ook witte ruis. Elke invoerreeks (X_i) levert hierbij volgens een afzonderlijke lineaire transferfunctie een component (U_i) van de uitvoerreeks (Y). De resterende ruis (N) vertegenwoordigt de component die wordt geleverd door alle dynamische invloedsfactoren die niet bij de modellering zijn betrokken. Deze ruis kan op zijn beurt worden weergegeven als de uitvoerreeks van een deelsysteem, dat volgens een Arima-model (in feite ook een transferfunctie) wordt aangedreven door witte ruis. De verschillende componenten geven door middel van lineaire superpositie de uiteindelijk waargenomen uitvoerreeks, volgens:

$$Y_t = U_{1t} + \dots + U_{mt} + N_t \quad 3$$

Figuur b1.1 toont het principe van het transfer-ruismodel, voor het geval een grondwaterstand wordt gemodelleerd als functie van het potentieel neerslagoverschot en een grondwaterwinning.

De kracht van het transfer-ruismodel is dat een groot aantal dynamische situaties beschreven kan worden met slechts een gering aantal modelparameters.

Figuur b1.1: Principe van het transfer-ruismodel van de grondwaterstand, hier met het potentieel neerslagoverschot en een winning als invoerreeksen.



Ontwikkelen van een Box-Jenkins model

Het ontwikkelen van een Box-Jenkins model is - net als bij de meeste modellen - een iteratief proces. Per ronde worden drie fasen doorlopen:

- 1) identificatie van de vorm het model;
- 2) schatting van de modelparameters en
- 3) verificatie van het model.

In het nu volgende beschrijven we de drie fasen afzonderlijk, met als toepassing de ontwikkeling van een transfer-ruismodel.

Identificeren van de modelvorm

Aan de hand van diagnostieken van de relaties tussen de uitvoerreeks en de invoerreeksen, bij voorkeur aangevuld met fysisch inzicht in deze relaties, wordt de vorm van het model gepostuleerd. De diagnostieken worden geleverd door de (steekproef) kruiscorrelatiefuncties van de uitvoerreeks en elke afzonderlijke invoerreeks. Zo'n functie geeft de kruiscorrelatiecoëfficiënt, een maat voor de samenhang tussen de waarde van de invoerreeks en de uitvoerreeks k tijdseenheden later, als functie van k (de "lag"). Een kruiscorrelatie functie wordt vaak geplotted weergegeven (het kruiscorrelogram), omdat dit de diagnose aanzienlijk vergemakkelijkt. Omdat het beeld van de relatie vertroebeld kan worden door de autocorrelaties binnen de reeksen, worden beide vóór het berekenen van de kruiscorrelatiefunctie door een filter gehaald, namelijk het inverse Arima-model voor de invoerreeks. Deze aanpak noemt men ook wel het "witten" [Box and Jenkins, 1976].

In deze identificatiefase dient tevens de vorm van het ruismodel te worden geformuleerd. Hiervoor zijn verschillende benaderingen mogelijk. Als er met grote waarschijnlijkheid van kan worden uitgegaan dat alle relevante dynamische invloedsfactoren in het transfer-ruismodel zijn opgenomen, kan het ruismodel worden geformuleerd als een stationair model. Zo'n model beschrijft een tijdreeks die voldoet aan een bepaald statistisch evenwicht en in feite met een constante spreiding rond een constante waarde schommelt. Als er daarentegen (vermoedelijk)

relevante dynamische invloedsfactoren ontbreken, kan het ruismodel in eerste instantie worden geformuleerd als het Arima-model voor de uitvoerreeks. In alle gevallen zal de juistheid of onjuistheid van de formulering van het ruismodel blijken in de verificatiefase.

Schatten van de modelparameters

De parameters van het geïdentificeerde model worden geschat volgens een optimalisatieprocedure, waarbij de som van de kwadraten van de residuën - een functie van de modelparameters - wordt geminimaliseerd. Dit noemt men ook wel de kleinste kwadraten-methode.

Verifiëren van het model

In de verificatiefase wordt nagegaan of het model de tijdreeks adequaat beschrijft en of voldaan wordt aan alle vooronderstellingen die aan het model ten grondslag liggen (residuën afkomstig uit dezelfde normale kansverdeling, onafhankelijk van elkaar en zonder relatie met de afzonderlijke invoerreeksen). Hierbij wordt gebruik gemaakt van diagnostieken van de residuën, zoals een plot, een histogram en enkele (steekproef-)correlatiefuncties, alsmede de schatting van de variantie van de residuën, de statistische significantie van de modelparameters en de correlatiematrix van de parameterschatters. Bij een onbevredigende diagnose kan een betere modelvorm geformuleerd worden aan de hand van de geconstateerde discrepanties. Het ontwikkelproces vervolgt dan weer met de schatting van de modelparameters.

Betrouwbaarheidsinterval van een geschatte parameter

De schatting van de parameter van een transfer-ruismodel kan opgevat worden als een waarde van een schatter, een stochastische (of toevals-)variabele met een bepaalde kansverdeling. Een schatting gaat gepaard met een maat voor de spreiding van deze kansverdeling: de standaardafwijking (de wortel uit de variantie). De standaardafwijking van een schatter wordt ook wel als standaardfout aangeduid. Als voldaan wordt aan de vooronderstellingen die ten grondslag liggen aan het transfer-ruismodel, heeft een schatter een normale kansverdeling waarvan het gemiddelde overeenkomt met de echte waarde van de te schatten parameter. Aan de hand van de schatting en de standaardfout kan dan het interval worden aangegeven waarbinnen de echte waarde van de parameter zal liggen met een bepaalde betrouwbaarheid (meestal wordt hiervoor 95% genomen). Dit noemt men het (95%) betrouwbaarheidsinterval. Als dit interval de waarde 0 bevat, is de modelparameter statistisch niet significant.

Formuleren van een transfer-ruismodel

De Box-Jenkins formulering van de dynamische relatie tussen een invoerreeks en een uitvoerreeks van een systeem is afgeleid van de algemene weergave van een discreet lineair dynamisch systeem met een lineaire differentie vergelijking. Voor een systeem met één invoerreeks geldt bijvoorbeeld de volgende differentievergelijking:

$$(1 + C_1 D + \dots + C_r D^r) Z_t = g(1 + D_1 D + \dots + D_s D^s) X_t \quad 4$$

met Z de afwijking van de uitvoerreeks ten opzichte van zijn gemiddelde, X de afwijking van de invoerreeks ten opzichte van zijn gemiddelde, t de discrete tijdsindex, $C_1 \dots C_r$ en $D_1 \dots D_s$ modelparameters, \tilde{N} de differentie-operator ($\tilde{N}Z_t = Z_t - Z_{t-1}$) en g de evenwichtsrelatie tussen Z en X (Engels: "steady-state gain"). De evenwichtsrelatie is een belangrijke karakteristiek van een dynamische relatie. Het is in feite de waarde die de uitvoerreeks aan zal nemen als de invoerreeks continu op de waarde +1 wordt gehouden. Of, in het geval van de modellering van de grondwaterstand aan de hand van een grondwaterwinning, de stationaire verlaging van de grondwaterstand bij een eenheid van die grondwaterwinning.

De oplossing van bovenstaande differentievergelijking kan worden uitgeschreven als:

$$Z_t = \sum_{u=0}^{\infty} n_u \cdot X_{t-u}$$

$$= n_0 X_t + n_1 X_{t-1} + n_2 X_{t-2} + \dots \quad 5$$

$$= (n_0 + n_1 B + n_2 B^2 + \dots) X_t$$

met n_0, n_1, n_2, \dots de impuls-respons gewichten (die samen de discrete impuls-respons functie vormen) en B de backshift-operator ($BX_t = X_{t-1}$, let wel: B is geen modelparameter, maar slechts een operator die dient om de notatie te vereenvoudigen!). Deze impuls-respons functie vertegenwoordigt de dynamica van het systeem en beschrijft de reactie van de uitvoerreeks Z op een eenheidspuls in de invoerreeks X . En als X bestaat uit een opeenvolging van verschillende pulsen, wordt Z beschreven als de som van door X geschaalde impuls-respons gewichten.

In Box-Jenkins notatie wordt een differentievergelijking met één invoerreeks uitgeschreven als:

$$(1 - d_1 B - \dots - d_r B^r) Z_t = (w_0 - w_1 B - \dots - w_s B^s) X_t \quad 6$$

met d_1, \dots, d_r de autoregressieve parameters, w_0, \dots, w_s de moving-average parameters en B de backshift-operator (hier gebruikt als substituuut voor \tilde{N} , aangezien $B=1-\tilde{N}$). Dit noemt men de transferfunctie van de orde (r,s) .

In nóg compactere vorm wordt dit ook wel als volgt uitgeschreven:

$$Z_t = \frac{w(B)}{d(B)} X_t \quad 7$$

met $w(B)$ de moving-average operator voor X , volgens:

$$w(B) X_t = w_0 X_t - w_1 X_{t-1} - \dots - w_s X_{t-s} \quad 8$$

en $d(B)$ de autoregressieve operator, die inwerkt op Z , volgens:

$$d(B) Z_t = Z_t - d_1 Z_{t-1} - \dots - d_r Z_{t-r} \quad 9$$

Het Box-Jenkins transfer-ruismodel is een model om een tijdreeks te beschrijven als een som van transferfuncties van de relevante dynamische invloedsfactoren, aangevuld met een ruis-term. In deze zin vormt het een discreet lineair dynamisch stochastisch model. In compacte vorm kan het transfer-ruismodel als volgt worden uitgeschreven:

$$Z_t = \frac{wI(B)}{dI(B)} X_{I_t} + \dots + \frac{wm(B)}{dm(B)} X_{m_t} + N_t \quad 10$$

met X_1, \dots, X_m de dynamische invloedsfactoren en N de ruis.

De tijdreeks van de ruis vertegenwoordigt de invloed van alle invloedsfactoren die niet in het model zijn opgenomen. Over het algemeen bevat zo'n reeks nog regelmatigigheden die beschreven kunnen worden met een univariaat model (ook wel Arima-model genaamd).

Het univariate model

Een univariaat model beschrijft een variabele als lineaire functie van voorgaande waarden van deze variabele en het modelresidu. De algemene vorm van een univariaat model is:

$$f(B)\Phi(B^S)(\nabla^d \nabla_S^D Z_t^l - c) = q(B)\Theta(B^S) a_t \quad 11$$

met Z de variabele, a het modelresidu, t de tijdsindex, \tilde{N} de differentie operator ($\tilde{N}Z_t = Z_t - Z_{t-1}$), d het aantal differenties, S de seizoensperiode (met dagelijkse waarnemingen en een weekcyclus is deze bijvoorbeeld 7), \tilde{N}_S de seizoensdifferentie-operator ($\tilde{N}_S Z_t = N_t - N_{t-S}$), D het aantal seizoensdifferenties, l de transformatieparameter ($l=1$ als Z normaal verdeeld is) en c een constante.

Verder bevat deze formulering de volgende operatoren:

- $f(B)$ de autoregressieve operator, volgens:

$$f(B)Z_t = Z_t - f_1 Z_{t-1} - \dots - f_p Z_{t-p} \quad 12$$

met $f_1 \dots f_p$ de autoregressieve parameters;

- $F(B^S)$ de autoregressieve seizoensoperator, volgens:

$$F(B^S)Z_t = Z_t - F_1 Z_{t-S} - \dots - F_P Z_{t-P \cdot S} \quad 13$$

met $F_1 \dots F_P$ de autoregressieve seizoenparameters;

- $q(B)$ de moving-average operator, volgens:

$$q(B)a_t = a_t - q_1 a_{t-1} - \dots - q_q a_{t-q} \quad 14$$

met $q_1 \dots q_q$ de moving-average parameters;

- en $Q(B^S)$ de moving-average seizoensoperator, volgens:

$$Q(B^S)a_t = a_t - Q_1 a_{t-S} - \dots - Q_Q a_{t-Q \cdot S} \quad 15$$

met $Q_1 \dots Q_Q$ de moving-average seizoenparameters.