DE ZOEKTOCHT NAAR EENVOUD

Betere multivariate statistiek én levenswetenschappen

door prof.dr. Cajo J.F. ter Braak

persoonlijk hoogleraar 'Multivariate statistiek voor de levenswetenschappen'

WAGENINGEN UNIVERSITEIT
WAGENINGEN UR

Inaugurele rede gegeven op 2 november 2006 in de Aula van Wageningen Universiteit

De zoektocht naar eenvoud:

betere multivariate statistiek én levenswetenschappen

Inleiding

Mijnheer de Rector Magnificus, dames en heren,

Wetenschap gaat uit van eenvoud, gaat uit van een wereld die ontleedbaar is in delen die relatief eenvoudig zijn of relatief eenvoudig bestudeerd kunnen worden. Juist als het ons doel is een systeem in zijn geheel te begrijpen, zullen we om ons doel te bereiken ergens moeten vereenvoudigen. Met het dogma van eenvoud is veel bereikt. Kijk maar naar de stand van kennis en techniek van nu ten opzichte van die in het jaar nul. Nogal wiedes zult u zeggen, maar les 1 uit de statistiek is dat als je een trend wilt detecteren je het best een lange periode kunt beschouwen.
Statistiek gaat over het analyseren, presenteren en interpreteren van gegevens en, gepaard daaraan, over hoe je die gegevens het best kunt verzamelen via experimenten of steekproeven. Trevor Hastie, Robert Tibshirani en Jerome Friedman geven in hun fantastische boek The Elements of Statistical Learning uit 2001 de korte omschrijving: Statistiek is het leren uit gegevens[1]. Het eindresultaat - de conclusie, de interpretatie van de gegevens - dient eenvoudig te zijn, dat wil zeggen: niet ingewikkelder dan nodig. Dit is een belangrijk doel van statistiek. Voor mij is statistiek een zoektocht naar eenvoud, eenvoud in het resultaat maar liefst ook in de methode om tot het resultaat te komen. De methode moet natuurlijk bovenal ook effectief zijn.

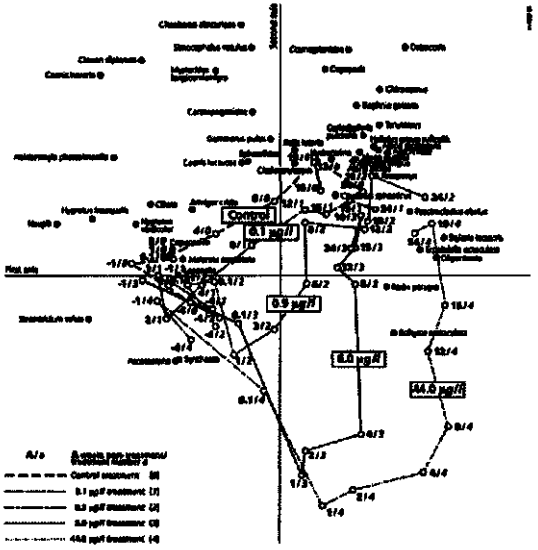
De zoektocht naar eenvoud en effectiviteit

Ik wil u iets van die zoektocht laten zien. Ik begin met een voorbeeld van zo'n zoektocht in de multivariate statistiek. Het voorbeeld zal, hoop ik, ook duidelijk maken wat multivariate statistiek inhoudt.

Naar een eenvoudige weergave van tijdsafhankelijke effecten

Bij de beoordeling van de risico's van bestrijdingsmiddelen in de landbouw zijn niet alleen de gezondheidsrisico's voor de mens van belang, maar ook de risico's voor het ecosysteem waarin het middel wordt gebruikt. Om voor toelating in aanmerking te komen, moeten de neveneffecten beperkt zijn en/of van beperkte duur. De mogelijke neveneffecten op waterorganismen worden ondermeer bestudeerd in semi-veldproeven. In zo'n proef worden experimentele sloten of aquaria behandeld met verschillende doses van het bestrijdingsmiddel en vervolgens wordt gekeken wat het verloop is in de tijd van de aantallen individuen van de verschillende soorten waterdieren. Dat kunnen wel honderden verschillende soorten zijn, dus de uitkomsten van de proef worden al snel onoverzichtelijk. In de literatuur van dit vakgebied (de ecotoxicologie) werden in de 90-er jaren van de vorige eeuw vele verschillende multivariate statistische methoden toegepast om de resultaten van zo'n proef weer te geven. De artikelen hierover blonken, op zijn zachtst gezegd, niet uit in eenvoud en overzichtelijkheid. Dat moet toch beter kunnen, dacht Paul van den Brink, de onderzoeker van Alterra met wie ik samen onderzoek naar het gebruik van multivariate technieken in de ecotoxicologie heb gedaan.

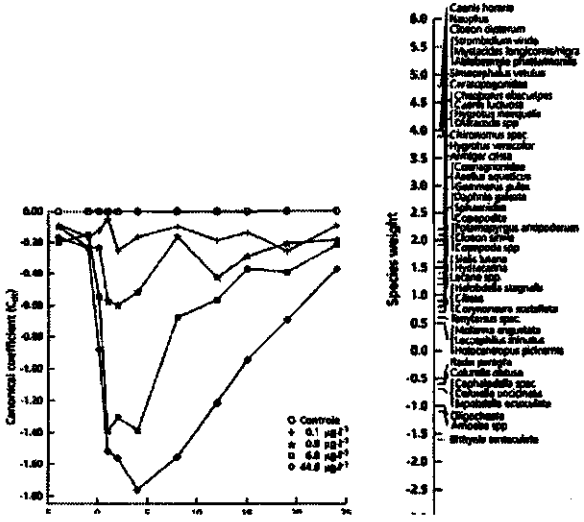
Fig. 1. Biplot op basis van redundantie-analyse die de effecten laat zien van een eenmalige toediening van het insecticide chlorpyrifos op invertebraten in proefsloten (bron: van den Brink *et al.* 1996).



Zijn eerste poging tot verbetering (van den Brink *et al.* 1996) was een redundantie-analyse met CANOCO, mijn computerpakket voor multivariate analyse van ecologische gegevens. Een resultaat was deze figuur (Fig. 1)- een biplot, niet bepaald het toppunt van eenvoud, al was het alleen maar omdat ik u eerst zou moet leren hoe u een biplot moet lezen (Gabriel 1971, Jongman *et al.* 1995 en Gower and Hand 1996). Ik vertel u alleen wát er staat en wat eruit valt af te lezen. In een biplot staan twee verschillende entiteiten, hier soorten (de namen) en monsters. Een monster is een sloot gemeten op een bepaalde dag. In de figuur zien we dat

de monsters vóór toepassing van het bestrijdingsmiddel dicht bij elkaar liggen, dus in die monsters komen ongeveer dezelfde soorten voor.

Fig. 2. Principal response curves met soortsgewichten (rechts) voor de gegevens in fig. 1. De canonische coëfficiënt is een maat voor hoeveelheid uitgedrukt als afwijking van de controle (bron: van den Brink *et al.* 1999).



De verbindingslijn voor de controle (de monsters met dosis nul) loopt van links-midden in de figuur naar rechts-boven. De verbindingslijn voor de monsters die de hoogste dosis hebben gekregen wijken het sterkst af van die voor de controle-monsters, maar aan het eind van het experiment is de afwijking niet meer zo groot. Aan de positie van de soorten valt af te lezen dat vele soorten sterk afnemen door het be-

strijdingsmiddel en dat een enkele soort (een slak) wat toeneemt. Voor zijn tijd (1996) was deze biplot al een toonbeeld van inzichtelijkheid. Toegepast op een ander gegevensbestand leidde de analyse helaas niet tot een inzichtelijke biplot. Het moest beter kunnen! Tegelijkertijd deed oud-collega Jan Oude Voshaar statistische analyses waarin steeds één soort afzonderlijk werd geanalyseerd. Elke analyse liet mooi zien hoe het behandelingseffect in de tijd verliep voor die ene soort. Mijn credo is: "Wat univariaat kan, moet ook multivariaat kunnen" Dat wil hier zeggen: "Wat voor één soort kan, moet voor alle soorten tezamen (voor de hele levensgemeenschap) kunnen". Na enige tijd leidde dit tot onze 'Principal Response Curve' (PRC) techniek (van den Brink and ter Braak 1999) waarvan hier het resultaat (Fig. 2). Horizontaal staat de tijd in weken na de toediening van het bestrijdingsmiddel en verticaal staat een maat voor hoeveelheid uitgedrukt als afwijking van de behandelde sloten met de controle-sloten. Net als in het vorige plaatje zien we dat de grootste afwijking zich voordoet bij de sloten met de hoogste dosis. De soorten staan nu langs een extra lijn (rechts). De afname van de soorten die hier bovenaan staan is het grootst en, let op het nulpunt, er is één soort (de slak) die toeneemt. In tegenstelling tot de vorige methode, levert de PRC methode wel overzichtelijke resultaten met andere gegevensbestanden. De PRC methode is inmiddels een wereldwijde standaard bij analyse van dergelijke proeven. De methode heeft ook potentie daarbuiten. Het is een algemene methode voor de grafische weergave van een hoofdeffect met interacties. Er ontbreekt ook nog iets, namelijk een aanduiding van de onzekerheid in de getoonde curven. Recent hebben Marieke Timmerman van Rijksuniversiteit Groningen en ik een bootstrap methode ontwikkeld waarmee de onzekerheid in

de curven kan worden berekend (Timmerman and ter Braak 2006). De zoektocht die ik zojuist beschreven heb, was een zoektocht naar eenvoud in het resultaat.

Van rekenregel naar statistisch model en terug

De zoektocht naar eenvoud is ook vaak een omweg naar eenvoud. Ik wil dat illustreren aan de hand van een voorbeeld. De eminente Duitse plantenecoloog Heinz Ellenberg (1913-1997) heeft zijn enorme veldkennis toegankelijk gemaakt voor anderen door een lijst van indicatiegetallen op te stellen. Het is een lijst van plantensoorten, waarbij hij aan elke plantensoort een getal voor vocht toekende op een schaal van 1 tot 12, waarbij 1 duidde op "extreem droog" en 12 op "extreem nat" (een onderwaterplant). Zo kende hij ook getallen toe voor bijvoorbeeld voedselrijkdom en zuurgraad. Hij schreef niet waar die getallen nu precies voor stonden, maar wél hoe de lijst gebruikt kon worden om de milieuomstandigheden van een plek te karakteriseren. Het voorschrift luidde: bepaal welke soorten er voorkomen, zoek voor elke soort het indicatiegetal op in de tabel en middel de zo gevonden indicatiegetallen. Wat moet je nu met zo'n voorschrift? Er zijn formele bezwaren tegen aan te voeren. Bijvoorbeeld, de indicatiegetallen zijn ordinaal en ordinale getallen 'mag' je niet zomaar middelen. Ik wist een doorbraak in deze discussie te bereiken door het probleem om te draaien. Neem de rekenregel (indicatiegetallen middelen) als uitgangspunt en onderzoek of er een model is waaronder deze rekenregel de best mogelijke is. Zo'n model is er, het is een 'species packing model' met Gaussische responscurven (ter Braak and Barendregt 1986). Collega Marten Scheffer heeft recent aannemelijk gemaakt dat dit model lang niet zo vanzelfsprekend is als tot voor kort werd aangenomen

(Scheffer and van Nes 2006), maar daar gaat het nu even niet over. Met een expliciet statistisch model hebben we wat houvast. Het geeft aan dat er op zijn minst één situatie is waarbij de rekenregel de beste is, dat de regel dus zo gek niet is. Voor wat betreft die ordinale schaal zegt de nieuwe theorie dat deze schaal prima is als de responscurven er op die schaal tenminste redelijk symmetrisch uitzien. Niemand heeft sindsdien behoefte gevoeld de schaal te transformeren om een grotere mate van symmetrie te verkrijgen. We kunnen nu ook gericht zoeken naar de statistische eigenschappen van de rekenregel. We kunnen ook proberen het model uit te breiden naar andere situaties, bijvoorbeeld naar de situatie dat we wel gegevens hebben over milieukenmerken maar geen of weinig kennis over de ecologische voorkeur van de soorten. Dit statistische model vormt de kern van mijn 'Theory of gradient analysis' (ter Braak and Prentice 1988) met als belangrijkste nieuwe techniek canonische correspondentie-analyse (ter Braak 1986). Dit werk uit 1986 vormt de basis van mijn computerpakket CANOCO voor multivariate analyse van ecologische gegevens. We vieren vandaag ook een beetje de twintigste verjaardag van CANOCO. Het computerpakket is in die twintig jaar door duizenden ecologen aangeschaft en met succes gebruikt voor het inzichtelijk maken van verbanden tussen het voorkomen van soorten en milieukenmerken.

De wetenschappelijke zoektocht die leidde tot CANOCO begon bij iets eenvoudigs, namelijk het middelen van indicatiegetallen, komt via 'harde wiskunde' (dat is een stukje van de omweg) naar iets nieuws, namelijk een statistisch model. Dat model is op zich ook weer eenvoudig. Het is een eenvoud op een ander niveau. Het Gaussische model past binnen een algemene klasse van modellen, de generaliseerde lineaire modellen, en daar kan elke getrainde statisti-

cus mee uit de voeten. We hoeven dus niets nieuws te leren, het is gewoon "een speciaal geval". En hoe kwam ik tot canonische correspondentie-analyse? Via de omweg van het statistische model. Deze omweg gaf houvast bij de precieze formulering van deze methode. In de methode spelen gewichten een rol, en intuïtief is niet meer te begrijpen waarom die precies zo gekozen moeten worden. Die gewichten volgen gewoon uit het model, uit de theorie. Korte tijd later heeft ook een Franse groep canonische correspondentie-analyse ontdekt (Lebreton *et al.* 1988). In hun aanpak is het mijns inziens minder overtuigend waarom de keuze van gewichten precies zó moet.
De omweg die ik beschreven heb is de omweg van rekenregel naar statistisch model, in het voorbeeld van middelen van indicatiegetallen naar het Gaussische responsmodel. De weg terug is die van een statistisch model naar nieuwe rekenregels. In het voorbeeld is het de weg van het Gaussische responsmodel met gegevens over het voorkomen van soorten en de waarden van milieukenmerken naar een rekenregel, een algoritme voor canonische correspondentie-analyse. Ik heb het nut van deze omweg proberen te schetsen. Een vervolgstap in deze lijn van onderzoek is het integreren van de gegevens over het voorkomen van soorten, de kenmerken van die soorten en de milieukenmerken van de plekken waar ze voorkomen (Dolédec *et al.* 1996).

Canonische correspondentie-analyse

Welke probleem heb ik met canonische correspondentie-analyse nu opgelost? Tot dan toe was het niet goed mogelijk het effect van een paar milieuvariabelen op een groot aantal soorten (dus op een levensgemeenschap) effectief in kaart te brengen als het aantal milieu kenmerken (p) plus het aantal

soorten (q) groter was dan het aantal monsters (n), waar die soorten en kenmerken gemeten waren. In canonische correspondentie-analyse mag het aantal soorten willekeurig groot zijn. Dat is een belangrijke vooruitgang omdat een typische dataset al gauw 20-200 soorten bevat. Daarnaast werd in de standaard multivariate statistiek uitgegaan van een rechtlijnig verband tussen soorten en milieuvariabelen. Dat in de ecologie onrealistische model is in canonische correspondentie- analyse vervangen door een eentoppig model, een model waarin elke soort zijn eigen niche[2] heeft.
Dit onderzoek heeft geleid tot betere multivariate statistiek voor het opsporen van verbanden in grote gegevensbestanden, het inschatten van effecten en risico's in een variabele wereld en het grafisch communiceren van multivariate kwantitatieve resultaten.
Met de technieken in CANOCO is het probleem van 'grote q (veel soorten) en kleine n (weinig objecten)' opgelost. Het probleem van veel milieukenmerken (p) en weinig objecten is jammer genoeg in CANOCO niet of maar zeer ten dele opgelost. Straks zal ik verder ingaan op dit 'grote p, kleine n' probleem.
De zoektocht die ik zojuist beschreven heb, was een zoektocht naar eenvoud en effectiviteit van de statistische methode.

Manieren om een statistische methode te definiëren

Regelmatig komt iemand me vragen, 'Ik zit met dit probleem en ik heb zo eens wat geprobeerd. We doen nu dit, maar mijn collega zegt dat we het anders moeten aanpakken. Wat is nu de beste wijze om dit probleem op te lossen.' Empirisch succes is natuurlijk een vereiste, maar vaak onbekend op het moment dat zo'n vraag gesteld wordt.

Bovendien is alleen empirisch succes een beetje mager. Je kunt niet alle type datasets verzinnen waarop de rekenregel zal worden toegepast. Doet de regel het wel net zo goed op een nieuw voorbeeld? Graag zou je willen begrijpen waarom de rekenregel het goed doet. Misschien is er wel een betere. Dit brengt mij er toe expliciet te maken dat er verschillende manieren zijn om een statistische techniek te definiëren namelijk via een

- beslissingstheoretisch model
- statistisch model
- doelfunctie (criterium)
- rekenregel of rekenregels (een algoritme)

Voorbeelden van methoden die alleen via rekenregels zijn gedefinieerd zijn Partial Least Squares (PLS), Agglomeratieve cluster analyse en Self Organizing Maps. Methoden die via een doelfunctie zijn gedefinieerd zijn bijvoorbeeld multidimensional scaling en penalized regression. In de cultuur van statistici hebben technieken een duidelijke pikorde. In die pikorde staat een techniek die alleen rekenregel is, lager in aanzien dan een techniek die afgeleid is van een doelfunctie en die technieken staan weer lager in aanzien dan de technieken die gedefinieerd zijn door een statistisch model. Het neusje van de zalm zijn dan die technieken die afgeleid zijn van een realistisch beslissingstheoretische model, waar alle kosten en baten van acties en beslissingen die genomen kunnen worden in onder zijn gebracht. Zo wordt PLS, een veel gebruikte rekenregel uit de chemometrie, vaak minder geacht dan ridge regressie, omdat ridge regressie een eenvoudige doelfunctie heeft (die ook nog volgt uit een Bayesiaans statistisch model), en PLS niet. Sijmen de Jong en ik zijn hard op zoek gegaan naar DE doelfunctie van PLS. We zijn een heel eind gekomen

(ter Braak and de Jong 1998), maar aan onze doelfunctie kleven toch nog wat schoonheidsfoutjes. Dit geringere aanzien is één van de oorzaken dat een bijzonder effectieve techniek als PLS maar langzaam doordringt in nieuwe gebieden als de bioinformatica (maar zie Boulesteix en Strimmer (2006)). Gelukkig is de cultuur aan het veranderen. Recent heeft Jerome Friedman een klasse van regressietechnieken gedefinieerd middels een rekenregel. Er is niet direct een doelfunctie of statistisch model! Net als bij PLS is de maatstaf voor succes de empirische voorspelkracht van de regel. Deze empirische voorspelkracht wordt berekend via kruisvalidatie. Je zou het zo kunnen zeggen. Als je slim genoeg bent een effectieve rekenregel te verzinnen, dan heb je de steun van het expliciete statistische model niet nodig. Ik wil graag nog benadrukken dat deze cultuuromslag mogelijk is geworden door onderzoek naar herbemonsteringsmethoden (Efron and Tibshirani 1993) als permutatie, bootstrap en kruisvalidatie. Dit zijn rekenintensieve methoden waarmee statistische significantie, standaardfouten en voorspelkracht kunnen worden berekend op basis van de combinatie van rekenregel en gegevens, zonder tussenkomst van een traditioneel statistisch model.

Herbemonsteringsmethoden hebben bovendien de charme van de eenvoud. Zo maken we in de naïeve bootstap een groot aantal maal een 'nieuw bestand van n monsters' door n keer aselect een monster uit het originele bestand te trekken. Sommige monsters zullen meer dan één keer in het nieuwe bestand zitten en andere zullen er niet in voorkomen. De manier waarop we via herbemonstering uit een gegevensbestand 'nieuwe bestanden' maken, laat direct zien wat de aannames zijn waarop de statistische analyse is gebaseerd. Omdat ze ook nog eens breed toepasbaar zijn, verdienen herbemonsteringsmethoden een belangrijke plaats

in het onderwijs. Ze helpen duidelijk maken wat de variatie in de uitkomst is.

Het grote *p*, grote *q*, kleine *n* probleem

Dan wil ik nu graag schetsen op welke gebieden mijn leerstoel de zoektocht naar eenvoud wil voortzetten. Mijn belangrijkste toepassingsgebieden zullen zijn statistische ecologie en statistische genomica.
Wat vroeger 'groot' betekende in de ecologie en milieuwetenschappen, zeg 20-200 soorten, is al lang niet meer als 'groot' aan te duiden in het nieuwe vakgebied genomics en aanverwante -omics gebieden. Met het Centrum voor Biosystems Genomics (CBSG) onder leiding van Willem Stiekema is dit nieuwe vakgebied goed vertegenwoordigd in Wageningen. Daarnaast participeert Wageningen-UR in een aantal andere centra van het Netherlands Genome Initiative. In dat nieuwe vakgebied wordt de expressie van vele genen – honderden tot duizenden- in kaart gebracht in verschillende organen, in verschillende groeistadia, onder verschillende groeiomstandigheden en, niet te vergeten, van planten met verschillend genotype. Interesse gaat ondermeer uit naar waardevolle inhoudsstoffen, zoals smaakmakers, geurstoffen of stoffen die gezondheidsbevorderend kunnen zijn. Daar is veel kennis van de stofwisseling in de plant voor nodig. Daarom wordt er op grote schaal en automatisch gemeten aan de metabolieten in de cel, en daar zijn er, met name in planten, ook duizenden van. De revolutie in de onderzoeksmethoden in de genomica heeft vergaande gevolgen voor de statistiek. Wat dat betreft herhaalt de geschiedenis zich. Nieuwe natuurkunde noopt tot nieuwe ontwikkelingen in de wiskunde en zo helpt de wiskunde de natuurkun-

de verder. Evenzo had de landbouw grote invloed op de statistiek en daarmee kreeg de statistiek grote invloed op hoe proeven werden opgezet en geanalyseerd. Nu is het de beurt aan de genomica en komt er een sterke wisselwerking tussen statistiek, bioinformatica en genomica.

Alleen al het uitrekenen van alle paarsgewijze correlaties tussen 10.000 genen is niet iets wat je zomaar doet. Het zijn er een slordige 50 miljoen, en dan rijst natuurlijk de vraag: wat zie je daar nog aan ?

De invloed van de genomica op de hoofdstroom van de statistiek is nog van recente duur. Zo schreven Leo Breiman en Jerome Friedman in 1997 in een discussiepaper voor de Royal Statistical Society naar aanleiding van mijn opmerking dat hun daar geïntroduceerde nieuwe multivariate regressie methode instabiel is als $q>n$ of $p+q > n$ (veel responsvariabelen en voorspellende variabelen en weinig monsters): 'Although this case seems somewhat unusual, some comments can be made.' (Breiman and Friedman 1997). Er wordt inmiddels internationaal hard gewerkt aan oplossingen voor wat kortheidshalve het 'grote p, kleine n' probleem heet. Het leidt bijvoorbeeld tot andere vormen van asymptotiek (Hall *et al.* 2005), vormen die vroeger niet eens serieus genomen zouden worden.

Wat is het grote p, kleine n probleem? Het is het probleem dat de hoogdimensionele ruimte héél erg leeg is. Neem een voetbalveld. Als er een vrije trap is net buiten het 16 meter gebied, dan staan alle tien spelers van de verdedigende partij bijna op één lijn elkaar te verdringen. De lengte van de lijn is, zeg maar voor het gemak, 10 meter. Na de vrije trap verspreiden ze zich. Als ze zich zouden verspreiden over een gebied van 10 bij 10 meter, hebben ze al veel meer ruimte; de dichtheid aan spelers is eerst 1 speler per meter en daarna ééntiende speler per vierkante meter. Zouden ze elk ook

nog tot 10 meter in de lucht kunnen springen dan is de dichtheid al éénhonderdste speler per cubieke meter. Daar kan een bal veel makkelijker doorheen! Op de lijn is er altijd tenminste één speler die bij de bal kan zonder zich te verplaatsen, in het 10 bij 10 meter vlak is dat al de vraag en in drie dimensies kan niemand meer bij de bal. De dichtheid aan spelers neemt dus schrikbarend af naarmate we het spel spelen in meer dimensies. Dat is één uiting van het grote p, kleine n probleem. Met 10.000 kenmerken hebben we het over een 10.000-dimensionale ruimte, die dus extreem leeg is. Het herkennen van structuren is dan heel lastig.

Stel we willen weten of een nieuw medicijn wel of niet zal aanslaan bij een nieuwe patiënt op basis van het expressiepatroon van 5000 genen. Zo'n patroon valt tegenwoordig snel en goedkoop vast te stellen. Op basis van het expressiepatroon en het wel of niet aanslaan van het medicijn bij 100 reeds behandelde patiënten hopen we een regel te kunnen opstellen die goed voorspelt of het medicijn zal aanslaan. Hier is p = 5000 en n = 100. Het probleem van schijncorrelaties - correlaties die alleen maar op toeval berusten - is hier levensgroot. Hoe vind ik de genen die er werkelijk toe doen, hoe maak ik daar een voorspelregel van, en wat is de kwaliteit van mijn voorspelling? Het is als het vinden van een speld in een hooiberg (Johnstone and Silverman 2004).

Ik heb hopelijk al duidelijk gemaakt dat het grote p, kleine n probleem ook voor Wageningen UR belangrijk en uitdagend is. Ik hoop en verwacht dat mijn leerstoel op een originele manier kan bijdragen met een nieuwe Bayesiaanse aanpak. Het werkpaard van veel mensen die aan dit probleem werken is de Lasso. Mooie nieuwe namen zijn door de invloed van John Tukey mode geworden in de statistiek. Lasso is een regressie-methode waarbij de som van de abso-

lute waarden van de regressie-coëfficiënten bestraft wordt en zo in toom gehouden wordt. Vandaar, de naam. Ik heb dit jaar een nieuwe methode gepubliceerd (Bayesian sigmoid shrinkage) die de lasso duidelijk verslaat in een bepaalde toepassing (ter Braak 2006). Eerlijkheidshalve moet ik erbij vermelden dat mijn methode niet de enige nieuwe methode is die de lasso verslaat, maar mijn vondst blinkt uit in eenvoud en snelheid. Bovendien staat de toepassing (wavelet denoising) nog veraf van waar ik uiteindelijk wil uitkomen: een effectieve regressietechniek voor data analyse in de statistische ecologie en statistische genomica. Deze zoektocht begon overigens in 2002 toen ik samen met Martin Boer en Ritsert Jansen probeerde interacties tussen genen op te sporen (Boer *et al.* 2002). Door in dit veld actief te zijn, kunnen we veelbelovende nieuwe methoden snel beoordelen en gebruik van goede methoden propageren in het onderzoek en onderwijs van Wageningen UR.

Ik ging in het voorgaande stilzwijgend uit van een standaardklasse van modellen in de statistiek, de generalized linear models (GLM). Genomics vraagt ook om modellen die niet zo standaard zijn en ik verwacht dat nieuwe vragen zullen leiden tot nieuwe modellen. Onze aandacht gaat onder meer uit naar toepassingen van Bayesiaanse netwerk modellen. Ik wil hierbij de samenwerking met bioinformatica verder uitbouwen. De projecten van de promovendi Yiannis Kourmpetis en Anand Gavai zijn nog maar het begin!

Dan wil ik u nu graag meenemen naar een situatie waar het veelvoud van eenvoud tot uitdagende complexiteit leidt. Ik doel op modelbouw.

Statistische aspecten van modelbouw

De stelling 'Statistiek is het leren uit gegevens' legt de nadruk wel erg op gegevens. Waar blijft de kennis die we al hebben over een systeem? Wageningen UR is door C.T. de Wit beroemd geworden met gewasgroeimodellen. Niet voor niets is er een onderzoekschool naar hem genoemd. Deze modellen proberen de groei van een gewas te verklaren op basis van een groot aantal op zich eenvoudige deelprocessen zoals hoeveel licht door de bladeren van een plant kan worden opgevangen, en hoeveel energie dat levert via de fotosynthese (van Ittersum *et al.* 2003, Yin and van Laar 2005). De uitkomst van een deelproces wordt bepaald door inputwaarden die deels in andere deelprocessen zijn berekend, door externe variabelen zoals het weersverloop gedurende het groeiseizoen en door de parameters van het proces, ook wel de modelparameters genoemd. Daaronder vallen bijvoorbeeld reaktieconstantes die de snelheid van chemische en enzymatische reacties bepalen. Op detailniveau bevat zo'n model allerlei empirische verbanden, waarvan sommige een goede en andere een mindere goede theoretische onderbouwing hebben. We weten dus wel iets van die parameters maar met onzekerheid. Vaak zijn er ook nog parameters waarvan we alleen de orde van grootte weten en tenslotte kunnen er parameters zijn die van geval tot geval iets anders kunnen zijn. De ene tarwevariëteit groeit tenslotte anders dan de andere. Het is duidelijk dat met zoveel onzekerheid statistiek een belangrijke rol speelt in de modelbouw. Modelbouw speelt ook een grote rol bij Systems Biology.

Een ander voorbeeld. Het Milieu- en Natuurplanbureau doet 'onafhankelijke evaluaties en verkenningen naar de kwaliteit van de fysieke leefomgeving en de invloed daarvan

op mens, plant en dier' (www.mnp.nl). De evaluaties, prognoses en verkenningen zijn allemaal gebaseerd op complexe modellen die bestaan uit op elkaar ingrijpende submodellen. Kwaliteitsborging van data en modellen is van groot maatschappelijk belang. Wiskunde en statistiek spelen daarbij een belangrijke rol. Statistiek draagt bij in de vorm van methoden voor onzekerheidsanalyse, gevoeligheidsanalyse en modelkalibratie. Michiel Jansen van Biometris heeft dit vakgebied helpen ontwikkelen en mede op basis van zijn kennis en ervaring geeft mijn DLO-collega Saskia Burgers nu de cursus 'Onzekerheids- en gevoeligheidsanalyse voor modelbouwers'. Ik wil dit vakgebied verder uitbouwen en daarbij in eerste instantie vooral aandacht geven aan modelkalibratie.

Wat is modelkalibratie? Ik geef een voorbeeld. Bij een gewasgroeimodel zou je gegevens kunnen hebben over de opbrengst van het gewas in 2006 op een zandgrondperceel. Als je echter het gewasgroeimodel draait op de computer met de dagelijkse temperatuur en zonneschijn in dat jaar, geeft het model een veel lagere opbrengst. Wat is nu de prognose voor de opbrengst in 2007? Houden we het op de te lage prognose die uit het model komt of verhogen we de prognose op basis van de gerealiseerde opbrengst in 2006? Een mogelijkheid is om de modelparameters iets aan te passen voor dit perceel en de variëteit die hier geteeld wordt. De modelparameters zijn als het ware knoppen waar je aan kunt draaien om een uitkomst te krijgen die beter past bij de gegevens. Dit is modelkalibratie, en wel modelkalibratie op zijn slechtst. Waarom op zijn slechtst? Omdat er vele mogelijke instellingen zijn van de modelparameters die als modeluitkomst allemaal precies de waargenomen opbrengst leveren, maar die onder *andere* temperatuur- en zonneschijnscenario's totaal verschillende uitkomsten zullen ge-

ven. Deze vorm van modelkalibratie is dan ook terecht verguisd. Ja, verguisd, maar daarom nog niet minder toegepast omdat het in de praktijk vaak een onmisbare stap is voor het verkrijgen van een in een beleidsstudie bruikbaar model. Dit kalibreren kan veel beter en op een wetenschappelijk verantwoorde manier binnen het Bayesiaanse modelraamwerk van Jansen en Hagenaars (2004). Een vereiste daarbij is dat we voorafgaand aan de kalibratie de onzekerheid in de modelparameters kwantificeren. Een stap in de goede richting is bijvoorbeeld het artikel van Marcel van Oijen in Tree Physiology (van Oijen *et al.* 2005). Terzijde merk ik op dat de modelkalibratie een ander voorbeeld is van het 'grote p, kleine n' probleem, omdat een model doorgaans veel onzekere parameters heeft en de modelkalibratie moet gebeuren op basis van maar enkele gegevens. Andermaal kan de Bayesiaanse statistiek uitkomst bieden. Het is de hoogste tijd om aandacht te geven aan Bayesiaanse statistiek.

Bayesiaanse statistiek

Ik stelde: "Statistiek is het leren uit gegevens". Bayesiaanse statistiek neemt die stelling heel serieus. Bayesiaanse statistiek gaat ervan uit dat we al wat geleerd hebben en nu willen dóórleren op basis van nieuwe feiten en gegevens. Daarentegen begint de klassieke statistiek, gechargeerd gezegd, alsmaar bij 'Af', alsof we nog niets weten als nieuwe gegevens beschikbaar komen. Bayesiaanse statistiek is genoemd naar de regel van Bayes, een stelling in de waarschijnlijkheidsrekening die bedacht is door de 18de eeuwse Engelse predikant Thomas Bayes. Het is een eenvoudige regel met

Fig. 3. Regel van Bayes voor normale verdelingen: de verdeling van de opbrengst volgens de huidige kennis (a priori verdeling, gestreepte lijn, links) en die van de nieuwe gegevens (likelihood, stippellijn, rechts) geven met de regel van Bayes een aangepaste verdeling voor de nieuwe kennis (*a posteriori* verdeling, doorgetrokken lijn, midden).

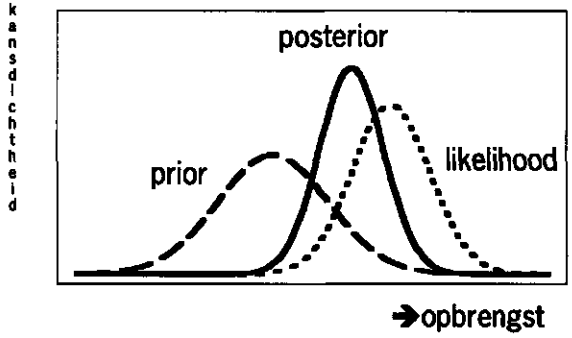
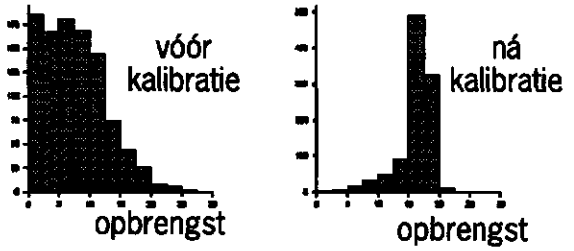


Fig. 4. Histogram van de *a priori* verdeling voor de voorspelde maïs opbrengst in kg/ha volgens het model (gebaseerd op onzekerheidsanalyse, links) en van de *a posteriori* verdeling van de voorspelling voor 2007 na kalibratie van het model voor een specifieke situatie op basis van de opbrengstgegevens uit 2006 (geactualiseerd naar Jansen en Hagenaars, 2004).



verreikende consequenties. Ik zal de regel van Bayes in woorden samenvatten:

"Kennis plus[3] nieuwe gegevens geeft nieuwe kennis"

De Bayesiaanse statistiek gaat over het aanpassen van onze kennis op basis van nieuwe gegevens. Kennis en gegevens worden hierbij beide weergegeven door kansverdelingen. Deze tak van statistiek kwam op in de vijftiger jaren van de vorige eeuw, maar leidde een sluimerend bestaan. Vele statistici moesten er niets van hebben omdat het niet 'objectief' was. Het probleem is vaak namelijk 'hoe kwantificeer je de kennis die je al hebt? ' Dat geeft een verafschuwd subjectief element in de statistische analyse. Er was een tweede reden dat Bayesiaanse statistiek niet van de grond kwam. Je kon er eigenlijk niets 'meer' mee dan met klassieke statistiek.

Hier is een leerboekvoorbeeld waarin zowel de kennis als de gegevens kunnen worden weergegeven door normale verdelingen (Fig. 3). De nieuwe kennis is dan ook een normale verdeling die een compromis is tussen de voorkennis en de gegevens. Ik geef nu een voorbeeld van modelkalibratie volgens de regel van Bayes (naar Jansen en Hagenaars, 2004). Links (Fig. 4) staat de voorspelde verdeling van opbrengst voor 2007 voor een willekeurig perceel. De verdeling volgt uit een onzekerheidanalyse van een gewasgroeimodel. De percelen van boer Jansen gaven relatief hoge opbrengsten in 2006. Rechts staat de verdeling van de voorspelde opbrengst voor zijn percelen voor 2007 nadat we het gewasgroeimodel hebben gekalibreerd op basis van de opbrengstgegevens uit 2006. U ziet welke winst in nauwkeurigheid we kunnen bereiken.

In toepassingen hebben we vaak te maken met op zich al rekenintensieve modellen met veel verschillende inputvariabelen en modelparameters. Dan is de Bayesiaanse analyse in theorie wel eenvoudig maar in de praktijk bijzonder reken-

intensief. De werkelijke doorbraak van de Bayesiaanse statistiek valt dan ook samen met de computerrevolutie. We hebben nog nooit zoveel rekenkracht tot onze beschikking gehad en die rekenkracht kunnen we goed gebruiken. Het is onmogelijk de nieuwe verdeling exact te berekenen. Daarom gaan we de verdeling simuleren. De verdeling wordt daarmee weergegeven door de verzameling van trekkingen uit die verdeling. Een multivariate verdeling van, zeg, 1000 variabelen wordt daarmee een tabel met 1000 kolommen en evenzoveel rijen als trekkingen. Maar hoe verricht je die trekking? Welke notaris kan dat onpartijdig doen? Als het probleem veel ingewikkelder is dan het trekken van een aantal winnaars uit alle inzendingen, hebben we daarvoor Markov Chain Monte Carlo methoden, zoals het Metropolis-Hastings algoritme. Dat algoritme is al in 1953 bedacht door Metropolis en collegae voor simulaties in de fysica (bij de ontwikkeling van de atoombom) en in 1970 veralgemeniseerd door Hastings. U ziet hier hoe lang het kan duren voor ontdekkingen hun nut bewijzen.
Stel, u wilt het beste beleid voeren (wie niet?) en u kunt daarvoor kiezen uit een combinatie van maatregelen. Idealiter hebt u een model, een computermodel, waarmee u de effecten van de maatregelen kunt doorrekenen. Kort door de bocht, voor elk combinatie van maatregelen rekent het model uit hoe goed het beleid is, bijvoorbeeld hoe veel extra geld het nieuwe beleid oplevert. In een dergelijke situatie kan een wiskundige optimalisatiemethode u de beste combinatie van maatregelen geven (tenminste als de methode niet blijft steken in een lokaal maximum). Maar wat gebeurt er nu als het model onzeker is? Dan wilt u toch zeker ook die onzekerheid verdisconteerd zien in de uitkomsten van het model. Dat kan nu precies met het Metropolis-Hastings algoritme. Het algoritme is te beschouwen als een

optimalisatie algoritme (zoals bijvoorbeeld simulated annealing) dat tevens de onzekerheid in de uitkomst laat zien.
Het Metropolis-Hastings algoritme is meer een richtlijn dan een rekenregel. Een concrete implementatie van de methode kan erg inefficiënt zijn. Er is dus nog steeds onderzoek nodig naar effectieve implementaties. Bij toeval heb ik er zelf ook één gevonden (ter Braak 2006). Het toeval betreft hier dat ik besloot deel te nemen aan een ééndagscursus 'Genetic algorithms with animal breeding applications' die de Wageningse onderzoekschool WIAS organiseerde. Julius van der Werf leerde me daar wat 'Differential Evolution' was en die avond nog was de essentie van de theorie achter 'Differential Evolution Markov Chain' rond. Toen was het nog drie maanden hard werken om te laten zien dat het ook echt werkt. Leve een goed onderzoeks- en onderwijsklimaat in Wageningen, waaraan ik met mijn leerstoel ook hoop bij te dragen. De methode is een toonbeeld van eenvoud en effectiviteit. De rekenregel is van dezelfde eenvoud als 'indicatiegetallen middelen'; de toepassingsmogelijkheden zijn vele malen groter.
Met deze ontdekking heb ik een 17 jaar oud probleem opgelost dat aan Michiel Jansen werd voorgelegd in het kader van kalibratie van een computermodel voor koolstofstromen in de Oosterschelde (Klepper 1989). Olivier Klepper dacht betrouwbaarheidsintervallen voor zijn modelparameters te kunnen verkrijgen op basis van het Price-algoritme (Price 1979), een bepaald optimalisatie-algoritme dat gebruikt wordt voor modelkalibatie. In een interne notitie liet Michiel Jansen netjes zien dat de claim ongegrond was. Daarna hebben Eligius Hendrix en Olivier Klepper aan alternatieven gewerkt (Klepper and Hendrix 1994). Met de truc die ik bedacht heb, kan ik het Price-algoritme zo aanpassen dat Kleppers' doel wel gehaald kan worden. Het

blijkt overigens ook dat in deze context Differential Evolution veel beter werkt dan het Price-algoritme.
Eén van de makers van WinBugs, een algemeen toepasbaar computer programma voor Bayesiaanse statistiek, schreef me recent in een e-mail 'The more I try the algorithm the more impressed I become!' en 'It works comparably well to Gibbs sampling but is so much simpler'. U begrijpt dat ik zijn woorden hier aanhaal omdat ik verder nog weinig concreets heb om het succes en de reikwijdte van de nieuwe methode te laten zien. Deze onderzoekslijn is in mijn visie erg vruchtbaar en heeft potentieel brede en grote impact, ook buiten de modelkalibratie.
Ik noem hierbij mijn postdoc Marc Rutten die momenteel onderzoekt in hoeverre Differential Evolution Markov Chain kan leiden tot snellere en effectievere Bayesiaanse algoritmes om QTLs op te sporen. QTLs zijn gebieden op een chromosoom, die een kenmerk van het individu beïnvloeden. Dergelijke QTLs kunnen gebruikt worden om betere veredelingsprogramma's op te stellen. Biometris is vanouds her sterk in QTL-analyse en die positie willen we graag behouden. Andere toepassingsgebieden zijn de farmacokinetica en farmacodynamica en meer in het algemeen de gegeneraliseerde gemengde niet-lineaire modellen.
Ik heb u iets van de zoektocht naar eenvoud laten zien. Deze leidt tot betere multivariate statistiek en daarmee tot betere levenswetenschappen.

Dankwoord
Graag wil ik afsluiten met een woord van dank. In de eerste plaats dank ik de Raad van Bestuur, het College voor Promoties en de leden van de toetsingscommissie persoonlijke hoogleraren, de directie van kenniseenheid Plant en het managementteam van Biometris voor het in mij gestel-

de vertrouwen. De benoeming tot persoonlijk hoogleraar ervaar ik als waardering voor mijn werk en werkwijze en geeft me een extra stimulans dóór te gaan. Ik blijf in dienst bij de stichting DLO. Als persoonlijk hoogleraar kan ik de kennis die ik heb opgebouwd op het gebied van de multivariate statistiek beter uitdragen binnen Wageningen UR en zo bijdragen aan de kwaliteitsborging van het onderzoek. Ik kan nu zelf AIO's aanvragen bij NWO en zo mijn visie op statistiek en onderzoek overdragen aan jonge mensen. In de tweede plaats wil ik graag mijn vroegere bazen bedanken. Jos Jansen, Kit Roes, Peter Finke en Gerie van der Heijden, jullie hebben het mogelijk gemaakt dat ik mijn wetenschappelijke ambitie kon volgen naast (en binnen) al het directe consultatiewerk. Wetenschapsbeoefening heb ik geleerd van Colin Prentice. Zonder hem zou het CANOCO-project nooit iets geworden zijn. John Birks heeft me ook altijd op bijzondere wijze gestimuleerd en geholpen. De populariteit van CANOCO is mede de verdienste van Petr Smilauer, die nu co-auteur is, Colin Prentice, John Birks en Paul van den Brink. Ik reisde zelf weinig; zij gaven als ware ambassadeurs overal in de wereld voordrachten en cursussen.

De vele onderzoekers bij Alterra en omgevingswetenschappen met wie ik onderzoek heb mogen doen, ik noem in het bijzonder professor Paul Opdam, Herman van Dam, Clair Vos, Piet Verdonschot, Han van Dobben en André Schaffers, jullie onderzoek en vragen hebben me bijzonder geïnspireerd. Zoals ik heb gezegd, statistische ecologie blijft een belangrijk aandachtsgebied van me. Met de hooggeleerden Oenema en Schaminée heb ik onlangs nieuwe samenwerking afgesproken. Hooggeleerde Leunissen, beste Jack en beste Roeland van Ham, onze samenwerking op het gebied van de genomica en bioinformatica is nog pril. Ik

hoop en verwacht dat ze mooie vruchten zal afwerpen. Hooggeleerden Grasman en Stein, beste Johan en Alfred, jullie hebben ervoor gezorgd dat de leerstoelgroep toegepaste wiskunde en statistiek zo goed presteerde dat Biometris nu weer twee 'full profs' heeft. Hooggeleerde van Eeuwijk, beste Fred, ik ben blij dat jíj de nieuwe reguliere hoogleraar toegepaste statistiek ben geworden. Ik denk dat we plezierig met elkaar zullen samenwerken en elkaar goed zullen aanvullen. De levenswetenschappen in Wageningen zullen er profijt van hebben. Hooggeleerde Molenaar, beste Jaap, als nieuwe reguliere hoogleraar toegepaste wiskunde speel je een belangrijke rol in het vormgeven van Systems Biology in Wageningen. Ik hoop daar ook aan bij te kunnen dragen.
Beste collegae bij Biometris, ik voel me thuis bij Biometris en dat komt door jullie. Jullie bijdrage aan mijn werk is veel groter dan tot uiting komt in co-auteurschappen en dankwoorden bij artikelen. Graag had ik ook júllie succesvolle toepassingen van statistiek in de levenswetenschappen laten zien! Het is nu een heel persoonlijke zoektocht geworden. Vanwege een combinatie van privé- en werkomstandigheden ben ik er rond de eeuwwisseling een jaar tussenuit geweest. Toen heb ik geleerd hoe belangrijk goed Bedrijfsmaatschappelijk werk is. Renée Hoevenaar coachte me door een moeilijke periode, waarvoor heel veel dank. Het is niet altijd gemakkelijk een gedreven wetenschappelijke onderzoeker te zijn in een verzakelijkte organisatie.
Tot slot dank ik Helmi voor wie creativiteit en eenvoud vanzelfsprekend zijn en met wie het leven iedere dag weer een plezier is.
Ik dank u allen voor uw aandacht.

Ik heb gezegd.

Referenties

Boer, M. P., ter Braak, C. J. F. and Jansen, R. C., 2002. A penalized likelihood method for mapping epistatic quantitative trait loci with one-dimensional genome searches. Genetics, 162, 951-960.

Boulesteix, A.-L. and Strimmer, K., 2006. Partial least squares: A versatile tool for the analysis of high-dimensional genomic data. Briefings in Bioinformatics, online bbl016.

Breiman, L. and Friedman, J. H., 1997. Predicting multivariate responses in multiple linear regression. J.R. Statist. Soc. B., 59, 3-54.

Dolédec, S., Chessel, D., ter Braak, C. J. F. and Champely, S., 1996. Matching species traits to environmental variables: A new three-table ordination method. Environmental and Ecological Statistics, 3, 143-166.

Efron, B. and Tibshirani, R. J., 1993. An introduction to the bootstrap. Chapman & Hall, London.

Hall, P., Marron, J. S. and Neeman, A., 2005. Geometric representation of high dimension, low sample size data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67, 427-444.

Gabriel, K. R., 1971. The biplot graphic display of matrices with application to principal component analysis. Biometrika, 3, 453-467.

Gower, J. C. and Hand, D. J., 1996. Biplots. Chapman & Hall, London.

Hastie, T., Tibshirani, R. and Friedman, J. H., 2001. The elements of statistical learning. Data mining, inference and prediction. Springer-Verlag, New York.

Johnstone, I. M. and Silverman, B. W., 2004. Needles

and straw in haystacks: Empirical bayes estimates of possibly sparse sequences. Annals of Statictics, 32, 1594-1649.

Jongman, R. H. G., ter Braak, C. J. F. and van Tongeren, O. F. R., 1995. Data analysis in community and landscape ecology. Cambridge University Press, Cambridge.

Klepper, O., 1989. A model of carbon flows in relation to macrobenthic food supply in the Oosterschelde estuary (S.W. Netherlands). PhD thesis, Agricultural University, Wageningen, 270.

Klepper, O. and Hendrix, E. M. T., 1994. A method for robust-calibration of ecological models under different types of uncertainty. Ecological Modelling, 74, 161-182.

Lebreton, J. D., Chessel, D., Prodon, R. and Yoccoz, N., 1988. L'analyse des relations especes-milieu par l'analyse canonique des correspondances. I. Variables de milieu quantitatives. Acta Oecologia Generalis, 9, 53-67.

Price, W. L., 1979. A controlled random search procedure for global optimisation. The Computer Journal, 20, 367-370.

Scheffer, M. and van Nes, E. H., 2006. Self-organized similarity, the evolutionary emergence of groups of similar species. PNAS, 103, 6230-6235.

ter Braak, C. J. F., 2006. A Markov Chain Monte Carlo version of the genetic algorithm differential evolution: Easy bayesian computing for real parameter spaces. Statistics and Computing, 16, 239-249.

ter Braak, C. J. F., 1986. Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis. Ecology, 67, 1167-1179.

ter Braak, C. J. F., 2006. Bayesian sigmoid shrinkage with improper variance priors and an application to wavelet denoising. Computational Statistics & Data Analysis, available online.

ter Braak, C. J. F. and de Jong, S., 1998. The objective function of partial least squares regression. Journal of Chemometrics, 12, 41-54.

ter Braak, C. J. F. and Prentice, I. C., 1988. A theory of gradient analysis. Advances in ecological research, 18, 271-317 (reprinted as Advances in ecological research *Classic Papers*, 34, 235-282).

Timmerman, M. E. and ter Braak, C. J. F., 2006. Bootstrap confidence intervals for principal response curves. submitted.

van den Brink, P. J., van Wijngaarden, R. P. A., Lucassen, W. G. H., Brock, T. C. M. and Leeuwangh, P., 1996. Effects of the insecticide dursban 4e (active ingredient chlorpyrifos) in outdoor experimental ditches: II. Invertebrate community responses and recovery. Environmental Toxicology and Chemistry, 15, 1143-1153.

van den Brink, P. J. and ter Braak, C. J. F., 1999. Principal response curves: Analysis of time-dependent multivariate responses of a biological commnuity to stress. Environmental Toxicology and Chemistry, 18, 138-148.

van Ittersum, M. K. *et al.*, 2003. On approaches and applications of the wageningen crop models. European Journal of Agronomy, 18, 201-234.

van Oijen, M., Rougier, J. and Smith, R., 2005. Bayesian calibration of process-based forests models: Bridging the gap between models and data. Tree Physiology, 25, 915-927.

Yin, X. and van Laar, H. H., 2005. Crop systems dynamics. An ecophysiological simulation model for genotype-by-environment interactions. Wageningen Academic Publsihers, Wageningen.

Noten

1 Deze omschrijving maakt duidelijk dat statistiek heel verwant is aan machine learning, artificial intelligence, patroonherkenning, data fusion en data mining. Vaak is de statistiek nog juist iets ambitieuzer dan de genoemde vakgebieden. De statistiek wil ook aangeven wat de onzekerheid in het geleerde is.

2 De niche is hier het gebied rond de top van de responscurve.

3 In de gebruikelijke formule staat een product van kansdichtheden. Dat wordt een optelling als we overgaan op de logaritme van de kansdichtheid.