Hoezo significant? Over het effect van een ingreep op de grondwaterstand

MARTIN KNOTTERS, PAUL BAGGELAAR EN EIT VAN DER MEULEN

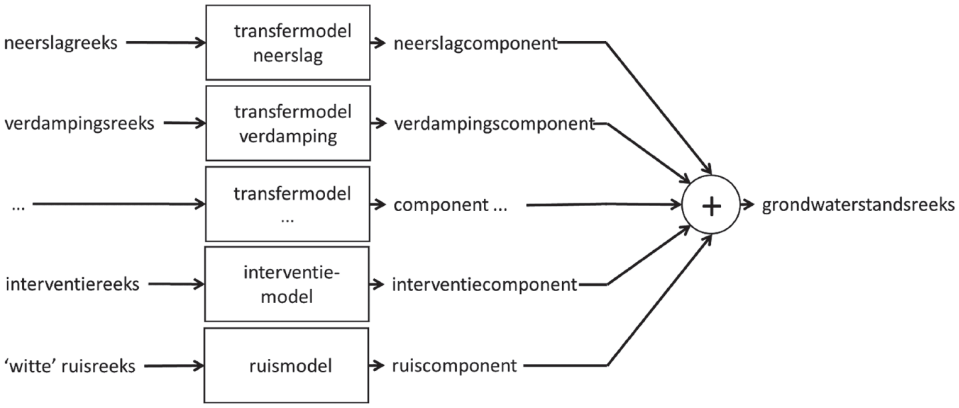
Artikel

Toegepast hydrologisch onderzoek dient niet zelden ter onderbouwing van beslissingen en bevindt zich daarom op het snijvlak van wetenschap en beleid. Op dit snijvlak zijn onzekerheden onvermijdelijk. Statistiek biedt mogelijkheden om met een aantal van deze onzekerheden om te gaan. Een voorbeeld hiervan is tijdreeksmodellering in combinatie met een statistische toets, ter ondersteuning van een beslissing over het uitkeren van schade door wateroverlast als gevolg van een ingreep in de waterhuishouding. Dit voorbeeld maakt duidelijk dat het aankomt op een goede vertaling van het beslisprobleem in een onderzoeksvraag en op een juiste toepassing van statistische methodiek. De gebruikelijke toetsrituelen die op 'significantie' gericht zijn blijken uit te kunnen monden in scheef verdeelde en grote risico's voor de betrokken partijen. In dit artikel laten we zien dat het ook anders kan.

Inleiding

Tijdreeksmodellen worden veel toegepast om de dynamiek van grondwaterstanden te verklaren uit neerslag en verdamping, eventueel samengetrokken tot neerslagoverschot en eventueel aangevuld met invloeden zoals oppervlaktewaterstanden en onttrekkingscijfers. In een transfer-ruismodel noemen we de delen van de dynamiek van de grondwaterstand die uit bekende invloeden kunnen worden verklaard transfercomponenten. Het resterende, onverklaarde, deel van de dynamiek noemen we de ruiscomponent. De som van de verschillende componenten is de grondwaterstand.

Als tijdreeksmodellering tot doel heeft het effect van een ingreep te kwantificeren, spreken we ook wel van interventiemodellering (Hipel e.a., 1975). Eén van de verklarende invloeden is dan een interventie: een tijdreeks met waarden 0 tot het moment van een ingreep en waarden 1 vanaf dat moment. Afbeelding 1 geeft schematisch een transfer-ruismodel weer dat verschillende transfercomponenten heeft, waaronder een interventiecomponent.



Afbeelding 1 Schematische voorstelling van een transfer-ruismodel voor de grondwaterstand, met meerdere transfer-componenten waaronder een interventiecomponent

Het meest eenvoudige effect van een interventie is een staptrend: na een ingreep stelt zich direct een nieuw gemiddeld niveau van de grondwaterstand in. In formulevorm ziet de interventiecomponent er dan als volgt uit:

$$I_t = \delta \cdot S_t$$

waarin S_t een reeks met waarde 0 tot het moment van de interventie is en waarde 1 vanaf dat moment. Het subscript t geeft discrete tijdstappen aan. Transfer-ruismodellen kunnen ook op continue tijd van toepassing zijn, herkenbaar aan de notatie (t). Het onderscheid tussen discreet en continu is voor de boodschap van dit artikel echter niet van belang. De waarde van parameter δ geeft het effect weer van de ingreep op het gemiddelde niveau van de grondwaterstand. Als een transfer-ruismodel met interventiecomponent wordt gefit op een grondwaterstandsreeks, dan levert dit een schatting, δ, op en een standaardfout die de onnauwkeurigheid van deze schatting aangeeft.

De geschatte parameter δ en zijn onnauwkeurigheid staan centraal in dit artikel, in het bijzonder de conclusies die er over de aanwezigheid van een effect kunnen worden getrokken op basis van δ en zijn standaardfout.

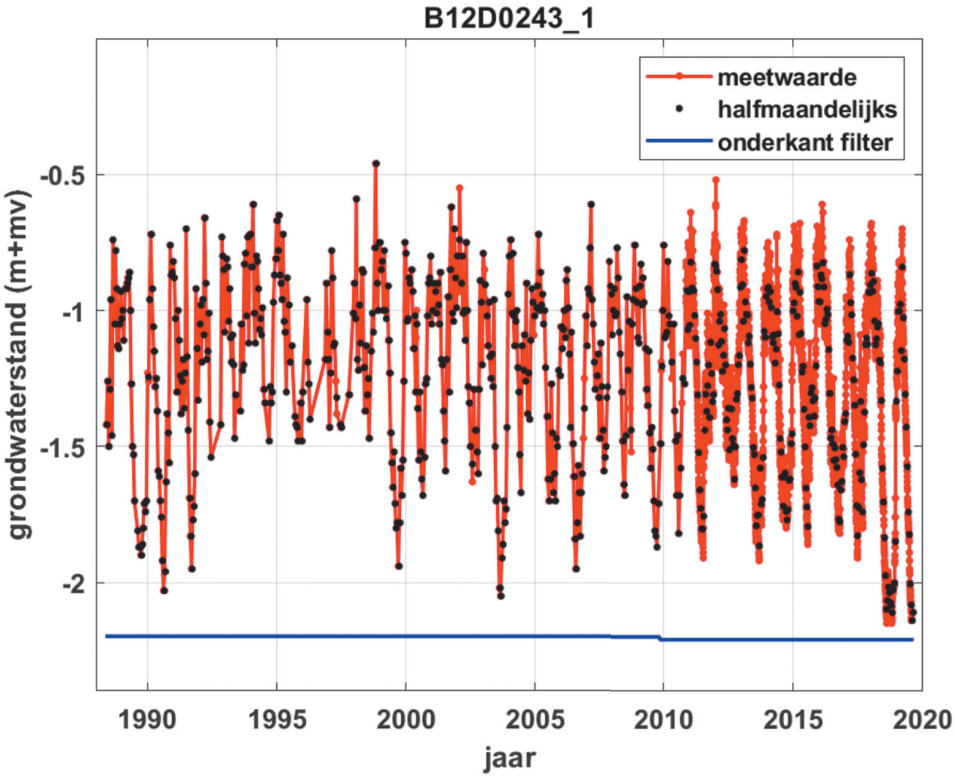
De aanleiding voor dit artikel is het geconstateerde gebruik om te concluderen dat een staptrend statistisch significant is wanneer de absolute waarde van δ groter is dan tweemaal zijn standaardfout. Feitelijk is er dan, zij het enigszins informeel, een statistische toets uitgevoerd. Maar geeft deze toets wel het antwoord op de eigenlijke onderzoeksvraag? Is die vraag goed gesteld? Welke hypothese wordt getoetst? Bij wie ligt de bewijslast? Wat veronderstel je zoal en wat maakt dat uit? Hoe groot zijn de kansen op het trekken van foute conclusies? Welke risico's lopen betrokken partijen? Hoe kunnen kansen op foute conclusies en daarmee samenhangende risico's van foute beslissingen worden beheerst? Het doel van dit artikel is om antwoorden te vinden op deze vragen.

Een grondwaterstandsreeks in Drenthe

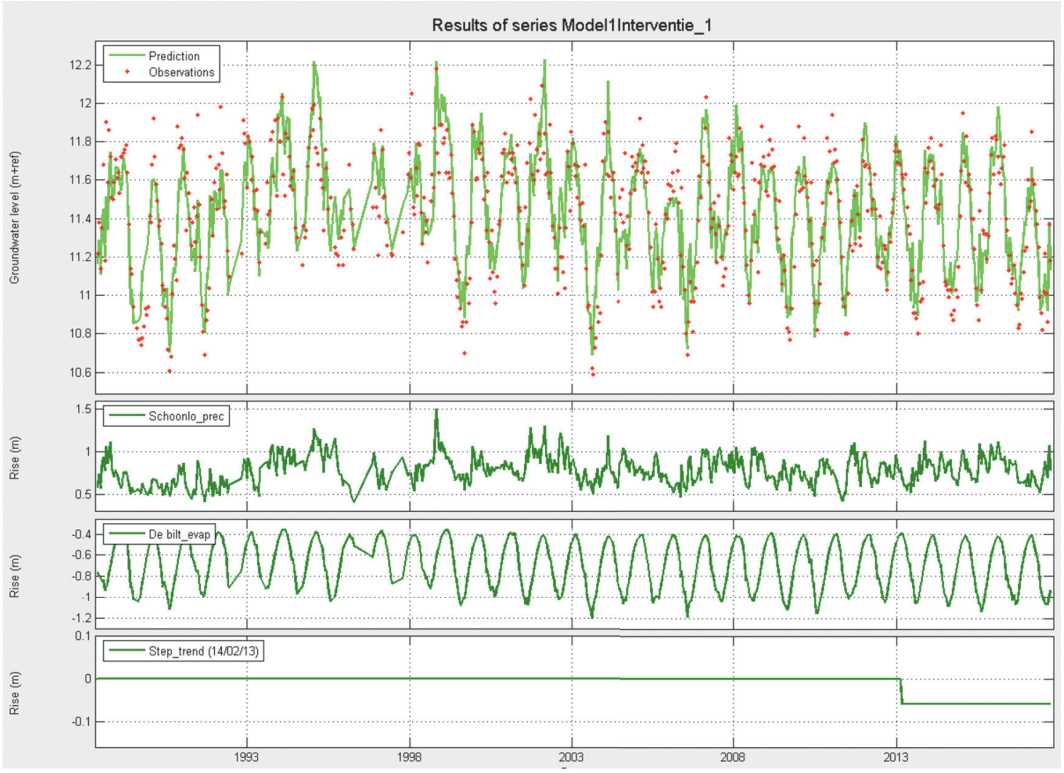
Ter illustratie gebruiken we de grondwaterstandsreeks van buis B12D0243_1

uit Drenthe (Afbeelding 2). Half februari 2013 zijn in de omgeving van deze buis maatregelen genomen voor natuurherstel. De vraag was of deze maatregelen hebben geleid tot hogere grondwaterstanden. In dat geval kunnen landgebruikers in aanmerking komen voor een schadevergoeding. Van de grondwaterstandsreeksen die in dit gebied zijn waargenomen, bleek alleen de reeks van buis B12D0243_1 geschikt voor tijdreeksmodellering, dat wil zeggen: voldoende lang en geen 'gaten'.

Afbeelding 2 laat zien dat aanvankelijk de grondwaterstand halfmaandelijks werd gemeten, maar sinds 2011 veel vaker, waarschijnlijk met een automatische drukopnemer. Om te voorkomen dat deze periode een onevenredig zwaar gewicht krijgt bij de interventiemodellering, hebben wij de gehele reeks 'verdund' tot een halfmaandelijkse frequentie. De interventiemodellering is in 2018 uitgevoerd met het programma 'Menyanthes' (Von Asmuth e.a., 2012). Afbeelding 3 geeft het resultaat hiervan grafisch weer. Het percentage verklaarde variantie bedraagt 78,96%.



Afbeelding 2 Tijdreeksgrafiek van reeks B12D0243_1



Afbeelding 3 Interventiemodellering van reeks B12D0243_1, uitgevoerd met het programma Menyanthes. Bovenste grafiek: rode stippen zijn grondwaterstandswaarnemingen, de lichtgroene lijn het deel van de grondwaterstand dat kan worden verklaard uit de invloeden "neerslag", "verdamping" en "staptrend", waarvan de componenten zijn weergegeven in de volgende drie grafieken.

De interventiecomponent (onderste grafiek in Afbeelding 3) laat zien dat sinds de ingreep in februari 2013 de grondwaterstand gemiddeld niet zou zijn gestegen en zelfs zou zijn verlaagd: Menyanthes schat deze verlaging op 6 cm (0,0589 m), met een standaardfout van 3 cm. Uitgaande van de 'gewoonte' om een parameterwaarde significant te noemen als zijn absolute waarde groter is dan tweemaal de standaardfout, is deze verlaging dus nét niet statistisch significant.

Wat concluderen we?

Hoe luidt nu de conclusie, en wat betekent deze conclusie voor een beslissing over uitkering van schade? Mogelijke conclusies over het effect van de maatregelen in februari 2013 op de grondwaterstand op locatie B12D0243_1 zijn onder meer:

1. De gemiddelde grondwaterstand is niet gestegen na de maatregelen;
2. Het is niet aangetoond dat de gemiddelde grondwaterstand is gestegen na de maatregelen;
3. Het is niet aangetoond dat de gemiddelde grondwaterstand is veranderd na de maatregelen;
4. De gemiddelde grondwaterstand is gedaald na de maatregelen.

Al deze conclusies zijn 'gegeven het model': ervan uitgaande dat aan de modelveronderstellingen van het gebruikte tijdreeksmodel wordt voldaan, trekken we een conclusie op basis van staptrendparameter en zijn standaardfout.
Op 'gegeven het model' komen we later terug, wanneer we de analyse herhalen met andere modellen. Eerst gaan we in op de conclusies en hun consequenties.

De vraag was "of de maatregelen hebben geleid tot hogere grondwaterstanden", gevolgd door een beslissing of landgebruikers in aanmerking kunnen komen voor een schadevergoeding. Aan het antwoord op de vraag zijn dus twee risico's verbonden: dat landgebruikers ten onrechte schade krijgen vergoed en dat landgebruikers ten onrechte geen schade krijgen vergoed.

Conclusie 1 en 4 klinken stelliger dan conclusie 2 en 3. Bij conclusie 1 zal worden besloten dat landgebruikers niet in aanmerking komen voor schadevergoeding en daarmee is de kous af. Hoe zeker zijn wij er echter van dat deze conclusie en beslissing juist zijn? Wat is de kans dat in werkelijkheid de grondwaterstand wél is gestegen en de landgebruikers dus ten onrechte niet in aanmerking komen voor vergoeding? Conclusie 2 en 3 klinken minder stellig en zouden kunnen leiden tot nader onderzoek. Conclusie 4 kan er zelfs toe leiden dat landgebruikers in aanmerking komen voor een vergoeding van droogteschade in plaats van de verwachte natschade en ook dan is de vraag hoe zeker we zijn dat deze conclusie juist is.

Tot nu toe gingen we uit van de vraag of de maatregelen hebben geleid tot hogere grondwaterstanden. Je kunt je echter afvragen of dit de goede vraag is in deze situatie. Want als je deze vraag probeert te beantwoorden, dan ligt de bewijslast feitelijk bij de landgebruikers. Zou de vraag niet moeten zijn: "Is verhoging van de gemiddelde grondwaterstand groter dan een bepaald minimum, waarboven landgebruikers in aanmerking komen voor schadevergoeding?" Zo gesteld moet de instantie die de maatregelen nam aantonen dat het effect ervan op de gemiddelde grondwaterstand binnen een bepaalde grens blijft.

Een ogenschijnlijk eenvoudige vraag, met een ogenschijnlijk eenvoudig antwoord, gebaseerd op de eenvoudige vuistregel dat een parameter statistisch significant is als de absolute waarde ervan groter is dan tweemaal de standaardfout, wordt nu toch nog ingewikkeld. De toets was gericht op het vinden van *bewijs* voor een verhogend effect van de maatregelen op de gemiddelde grondwaterstand. Doordat we tweemaal de standaardfout als criterium hanteerden, beperkten we de kans op ten onrechte concluderen dat er een verhogend effect tot circa 0,025. Nu we beseffen dat uit de conclusie een *beslissing* volgt, met risico's voor twee partijen, zijn we ook geïnteresseerd in de kans dat we ten onrechte *niet* concluderen dat de maatregelen een verhogend effect hebben op de gemiddelde grondwaterstand. Op het verschil tussen toetsen om te *bewijzen* en toetsen om te *beslissen* gaan we in de volgende paragraaf in.

Twee paradigma's

Het gebruik van statistische toetsen staat regelmatig ter discussie. Deze discussie gaat terug tot een controverse die in de jaren vijftig en zestig van de vorige

eeuw speelde tussen Sir Ronald Aylmer Fischer enerzijds en Jerzy Neyman en Egon Pearson anderzijds. Enig inzicht in deze controverse kan heel verhelderend zijn wanneer je een statistische toets wilt toepassen. Het antwoord op de vraag 'Waarom toets je?' blijkt belangrijk te zijn. Wij lichten een tipje van de sluier op, voor een uitvoerige beschrijving verwijzen we naar Hubbard (2004).

Sir Ronald Aylmer Fisher ontwikkelde bijna een eeuw geleden significantietoetsen met als doel bewijs tegen de nulhypothese te vinden in de resultaten van een experiment of ander onderzoek (Fisher, 1925; 1935a,b). Hierbij speelt de *p value* een cruciale rol: de kans op de resultaten, of extremere, als de nulhypothese, die luidt dat er geen effect of relatie is, waar zou zijn. Hoe kleiner deze *p value*, hoe meer twijfel over de houdbaarheid van de nulhypothese. Het vertrouwen in de nulhypothese wordt vaak opgezegd als de *p value* kleiner is dan 0,05. Dit noemen we het significantieniveau. Komt de *p value* onder het significantieniveau, dan spreken we van *evidence*, bewijs, en concluderen we dat een statistisch significant effect of een significante relatie is aangetoond.

Ongeveer tegelijkertijd met Fisher werkten Jerzy Neyman en Egon Pearson aan de ontwikkeling van hypothesetoetsen (Neyman en Pearson, 1928a en b, 1933). Zij stelden tegenover een hoofdhypothese, H_H (H=hoofd), een alternatieve hypothese, H_A: als de resultaten niet waarschijnlijk zijn onder H_H, dan moet er een H_A zijn die de uitkomsten verklaart. Verder introduceerden zij de kans op twee typen fouten die je kunt maken bij een hypothesetoets: een type-I-fout (ten onrechte verwerpen van H_H) en een type-II-fout (ten onrechte accepteren (!) van H_H).

Nu zal menigeen hebben geleerd dat je de nulhypothese niet kunt accepteren. Volgens Fisher kun je dat inderdaad niet, maar volgens Neyman en Pearson kan dat wel met de hoofdhypothese. Hun hypothesetoetsen moesten helpen bij het nemen van beslissingen: onderneem ik actie A of actie B? Keur ik iets goed of af, bijvoorbeeld. Deze beslissing staat los van het vertrouwen dat een onderzoeker hecht aan een hypothese. Wel worden de eventuele gevolgen van de beslissing onder ogen gezien: de kans op een verkeerde beslissing wordt uitgedrukt in de *error rates* α (de kans op een type-I-fout) en β (de kans op een type-II-fout). Ook intruduceerden Neyman en Pearson de *power* of onderscheidingsvermogen van een toets: $1-\beta$ (de kans dat een H_H die niet opgaat ook daadwerkelijk wordt verworpen). De *error rates* α en β worden voorafgaand aan het onderzoek gekozen, bijvoorbeeld op basis van een kosten-batenanalyse, en zijn samen met het kleinste relevant geachte effect bepalend voor de steekproefomvang.

Vanwege de directe betekenis voor het nemen van beslissingen, gedrag dus, worden hypothesetoetsen volgens Neyman en Pearson *behavioral* genoemd, ter onderscheid van de significantietoetsen volgens Fisher, die *evidential* worden genoemd. *Behavioral* hypothesetoetsen vinden bijvoorbeeld hun toepassing in kwaliteitscontroles (zie volgende paragraaf), terwijl *evidential* significantietoetsen wetenschappelijk onderzoekers helpen bij de richting van hun onderzoek.

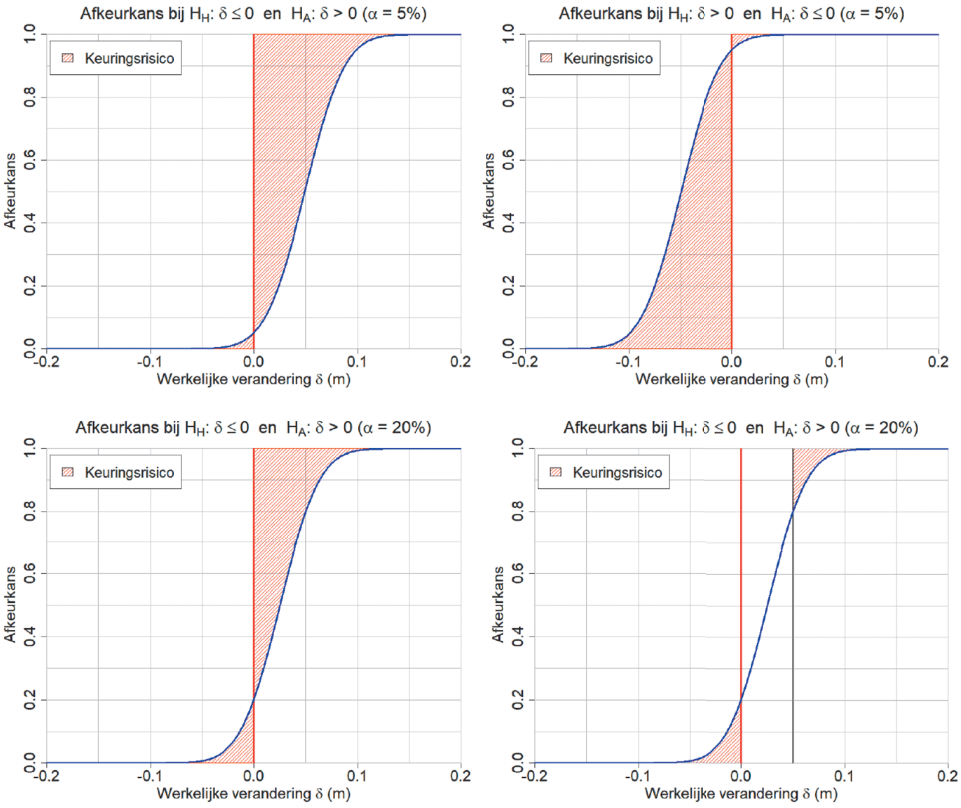
Een parallel met kwaliteitskeuring

In de lijn van *behavioral* hypothesetoetsen (Neyman-Pearson) werken we nu de vraagstelling of de grondwaterstand bij B12D0243_1 is gestegen door de inrichting van een natuurgebied uit als een keuringsprobleem. Bij een keuring zijn twee partijen betrokken met tegenstrijdige belangen: hier de agrariërs en de natuurbeheerder. We keuren de situatie af als de ingreep de grondwaterstand heeft verhoogd (de agrariërs hebben dan recht op een vergoeding) en we keuren goed als dat niet het geval is. Beide partijen lopen keuringsrisico's: de agrariërs op onterecht goedkeuren van de situatie (het concluderen dat er geen verhoging is, terwijl die er wel is) en de natuurbeheerder op onterecht afkeuren van de situatie (het concluderen dat er verhoging is, terwijl die er niet is).

Het zou voor de hand liggen deze keuringsrisico's niet alleen zo beperkt mogelijk te houden, maar ook om ze evenredig te verdelen over de twee partijen. Keuringsrisico's kunnen vooraf worden beperkt door de ruimtelijke en temporele meetinspanning af te stemmen op de risico's die nog als aanvaardbaar worden geacht, maar daar is het nu uiteraard te laat voor en moeten we het doen met één meetreeks. Het is nog wél mogelijk de keuringsrisico's gelijk te verdelen over de partijen wanneer een hypothesetoets (Neyman-Pearson) wordt toegepast. Wordt echter een significantietoets (Fisher) toegepast, met een nulhypothese "er is geen effect" en een significantieniveau van bijvoorbeeld 0,05, dan kan de ene partij veel meer risico lopen dan de andere. Wellicht is men er zich niet van bewust dat een evenredige verdeling van risico's mogelijk is, of wil men het belang van één partij zwaarder laten meewegen. Bij milieudisputen kan het voorzorgsprincipe bijvoorbeeld een reden zijn om asymmetrische keuringsrisico's te rechtvaardigen. Het is echter de vraag of beide partijen wel weten welke risico's zij precies lopen. Keuringskarakteristieken maken die risico's duidelijk.

Een keuringskarakteristiek geeft de relatie weer tussen de afkeurkans en de werkelijke waarde van de te keuren variabele. Afbeelding 4 toont vier voorbeelden van keuringskarakteristieken voor het geval B12D0243_1. Elk toont de afkeurkans (verticale as) als functie van de werkelijke verandering van de grondwaterstand door de ingreep (δ, horizontale as). Voor $\delta \leq 1$ is er geen sprake van een verhoging door de ingreep en voor $\delta > 0$ wel. De gearceerde deelgebieden zijn de keuringsrisicogebieden van de twee partijen. Het gearceerde deelgebied links van $\delta = 0$ vertegenwoordigt het keuringsrisico voor de natuurbeheerder (risico op onterecht afkeuren, type-I-fout of α) en het gearceerde deelgebied rechts van $\delta = 0$ dat voor de agrariërs (risico op onterecht goedkeuren, type-II-fout, of β). In het ideale geval wordt er bij $\delta \leq 0$ nooit afgekeurd en bij $\delta > 0$ altijd, zodat er helemaal geen keuringsrisico's optreden. Maar zo'n risicoloze keuring is alleen mogelijk als de verandering door de ingreep volledig foutloos wordt geschat, wat onhaalbaar is. De weergegeven keuringskarakteristieken zijn realistischer, omdat die de onnauwkeurigheid van het berekende effect van de ingreep verdisconteren (standaardfout bedraagt 3 cm, met 450 vrijheidsgraden, zie voorgaande). Tabel 1 vermeldt hun kenmerken. De bovenste twee gaan uit van een kans op type-I-fout $\alpha = 5\%$ en de onderste twee van $\alpha = 20\%$.

Uit de grafiek linksboven blijkt dat als de hoofdhypothese van géén verhoging wordt gehanteerd de keuringsrisico's vrijwel volledig bij de agrariërs liggen. Pas bij een werkelijke verhoging van 10 cm of meer wordt de situatie vrijwel zeker (dat wil zeggen met minstens 95% kans) afgekeurd, wat inhoudt dat er voor het traject 0 tot 10 cm werkelijke verhoging een substantieel risico is op onterechte goedkeuring, i.c. geen schadeuitkering. Maar als daarentegen de hoofdhypothese van wél verhoging wordt gehanteerd, blijkt dat de keuringsrisico's sterk verschuiven naar links, naar de natuurbeheerder (grafiek rechtsboven).



Afbeelding 4 Keuringskarakteristiek van vier toetsuitvoeringen, met de keuringsrisico's voor de twee betrokken partijen.

Tabel 1 Kenmerken van de vier keuringskarakteristieken van afbeelding 4

Afbeelding 4	Combinatie van hypothesen	α	Afkeuren situatie bij
Linksboven	H_H: $\delta \leq 0$ en HA: $\delta > 0$	5%	Verwerpen H_H
Rechtsboven	H_H: $\delta > 0$ en HA: $\delta \leq 0$	5%	Niet verwerpen H_H
Linksonder	H_H: $\delta \leq 0$ en HA: $\delta > 0$	20%	Verwerpen H_H
Rechtsonder[1]	H_H: $\delta \leq 0$ en HA: $\delta > 0$	20%	Verwerpen H_H

1 De partijen hebben hier afgesproken dat alleen een werkelijke verhoging van meer dan 5 cm praktisch relevant wordt geacht, zodat er voor het traject 0 t/m 5 cm werkelijke verhoging nog geen sprake is van een keuringsrisico voor de agrariërs (onterecht goedkeuren)

Een evenwichtigere verdeling van de keuringsrisico's kan worden bereikt door α te verhogen, zoals te zien in de grafiek linksonder, waar deze is verhoogd naar 20%. Er is zelfs een volledig symmetrische verdeling van de keuringsrisico's te bewerkstelligen door α op 50% in te stellen.
Als de twee partijen afspreken dat er sprake is van een kleinste relevant geachte verhoging – zoals 5 cm - waarboven de keuringsrisico's voor de agrariërs pas consequenties hebben en daarmee dus ook pas gaan meetellen, gaat de grafiek linksonder over in de grafiek rechtsonder en is er sprake van een symmetrische verdeling van de keuringsrisico's over de twee partijen. Het keuringsrisico van de natuurbeheerder (α) bedraagt dan 20% voor δ = 0 cm en dat van de agrariërs (β) bedraagt dan 20% voor δ = 5 cm.

Uit bovenstaande leiden wij af dat bij het formuleren van de combinatie van hoofdhypothese en alternatieve hypothese voor een keuringsprobleem zoals bij reeks B12D0243_1, mede dient te worden afgegaan op de bijbehorende keuringskarakteristieken. Bij voorkeur gebeurt dit in samenspraak met de twee betrokken partijen. Anders is er geen garantie dat de keuringsrisico's eerlijk zijn verdeeld.

'Gegeven het model'

Reeks B12D0243_1 modelleerden we met een Pirfict-model met behulp van Menyanthes, met een halfmaandelijkse frequentie van grondwaterstanden en voor een periode van 27 mei 1988 tot 28 september 2017, zie afbeelding 3. Alle voorgaand beschreven analyses zijn gegeven dit model. De keuze voor een ander model kan echter leiden tot een andere schatting voor δ en voor een andere standaardfout. Ook de keuzes voor modelperiode en -frequentie kunnen van invloed zijn. Om inzicht te krijgen in de invloed van keuzes bij het modelleren op het toetsresultaat modelleerden we de reeks ook met Pastas (Collenteur e.a., 2019) en TRG (Van der Meulen en Baggelaar, 2019), voor periodes van verschillende lengte en voor reeksen met verschillende frequenties. Tabel 2 vat de verschillende resultaten voor δ en zijn standaardfout samen. Duidelijk is dat verschillende keuzes leiden tot verschillende uitkomsten: de geschatte staptrends variëren van circa 6 tot 9 cm en de standaardfouten schommelen tussen 1,39 en 3,85 cm.

Tabel 2 Staptrendparameter en bijbehorende standaardfout (s.e.), geschat met verschillende software en tijdreeksmodellen, verschillende invoerreeksen, periodes en frequenties. Invoerreeksen: neerslag (N), potentiële referentiegewasverdamping (V) en potentieel neerslagoverschot (PNO=N-V). De periode start op 27 mei 1988. De staptrend is gemodelleerd voor de datum 4 februari 2013. h=halfmaandelijkse frequentie, m=maandfrequentie. De residuen zijn getoetst op afwijking van normaliteit (Lillieforstoets) en aanwezigheid van autocorrelatie (Portmanteautoets). "" betekent dat de nulhypothese van normaal verdeelde resp. ongecorreleerde residuen wordt verworpen (bij een significantieniveau van 0,05). "?": geen toetsresultaat bekend.*

Software, model	invoer	einde periode	frequentie	δ (cm)	s.e. cm)	Normaliteits -toets	Toets op autocorrelatie
Menyanthes, Pirfict	N, V	28-9-2017	h	-5,89	3,00	?	?
Pastas, Pirfict	N, V	28-9-2017	h	-6,66	2,75	*	
Pastas, Pirfict	PNO	28-9-2017	h	-7,65	3,13	*	
Pastas, Pirfict	N, V	28-9-2017	m	-6,81	2,97		
Pastas, Pirfict	PNO	28-9-2017	m	-7,80	3,32		
Pastas, Pirfict	N, V	17-8-2019	h	-8,12	2,60	*	
Pastas, Pirfict	PNO	17-8-2019	h	-8,97	1,39		*
Pastas, Pirfict	N, V	17-8-2019	m	-8,58	2,81		
Pastas, Pirfict	PNO	17-8-2019	m	-9,55	1,88		*
TRG, Box-Jenkins	N, V	28-9-2017	h	-7,04	3,19	*	
TRG, Box-Jenkins	PNO	28-9-2017	h	-7,20	3,48		
TRG, Box-Jenkins	N, V	28-9-2017	m	-7,43	3,33		
TRG, Box-Jenkins	PNO	28-9-2017	m	-8,15	3,85		
TRG, Box-Jenkins	N, V	17-8-2019	h	-9,01	2,78		
TRG, Box-Jenkins	PNO	17-8-2019	h	-9,10	3,00		
TRG, Box-Jenkins	N, V	17-8-2019	m	-8,85	2,90		
TRG, Box-Jenkins	PNO	17-8-2019	m	-9,59	3,31		

Als de gemodelleerde periode doorloopt tot 17 augustus 2019 dan worden grotere negatieve staptrends geschat. Mogelijk komt dit door de niet-lineaire effecten in de droge zomers van 2018 en 2019 die niet in de lineaire transfercomponent voor neerslag, verdamping of potentieel neerslagoverschot zijn beschreven en 'terechtkomen' in de staptrend. Dit maakt duidelijk dat modelveronderstellingen, in dit geval over lineaire samenhang, van invloed zijn op het resultaat van een toets op het effect van een ingreep. Toetsen op afwijking van normaliteit en aanwezigheid van autocorrelatie indiceren dat in een aantal gevallen niet aan de modelveronderstellingen van normaal verdeelde, onafhankelijke residuen wordt voldaan. De geldigheid van deze veronderstellingen is van belang wanneer je conclusies trekt of beslissingen neemt op basis van de uitkomsten van statistische toetsen. Als immers niet aan deze veronderstellingen wordt voldaan is statistische significantie niet goed te beoordelen en zijn de risico's van foute beslissingen niet goed in te schatten.

Enkele concluderende opmerkingen

Dit artikel laat zien dat kennis over statistische toetsen belangrijk is wanneer een beslissing over schadeuitkering moet worden genomen op basis van de resultaten van interventiemodellering. Het ritueel toepassen van een significantietoets, met 'er is geen effect' als nulhypothese en een significantieniveau dat

traditiegetrouw op 0,05 staat, blijkt te leiden tot een scheve verdeling van risico's van verkeerde beslissingen tussen de twee betrokken partijen.

Een hypothesetoets, zoals toegepast bij kwaliteitscontroles, maakt het mogelijk om risico's van verkeerde beslissingen te kwantificeren en weloverwogen te verdelen over de betrokken partijen, bijvoorbeeld gelijkmatig. Dit vereist dat voorafgaand aan het onderzoek het beslisprobleem wordt geformuleerd in het raamwerk van een hypothesetoets. Dit komt overeen met de aanbeveling die Knotters e.a. (2017) en Poortvliet e.a. (2019, 2020) doen om de benutting van statistische informatie over onzekerheid te verbeteren bij het nemen van beslissingen in het kwantitatief waterbeheer.

Naast een juiste toepassing van statistische toetsen is het van belang bij de modellering van het effect van een ingreep met een tijdreeksmodel wordt voldaan aan de modelveronderstellingen, zoals lineariteit en onafhankelijke, normaal en gelijk verdeelde residuen (Baggelaar en Van der Meulen, 2020). Onnauwkeurigheid en daarvan afgeleide risico's bij toetsen zouden anders kunnen worden over- of onderschat. Bovendien is het van belang dat de tijdreeksen gescreend zijn en vrij zijn van trends die aan het meetproces zijn toe te schrijven (Baggelaar en Van der Meulen, 2019), teneinde effecten van ingrepen die buiten het meetproces liggen goed te kunnen kwantificeren met interventiemodellering.

Samengevat illustreert dit artikel dat kennis van statistische technieken en hoe die toe te passen bijzonder nuttig kan zijn bij zowel de uitvoering van tijdreeksmodellering als de toepassing van de resultaten ervan bij het trekken van conclusies en het nemen van beslissingen.

Literatuur

Asmuth, J.R. von, K. Maas, M. Knotters, M.F.P. Bierkens, M. Bakker en T.N. Olsthoorn (2012) Menyanthes: software for hydrogeologic time series analysis, interfacing data with physical insight; in: *Environmental modeling and software*, vol 38, pag 178-190.

Baggelaar P.K. en E. van der Meulen (2019) Kenmerken meetfout bij automatisch meten grondwaterstand door Provincie Overijssel; Rapport van AMO (Hengelo) en PB Icastat (Amstelveen).

Baggelaar, P. en E. van der Meulen (2020) Naar betere tijdreeksmodellering met Pastas; Rapport van AMO (Hengelo) en PB Icastat (Amstelveen).

Collenteur, R.A., M. Bakker, R. Caljé, S.A. Klop en F. Schaars (2019) Pastas: open source software for the analysis of groundwater time series; in: *Groundwater*, vol 57, pag 877-885; doi: 10.1111/gwat.12925.

Fisher, R.A. (1925) Statistical methods for research workers; Oliver & Boyd, Edinburgh.

Fisher, R.A. (1935a) The design of experiments; Oliver & Boyd, Edinburgh.

Fisher, R.A. (1935b) Statistical tests; in: *Nature*, vol 136, pag 474.

Hipel, K.W., W.C. Lennox, T.E. Unny en A.I. McLeod (1975) Intervention analysis in water resources; in: *Water Resources Research,* vol 11, pag 855-861.

Hubbard, R. (2004) Alphabet soup. Blurring the distinctions between *p*'s and α's in psychological research; in: *Theory & Psychology*, vol 14, pag 295-327.

Knotters, M., P.M. Poortvliet, J. Verstoep, J. van Wijk en P. Bergsma (2017) Communiceren van statistische informatie over onzekerheid in het tactisch-strategische waterkwantiteitsbeheer; STOWA rapport 2017-21, Amersfoort.

Meulen, E.C.J. van der en P.K. Baggelaar (2019) Tijdreeksanalist (TRGnet en TRG) – Tijdreeksanalyse met uitgebreide Box-Jenkins modellering – Versie 6; AMO, Hengelo.

Neyman, J. en E.S. Pearson (1928a) On the use and interpretation of certain test criteria for purposes of statistical inference. Part I; in: *Biometrika*, vol 20A, pag 175–240.

Neyman, J. en E.S. Pearson (1928b) On the use and interpretation of certain test criteria for purposes of statistical inference. Part II; in: *Biometrika*, vol 20A, pag 263–294.

Neyman, J. en E.S. Pearson (1933) On the problem of the most efficient tests of statistical hypotheses; in: Philosophical Transactions of the Royal Society of London, Ser. A, vol 231, pag 289–337.

Poortvliet, P.M., M. Knotters, P. Bergsma, J. Verstoep en J. van Wijk (2019) On the communication of statistical information about uncertainty in flood risk management; in: Safety Science, volume 118, pag 194-204.

Poortvliet, P.M., M. Knotters, P. Bergsma, J. Verstoep en J. van Wijk (2020) Communicatie van statistische informatie over onzekerheid bij de beheersing van risico's van wateroverlast; in: Stromingen, volume 2020(2), pag 49-63.

Summary What do you mean 'significant'? On the effect of interventions on water table depths

Applied hydrological research often aims to support decision making and thus balances at the edge of science and policy, where uncertainties are unavoidable. Statistics might offer ways to deal with at least some of these uncertainties. An example is time series modelling combined with statistical testing to support a decision on compensation to farmers for financial losses due to water level rises after an intervention in the groundwater regime. Both a proper translation of the decision problem into a research question and a prudent application of statistical methods appear to be crucial in this case. The usual ritual of null hypothesis significance testing, using a significance level of 0.05, appears to end up in risks that can be unacceptably large for one of the parties involved. Behavioral hypothesis tests, that are widely applied in quality control, are more appropriate to support decision making, since they allow for a good balance of the error rates, i.e. the probabilities of wrongly rejecting or accepting the main hypothesis. It is concluded that basic statistical knowledge is indispensable in hydrological time series modelling and making inference from modelling results.

Auteurs

MARTIN KNOTTERS
Wageningen Environmental Research
martin.knotters@wur.nl

PAUL K. BAGGELAAR
PB Icastat
paul.baggelaar@planet.nl

EIT VAN DER MEULEN
AMO
amo@home.nl