geaccumuleerd of herberekend. Bereken voor nutriëntenconcentraties de gewenste waarden (bijvoorbeeld gemiddelde, minimum, maximum, voorjaarswaarde).
Bepaal aan de hand van de biotische variabelen het gewenste biologische waterkwaliteitsniveau (of het type levensgemeenschap) en codeer deze als een nominale (0/1) variabele. Zorg voor een handige codering die aansluit bij het doel van de analyse. Wanneer u bijvoorbeeld 5 kwaliteitsniveaus heeft, codeer dan 0 indien het heersende kwaliteitsniveau lager is dan het gewenste en 1 als het heersende kwaliteitsniveau gelijk of hoger is aan het gewenste.
Bereken voor beheersmaatregelen voor elk tijdstip in de meetreeks hoe lang geleden ze zijn uitgevoerd. Combineer de gegevens in één bestand. Voor de meeste (statistiek)programma's is gewenst dat de gegevens georganiseerd zijn per object, m.a.w. dat alle variabelen per monsterpunt per tijdstip in een rij staan. Een enkel programma verwacht de gegevens per variabele. Zorg ervoor dat ontbrekende waarden goed gecodeerd zijn; veel statistiekprogramma's verwachten een asterisk (*) of een waarde als bijvoorbeeld -9999 voor ontbrekende waarden, controleer dit van tevoren in de handleiding.

2.3.5 Eerst tekenen, dan rekenen

Maak eerst grafieken van de responsvariabele tegen de stuurvariabelen om enig idee te krijgen over de verbanden en over de spreiding van de gegevens. Let er vooral op, dat er geen uitbijters zijn in de stuurvariabele. Verwijder uitbijters en bestudeer nogmaals de grafiek. Als u geen relatie ziet, zal deze meestal ook niet worden gevonden door de toepassing van een model. Gebruik als u geen relatie ziet met bijvoorbeeld nutriënten verschillende symbolen voor monsterpunten met verschillend beheer en bekijk de grafiek nogmaals.

2.4 Uitvoering statistische analyse

Hieronder zijn een aantal algemene tips en opmerkingen vermeld die handig zijn bij het uitvoeren van de analyse. Voorbeelden van het uitvoeren van de analyses in Excel en SPSS kunnen worden gevonden in Hoofdstuk 3.

2.4.1 Het inlezen van de data

Lees de gegevens in het (statistiek)programma in en controleer of ze juist zijn ingelezen. Fouten met inlezen treden vooral op, wanneer het programma een strakke invoer vereist, maar ook wanneer gegevens in vrijere vorm worden ingelezen en gegevens ontbreken. Veel statistiekprogramma's lezen gegevens direct uit een spreadsheet. Ook in dit geval kunnen problemen optreden met lege cellen, die soms als waarde 0 worden geïnterpreteerd.

2.4.2 Logistische regressie

Gebruik de procedure voor logistische regressie. Sommige programma's, bijvoorbeeld SPSS en Statistix, hebben een aparte procedure, in andere programma's, bijvoorbeeld CSS-Statistica vindt u logistische regressie onder niet-lineaire modellen

(non-linear models), en tenslotte kunt u logistische regressie ook vinden onder de GLM's, gegeneraliseerde lineaire modellen (bijvoorbeeld in Genstat, SAS, maar niet in SPSS, waar GLM wordt gebruikt als General Linear Model, dus niet als Generalized Linear Model). In het laatste geval dient u de link-functie (logit) op te geven en de foutstructuur (binomiaal). Bij enkele programma's kunt u bovendien opgeven wat dient te worden geoptimaliseerd. Kies hiervoor altijd -2ln(likelihood) (zie paragraaf 2.1). Omdat de parameters worden geschat met een optimalisatieprocedure die soms niet tot een optimale oplossing komt, is het in een aantal programma's mogelijk startwaarden voor de parameters op te geven. Probeer hiervoor zinnige waarden te bedenken of laat het aan het programma over.

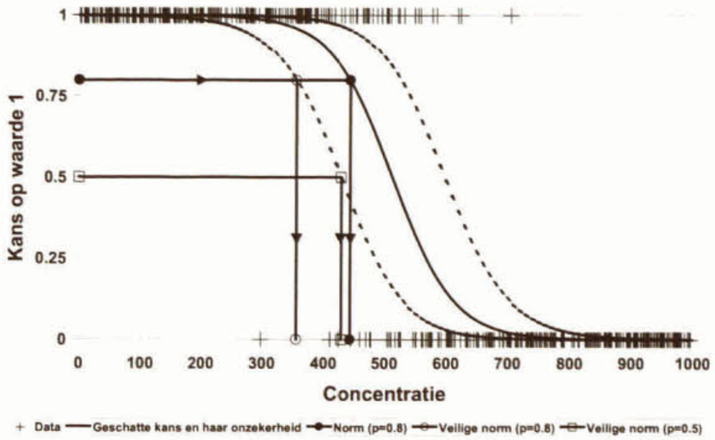
2.4.3 Modelspecificatie

Begin de analyse met het model als beschreven in paragraaf 2.1. Als dit model geen relatie lijkt op te leveren of als het duidelijk is dat andere factoren dan nutriëntconcentratie ook een rol spelen stap dan pas over naar een moeilijkere modelstructuur. In een gebaggerde diepe sloot bijvoorbeeld, is het effect van verhoging van de fosfaatconcentratie in het algemeen minder dan in een ondiepe sloot. Het is dus raadzaam om het diep of ondiep zijn van een sloot in de analyse mee te nemen als duidelijk is dat de dataset getallen afkomstig uit beide typen systemen bevat. Voor een beschrijving van deze wat meer gecompliceerde relaties kunnen in het regressiemodel naast de respons- en stuurvariabelen ook nog kwadratische en interactietermen toegevoegd worden. Kwadratische termen worden in de logistische regressie onder andere gebruikt om een optimumcurve (klokvormige curve) te modelleren (zie Bijlage 1 voor een voorbeeld). Interactietermen tussen twee nutriënten worden gebruikt om te beschrijven hoe het effect van verandering in concentratie van een van beide nutriënten veranderd wordt door verandering in concentratie van het andere nutriënt (zie Bijlage 2 voor een voorbeeld). In de hoofdtekst van dit rapport gaan we uit van een stuurvariabele (bijvoorbeeld nutriëntconcentratie) en een responsvariabele (bijvoorbeeld aan- of afwezigheid van macrofyten).

2.5 Interpretatie van het resultaat

We kunnen het resultaat op verschillende manieren interpreteren (zie figuur 5). Hoe we het interpreteren en welke regels worden gebruikt voor het vaststellen van een gedifferentieerde norm is in feite een beleidsmatige keuze. De verschillende keuzemogelijkheden worden in het onderstaande nader toegelicht.
Het uitgangspunt is dat de kans op het behalen van het van tevoren bepaald kwaliteitsniveau bij de normconcentratie minimaal een zekere waarde heeft.
De interpretatie is dan recht toe recht aan en direct uit de grafiek af te lezen.
We kiezen een bepaalde kans dat de gewenste biologische waterkwaliteit wordt gehaald, en lezen in de figuur af welke concentratie daarbij hoort. Met de gekozen kans wordt ook het geschatte risico op het niet behalen van de gewenste kwaliteit, hoewel de normconcentratie wordt behaald, bepaald. Het verband tussen de nutriëntconcentratie en de waterkwaliteit is echter geschat en dus is er ook nog het probleem van de fout in de schatting.

We kunnen rekening houden met de onzekerheid van de geschatte relatie als gevolg van deze fout; we kunnen deze onzekerheid ook gewoon accepteren en daarom in de normstelling negeren.



Figuur 5: Het aflezen van de norm, wanneer voor elk water minimaal een bepaalde kans op behalen basiskwaliteit gewenst is.

Keuzemogelijkheid I. Onzekerheid accepteren.
De meest aannemelijke waarde van de kans op behalen basiskwaliteit is weergegeven als de dikke lijn in de figuur. Als we de kans op het behalen van de gewenste biologische waterkwaliteit zetten op 80%, komt de norm uit op 443 voor de concentratie van nutriënt X (Figuur 5). Een hogere kans leidt tot een strengere norm, een lagere kans tot een minder strenge norm. Overigens is het geenszins zeker dat deze 80% altijd wordt gehaald, het is alleen de meest aannemelijke schatting.

Keuzemogelijkheid II. Op zeker spelen.
Waneer men zeker wil weten, zeg met een onzekerheid van 2,5%, dat in 80% van de gevallen de biologische waterkwaliteit wordt gehaald, zal men de onzekerheid van de S-curve in beschouwing moeten nemen. Een andere waarde dan 2,5% is ook mogelijk wanneer we een ander betrouwbaarheidsinterval berekenen, maar in ons voorbeeld is het 95% betrouwbaarheidsinterval berekend. We nemen nu als normconcentratie de concentratie, waarbij de ondergrens van het betrouwbaarheidsinterval wordt bereikt, in het voorbeeld zal dus de normconcentratie ongeveer 355 worden.

Keuzemogelijkheid III. Combinatie
Beperk het risico van het niet behalen van de doelstelling door een lager doel te stellen waar met een hoge betrouwbaarheid aan dient te worden voldaan. Zo'n minder stringente doelstelling is bijvoorbeeld dat in 50% van de gevallen de biologische waterkwaliteit moet worden gehaald. Daartegenover wil men wel redelijk zeker zijn (onzekerheid 2,5%) dat deze doelstelling wordt gehaald. In het voorbeeld ligt deze norm bij 428 (Figuur 5) Wanneer het betrouwbaarheidsinterval relatief smal is, kan concentratie III hoger zijn dan concentratie I.

Zorg daarom tenminste voor representativiteit en onafhankelijkheid van de gegevens (zie paragraaf 2.2). Indien u uit een bestaande dataset selecteert en er voldoende gegevens zijn, neem dan een aselecte steekproef van voldoende grootte. Als vuistregel kan worden aangenomen dat voor deze methode de responsvariabele in tenminste 15-20 gevallen 0 moet scoren (gewenste kwaliteit niet gehaald) en in tenminste 15-20 gevallen 1 moet scoren (gewenste kwaliteit gehaald), dus dat tenminste 30-40 waarnemingen voor de analyse nodig zijn.

2.3.2 Onderzoeksopzet en bemonsteringsschema

Bepaal aan de hand van de criteria waaraan de data moeten voldoen (zie paragraaf 2.2), waar zal worden gemeten (niet te dicht bij elkaar gelegen monsterpunten) en met welke frequentie (niet te vaak). Bedenk tevoren welke andere variabelen (bijvoorbeeld beheer of bestrijdingsmiddelen) ook van invloed zouden kunnen zijn op de biologische waterkwaliteit en neem deze ook op in de bemonstering. Noteer bijvoorbeeld hoe lang geleden voor het laatst is gebaggerd of geschoond e.d. De frequentie, waarmee biotische gegevens worden verzameld en de frequentie waarmee andere gegevens worden verzameld, dient voor de analyse gelijk te zijn. Het verdient echter aanbeveling om nutriëntengegevens vaker te verzamelen dan biologische gegevens, daar zij een grotere variabiliteit in tijd en ruimte vertonen. Wanneer nutriënten vaker worden gemeten dan biotische gegevens, zorg dan voor aggregatie: neem bijvoorbeeld het (geometrisch) gemiddelde van de nutriëntenconcentratie, de mediane concentratie of de concentratie op een relevant tijdstip. Wanneer een beheersmaatregel met een lagere frequentie wordt uitgevoerd, noteer dan toch bij elke bemonstering wanneer deze voor het laatst uitgevoerd is.

2.3.3 Opslag ruwe gegevens

Sla alle relevante gegevens op in een spreadsheet of database Hanteer hierbij het basisprincipe, dat elke meting dient te worden bewaard, om eventueel later de basisgegevens op een andere wijze te kunnen gebruiken. Bewaar dus niet uitsluitend jaargemiddelden e.d. (later zou bijvoorbeeld kunnen blijken dat u beter het geometrisch gemiddelde had kunnen nemen). Minimaal benodigde gegevens zijn: code monsterpunt, datum, gemeten concentraties, uitgevoerde beheersmaatregelen, gemeten biotische variabelen. Zorg er te allen tijde voor dat een eenvoudige koppeling mogelijk is tussen de responsvariabelen en de stuurvariabelen, m.a.w. dat de codes van de monsterpunten uniek zijn, dus overeenkomend in alle bestanden, of sla alle gegevens in één bestand op.
Het verdient de voorkeur gebruik te maken van een relationele database, waarin alle gegevens worden opgeslagen per monsterpunt, met afzonderlijke bestanden voor biotische variabelen, nutriënten en beheer.

2.3.4 Voorbewerking gegevens

Selecteer de relevante gegevens uit elk van de spreadsheets of databases. Zorg ervoor dat in elk tussenbestand tenminste de monstercode en datum (of jaar en/of maand) vermeld staan. Bedenk op welke temporele schaal gegevens moeten worden

dat wil zeggen dat binnen het watertype (bijvoorbeeld sloten) verschillende deeltypes (bijvoorbeeld zand- en kleisloten) aselect worden bemonsterd. Het is in het laatste geval mogelijk om verschillen tussen deeltypes te toetsen, terwijl ook een algemeen resultaat kan worden verkregen. Tenslotte is het ook toegestaan om een meer evenwichtige verdeling van de waarden van de stuurvariabelen (bijvoorbeeld nutriëntenconcentraties) na te streven. Wanneer echter een range van nutriëntenconcentraties buiten beschouwing gelaten wordt is het geschatte model op zijn minst minder betrouwbaar in die range.

2.2.3 Homogeniteit

Binnen een set te analyseren gegevens moeten bemonsteringsmethode, bemonsteringsfrequentie en analysemethoden hetzelfde zijn. Indien er verschillen zijn kan hiervoor bij de analyse eventueel worden gecorrigeerd, maar het voert in het kader van deze handleiding te ver om deze correcties uitvoerig te beschrijven. Hier is het voldoende te vermelden dat in feite geen correctie wordt geschat om voor de verschillende analysemethoden te corrigeren, maar een apart relatie met de waterkwaliteit per analysemethode. Problemen ontstaan, wanneer een andere analysemethode samengaat met bepaalde watertypen. Als bijvoorbeeld de pH in alle sloten anders is gemeten dan in alle meren is niet te achterhalen of een eventueel verschil in pH en een eventueel verschil in respons op de pH worden veroorzaakt door het verschil in watertype dan wel door het verschil in meetmethode.

2.2.4 Representativiteit

Om een goed beeld te krijgen van de relatie tussen stuur- en responsvariabelen dienen de gegevens goed gespreid te zijn in ruimte en tijd. Dat wil zeggen dat bij voorkeur gegevens uit meerdere jaren en van voldoende monsterpunten dienen te worden gebruikt en dat het aantal monsterpunten, wanneer verschillende watertypen in de analyse gecombineerd zijn, naar rato van oppervlak (meren) of lengte (beken, sloten) is verdeeld over de verschillende watertypen. In het algemeen zal, wanneer aan goede spreiding in ruimte en tijd gedacht is, ook de spreiding over nutriëntenconcentraties en biotische kwaliteit voldoende zijn.

2.3 Gegevensverzameling en voorbewerking

Hieronder volgen een aantal vuistregels, tips en opmerkingen die belangrijk zijn voor een juiste gegevensverzameling en voorbewerking van de data.

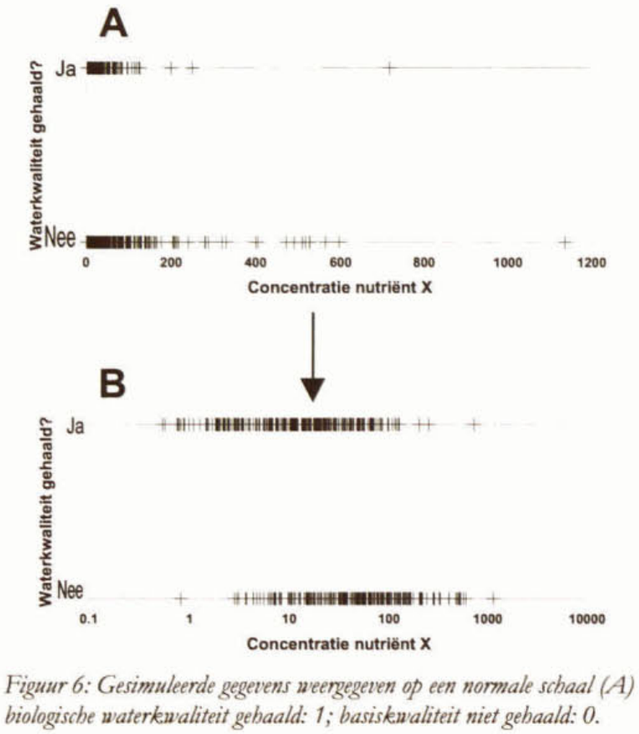
2.3.1 Begin bij het einde

Bedenk, voordat u met gegevens verzamelen begint, met welk doel de gegevens moeten worden verzameld. In het geval dat u gegevens uit een bestaande set selecteert, wat vaak het geval zal zijn, geldt hetzelfde.
Het doel van de analyse is het vaststellen van de relatie tussen biotische kwaliteit en nutriëntenconcentraties en eventueel beheers- of inrichtingsmaatregelen.

3 Uitgewerkte voorbeelden in Excel en SPSS

3.1 De dataset

Figuur 6 geeft de data van de voorbeeld dataset weer. Wanneer we de concentraties uitzetten in een grafiek zien we direct dat er vele metingen beschikbaar zijn in het lage concentratietraject (zie bijvoorbeeld Figuur 6A, concentratiebereik 0 t/m 200) en maar enkele in het hoge bereik (bijvoorbeeld maar 2 waarnemingen in Figuur 6A tussen 600 en 1200). De figuur geeft dus maar de informatie van een paar uitschieters weer en niet de informatie van de meerderheid van de getallen. Door de concentraties te transformeren kan een betere verdeling van alle metingen over de as worden verkregen (zie bijvoorbeeld Figuur 6B). Gekozen is voor de logaritme van de concentratie omdat ook in de praktijk concentraties vaak log-normaal verdeeld zijn en door deze transformatie de concentraties regelmatiger langs de X-as verspreid komen te liggen en daarom de invloed van extreem hoge waarden op het resultaat van de analyse wordt verminderd. Het voert hier te ver om transformatie van data uitvoerig te bespreken, voor meer informatie wordt verwezen naar Oude Voshaar (1994).
Dat de concentraties in deze dataset log-normaal verdeeld zijn komt omdat deze zo geconstrueerd zijn. Voor het maken van deze dataset zijn eerst uit de normale verdeling 500 random getallen getrokken met een gemiddelde van 3 en een standaarddeviatie van 1.4. . Deze waarden zijn gebruikt als Ln(concentratie) zodat 95 % van de Ln(concentraties) van het virtuele nutriënt ligt tussen ongeveer 0.2 (3-(2*1.4)) en 5.8 (3+(2*1.4)) Ln(eenheden). Bij benadering 95 procent van de nutriëntenconcentraties ligt dus tussen de 1.2 (exp(0.2))en 330 (exp(5.8)) eenheden (bijvoorbeeld $\mu g\ l^{-1}$).
Met behulp van deze Log(concentraties) en twee gekozen parameters voor b_0 en b_1 (b_0=3 en b_1=-1) is met behulp van formule (1) voor ieder Log(concentratie)-getal de verwachte kans op het behalen van de waterkwaliteit berekend. Deze verwachte kans is omgezet naar een 0 (niet behalen van de waterkwaliteit) of een 1 (wel behalen van de waterkwaliteit) door iedere kans te vergelijken met een random getal tussen de 0 en 1. Als het random getal kleiner is dan de verwachte kans wordt een 1 toegekend, als deze groter is een 0. Hierdoor worden verwachte kansen die de 1 benaderen meestal vervangen door een 1 en gevallen waar de verwachte kans heel laag is meestal door een 0. Dit leidt tot een dataset waar lage nutriëntconcentraties over het algemeen een 1 scoren (wel behalen van de waterkwaliteit) en hoge concentraties een 0 (niet behalen van de waterkwaliteit).



Figuur 6: Gesimuleerde gegevens weergegeven op een normale schaal (A) en een log-schaal (B). biologische waterkwaliteit gehaald: 1; basiskwaliteit niet gehaald: 0.

Wanneer we uitsluitend naar de monsterpunten in de grafiek kijken, zien we duidelijk dat bij lage concentraties de vereiste kwaliteit behaald is, bij hoge concentraties niet en dat ertussenin een overgangszone is, waarin zowel monsters aanwezig zijn die voldoen aan de biologische kwaliteitsnorm als monsters die daar niet aan voldoen. Het ligt dus voor de hand dat de relatie tussen de kans dat het vereiste kwaliteitsniveau wordt behaald en de concentratie een dalende functie is.

3.2 Logistische regressie m.b.v. Excel

We hebben enerzijds de dataset (Figuur 6B) en anderzijds de verwachte relatie als beschreven in formule (1), welke herschreven kan worden m.b.v. de logit transformatie tot:
logit $\{p(\text{Kwaliteit}=1)\} = b_0 + b_1 * \text{Ln}(\text{Concentratie})$
Het is niet mogelijk deze functie op te lossen met behulp van simpele lineaire regressie daar we geen kansen verzameld hebben in onze dataset maar gerealiseerde uitkomsten (zie paragraaf 2.1). Een manier om de parameters b_0 en b_1 te bepalen is ze net zolang te variëren in waarde totdat de afwijking van de verwachte kans op een goede biologische waterkwaliteit en de gerealiseerde biologische waterkwaliteit zo klein mogelijk is. In het geval van Figuur 6B bijvoorbeeld moet de gefitte kans op het behalen van de biologische waterkwaliteit erg hoog zijn tussen de concentraties

regressie geanalyseerd worden. Een voorbeeld is opgenomen in de bijlagen. Tevens is een voorbeeld opgenomen waarbij het voorkomen van de gewenste waterkwaliteit een optimum heeft, dat wil zeggen dat deze bij een toenemende nutriëntconcentratie eerst toeneemt en dan weer afneemt. Theoretisch kan zo'n relatie verwacht worden voor bijvoorbeeld hogere waterplanten; bij een te lage nutriëntconcentratie kunnen zij niet groeien, maar bij een te hoge nutriëntconcentratie worden zij weggeconcurreerd door de algen. In dit voorbeeld is tevens verwerkt hoe binnen de analyse rekening gehouden kan worden met de invloed van een bepaald beheer van de monsterpunten (bijvoorbeeld baggeren).

2.2 Randvoorwaarden van de invoergegevens

Om de bovenstaande analyse, maar ook andere statistische methoden, te mogen toepassen moeten de invoergegevens aan een aantal basisrandvoorwaarden voldoen. Hieronder zullen de belangrijkste worden besproken.

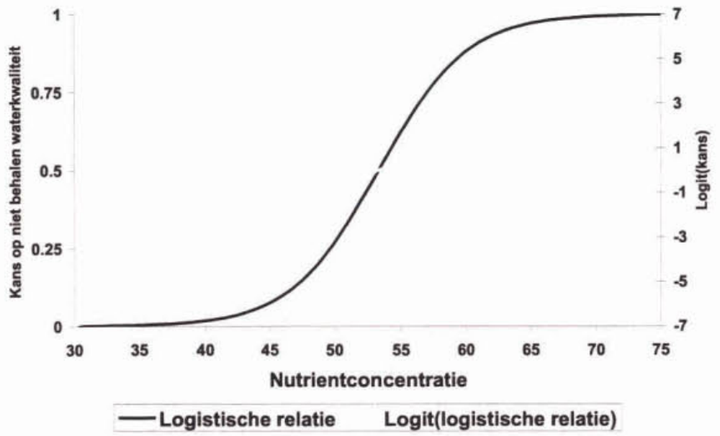
2.2.1 Onafhankelijkheid

Ten eerste dienen de monsters, maar dit geldt voor elke regressiemethode, onafhankelijk te zijn. Dat wil zeggen, dat de monsters in tijd en ruimte zodanig gespreid zijn dat de waarde van het ene monster geen directe relatie heeft met de waarde van een ander. Dit geldt bijvoorbeeld voor monsters die in dezelfde sloot (een paar meter verderop) zijn genomen of monsters die op dezelfde locatie zijn genomen met bijvoorbeeld een week tussenpoos. Immers, als ergens bijvoorbeeld ondanks lage nutriëntenconcentraties de basiskwaliteit niet wordt gehaald, is de kans groot dat deze situatie korte tijd later ook nog bestaat. Preciezer gezegd: de waarde van een bepaald monster mag niet beter voorspeld worden door de waarde van in tijd en/of ruimte nabij gelegen monsters die ook bij de schatting van de parameters gebruikt zijn, dan door de algemene relatie die geschat is. Wanneer een monsterpunt bijvoorbeeld kort na elkaar twee maal wordt bemonsterd, kan verwacht worden dat de metingen die aan deze monsters worden uitgevoerd op dezelfde wijze zullen afwijken van het algemene verband.

2.2.2 Aselecte keuze van monsterpunten

Ten tweede dienen de monsterpunten aselect te worden gekozen, dat wil zeggen, dat de keuze van monsterpunten niet afhankelijk mag zijn van de waarde van de responsvariabelen (in ons geval de behaalde biotische kwaliteit, bijvoorbeeld aan- of afwezigheid van waterplanten) of van andere variabelen die mogelijk van invloed zijn op het resultaat. Met andere woorden: er mogen niet uitsluitend "mooie" of "slechte" punten worden gekozen, en de locatie van de monsterpunten mag niet direct worden bepaald door factoren als bereikbaarheid, situering binnen een water (oever, midden, bovenloop, benedenloop etc.). Het is echter wel toegestaan om het watertype waarvoor de gedifferentieerde normen moeten worden opgesteld nauwer te omschrijven (bijvoorbeeld middenlopen van beken, oevers van sloten etc.). Hierdoor worden de resultaten van de analyse echter ook beperkt tot het in de onderzoeksopzet omschreven watertype. Ook is het toegestaan om te stratificeren,

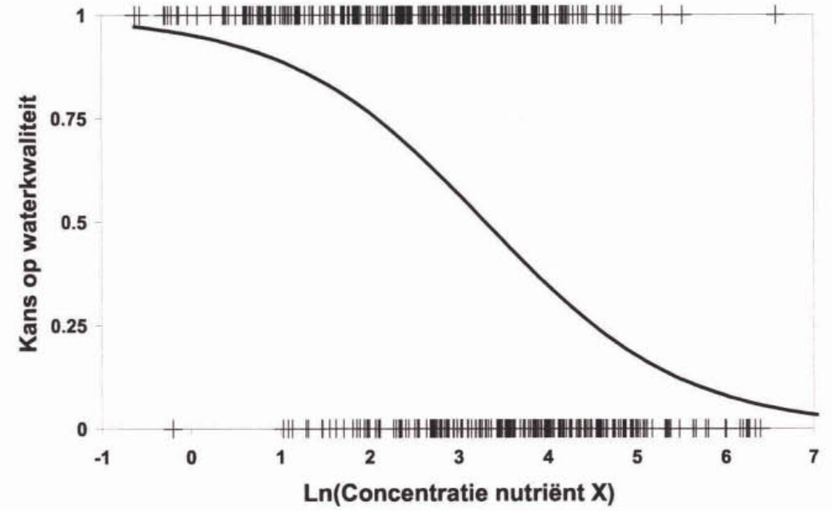
Het probleem wordt opgelost door niet de waargenomen fracties te transformeren om een lineair model te krijgen, maar de bij het model verwachte fractie gegeven de gekozen parameters. De parameters kunnen worden geschat door in een iteratief proces (stapsgewijze benadering) steeds de parameters van de lineaire functie bij te stellen en dan te berekenen hoe groot de aannemelijkheid is dat, gegeven de geschatte kansen de resultaten zo zijn als waargenomen (zie Oude Voshaar, 1994 voor meer details). Deze methode wordt logistische regressie genoemd, naar de logistische (sigmoïde) curve die ermee wordt beschreven. Omdat we in feite een lineair model aanpassen, maar gebruik maken van een andere, namelijk binomiale, verdeling van de fouten en van een transformatie van de verwachte waarden in plaats van een transformatie van de waarnemingen valt dit model binnen de klasse van gegeneraliseerde lineaire modellen (GLM's). Deze manier van optimaliseren maakt gebruik van de 'meest aannemelijke schatter' (maximum likelihood estimator), waarbij de likelihood (aannemelijkheid, dat wil zeggen de kans dat het geschatte model leidt tot de waarnemingen die in het veld gedaan zijn) wordt gemaximaliseerd door aanpassing van de te schatten parameters. Dit wordt gedaan door minimalisatie van de som van –2ln(likelihood) over alle monsters. Hierna kan dit model worden getoetst t.o.v. het nul-model (een model waarin geen invloed is van de concentratie op de kans op behalen van de kwaliteit, dus waarin er over de gehele concentratierange een gelijke kans is dat de biologische waterkwaliteit wordt gehaald) met behulp van de chi-kwadraat verdeling (zie Oude Voshaar, 1994 voor meer details).



Figuur 4: Voorbeeld van een logistische relatie en haar getransformeerde lineaire logit relatie.

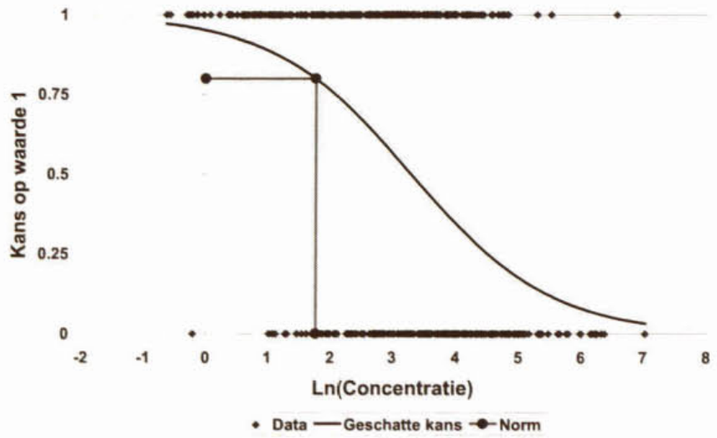
Hierboven is de aanname gedaan dat de waterkwaliteit door de concentratie van één nutriënt bepaald wordt. Het is daarentegen zeer wel denkbaar dat de waterkwaliteit niet bepaald wordt door de concentratie van èèn nutriënt maar door een combinatie van twee (bijvoorbeeld N en P) of meerdere. In dat geval kunnen de effecten van de nutriënten onderling onafhankelijk optreden, maar ze kunnen elkaar ook beïnvloeden. Deze zogenaamde interactie kan ook met behulp van logistische

0.1 en 2 en erg laag tussen de concentraties 500 en 10000 (Figuur 6B). Door het minimaliseren van de afwijkingen tussen de datapunten en de gefitte curve (Figuur 7) worden de meest aannemelijke schatters (maximum likelihood estimates) van b_0 en b_1 bepaald (zie paragraaf 1.2).



Figuur 7: Gesimuleerde gegevens weergegeven op een log-schaal en de fit als uitgevoerd in Excel.

De afwijkingen worden uitgedrukt in -2ln(likelihood). De likelihood is de aannemelijkheid van de gerealiseerde uitkomst (biologische waterkwaliteit = 1 of 0). Met ander woorden: voor een monsterpunt waar de vereiste kwaliteit gehaald is, is de likelihood gelijk aan de kans op behalen van de kwaliteit; voor een monsterpunt waar de vereiste kwaliteit niet behaald is, is deze gelijk aan de kans op niet behalen van de kwaliteit. Deze kans is gelijk aan 1 minus de kans op behalen van de kwaliteit. Voor ieder monster wordt deze –2ln(likelihood) berekend en hierna gesommeerd over alle monsters (zie paragraaf 2.1). Voor iedere combinatie van b_0 en b_1 kan dus een likelihood worden berekend. De minimalisatie van de –2ln(likelihood) (hetgeen neerkomt op maximalisatie van de aannemelijkheid, het gebruik van de factor 2 heeft te maken met de relatie met de Chi-kwadraat verdeling) kan binnen Excel worden uitgevoerd met behulp van de "oplosser" optie. Tevens kunnen de nutriëntenconcentraties bij verschillende kwaliteitsdoelstellingen worden bepaald. Als bijvoorbeeld de kwaliteitsdoelstelling is dat in 80% van de gevallen de biotische kwaliteit gelijk is aan de basiskwaliteit moet de nutriëntenconcentratie behorende bij p(behalen biologische waterkwaliteit)= 0.80 worden berekend. Dit kan ook met behulp van de "oplosser" optie binnen Excel.
Een beschrijving van de analyse in Excel is gegeven in de Excel file "LogistischeRegressieExcel.xls". In deze file is tevens beschreven hoe deze analyse met nieuwe data moet worden uitgevoerd. In de file wordt in het werkblad "grafiek" de output in de vorm van een grafiek weergegeven (zie Figuur 8), in het werkblad "Resultaat" zijn de belangrijkste resultaten in tabelvorm samengebracht (zie Tabel 1).

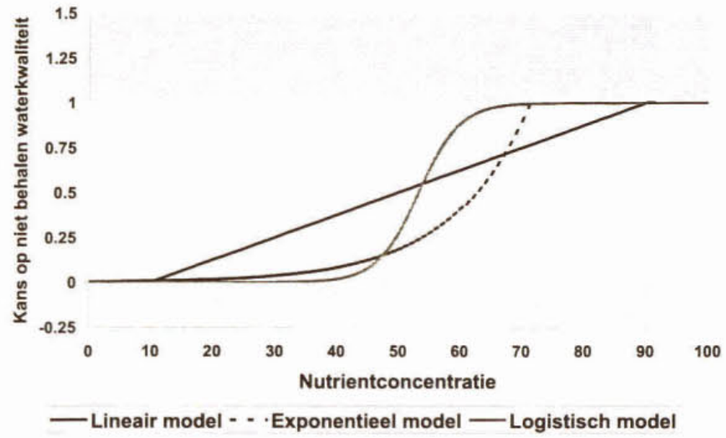


Figuur 8: Voorbeeld van output Excel-file als gebruikt voor de logistische regressie. De norm geeft de nutriëntenconcentratie waarbij is geschat dat in 80% van de gevallen de biologische waterkwaliteit wordt gehaald.

Tabel 1: Voorbeeld van output Excel-file als gebruikt voor logistische regressie. De tabel geeft de meest belangrijke resultaten van de analyse weer

Aantal punten	500
Waarvan waarde 1	271
Waarvan waarde 0	229
Gemiddelde kans op waarde 1	0.542
-2ln(likelihood) null model	689.62
-2ln(likelihood) model	557.06
Chi-kwadraat	132.55
Significantie	1.13E-30
Intercept	2.982
Hellingshoek	-0.905
Dus de lineaire predictor: logit { p(Kwaliteit=1) } = 2.982 + -0.905 * Log (Concentratie)	

Een nadeel van de analyse m.b.v. Excel is dat wel een juiste beschrijving van de relatie word verkregen maar dat geen standaardafwijking van de parameters b_0 en b_1 wordt verkregen. Oftewel Keuzemogelijkheid I binnen de interpretatie als beschreven in paragraaf 2.5 is wel mogelijk maar de interpretatie als beschreven bij Keuzemogelijkheid II en III niet. We hebben daarom het bovenstaande voorbeeld ook met behulp van een gespecialiseerd statistisch pakket (SPSS) geanalyseerd.



Figuur 3: Verschillende modellen voor de beschrijving van de relatie tussen het niet behalen van de biologische waterkwaliteit en nutriëntconcentratie. De grijze gebieden geven onmogelijke kanswaarden aan.

Wanneer het verband tussen nutriëntconcentratie en kans op behalen van de gewenste kwaliteit een sigmoïde curve is, kunnen we dus de zogenaamde logit transformatie toepassen (welke de omgekeerde bewerking van formule 1 is):

$$\text{Logit}(p)=\text{Ln}(p/(1-p)), \qquad (2)$$

Wanneer de formules (1A) en (2) gecombineerd worden ontstaat een lineair verband tussen de getransformeerde kans en de concentratie:

$$\text{logit}(p) = b_0 + b_1 x \qquad (3)$$

De x-waarden in dit model zijn de nutriëntenconcentraties, de y-waarden de logit van de fractie van de monsterpunten waarbij de waterkwaliteit behaald wordt gegeven de nutriëntenconcentraties. Wanneer we dus (na transformatie) een gewone lineaire regressie uitvoeren beschrijven we een sigmoïde verband tussen concentraties en kans op behalen van de gewenste kwaliteit.
Om de fractie te kunnen berekenen is het echter wel nodig dat meerdere gegevens omtrent het al of niet behalen van de waterkwaliteit beschikbaar zijn voor een bepaalde nutriëntconcentratie. Deze zullen in de praktijk niet beschikbaar zijn, daar we uitsluitend over gegevens betreffende aparte monsterpunten beschikken met ieder hun eigen, unieke nutriëntconcentratie. Hierdoor kan deze transformatie niet simpel worden toegepast om vervolgens de parameters b_0 en b_1 met behulp van eenvoudige lineaire regressie te bepalen. Een oplossing zou zijn om een aantal waarnemingen die bijna dezelfde nutriëntconcentratie vertegenwoordigen te groeperen en voor deze groep de fractie te berekenen, maar deze methode heeft als nadeel dat het aantal waarnemingen sterk wordt gereduceerd.

waterkwaliteit. Tevens nemen wij aan dat er in relatie tot de waterkwaliteit geen interacties tussen verschillende nutriënten optreden. In de bijlagen komen echter wel ingewikkelder gevallen aan de orde.
De modelmatige beschrijving van de relatie tussen het wel of niet behalen van de gewenste biologische waterkwaliteit en de concentratie van een nutriënt wordt gedefinieerd als een kans. De kans om de biologische waterkwaliteit te halen is bijvoorbeeld bijna 100% bij een nutriëntconcentratie van 200 en bijna 0% bij een nutriëntconcentratie van 800 (Figuur 2). De relatie loopt bijna horizontaal in de concentratie ranges van 0 tot 300 en 700 en 1000 en loopt naar beneden tussen concentraties van 300 tot 700. Verder is het duidelijk dat de kans op het behalen van de biologische waterkwaliteit niet kleiner dan 0 en niet groter dan 1 kan zijn.
Uit de hierboven beschreven uitgangspunten blijkt dat een lineair model als
$p = b_0 + b_1 x$,
waarin p de kans is op behalen van de vereiste biologische waterkwaliteit, x de nutriëntconcentratie, b_0 het intercept (snijpunt met y-as) en b_1 de helling van het verband, niet geschikt is voor de beschrijving van de relatie tussen het behalen van de gewenste biologische waterkwaliteit en nutriëntconcentratie. Dit omdat waarden lager dan 0 en hoger dan 1 verkregen kunnen worden (zie Figuur 3). Het eerste probleem kan worden verholpen door een exponentieel model te gebruiken:
$p = exp(b_0 + b_1 x)$.
Dit model zorgt ervoor dat negatieve kansen worden vermeden, maar voorkomt niet dat kansen boven de 1 worden verkregen (Figuur 3). Door het exponentiele model door zichzelf plus 1 te delen wordt een sigmoïde curve verkregen welke door het volgende model wordt beschreven:

$$p = \frac{e^{b_0+b_1 x}}{(1+e^{b_0+b_1 x})} \qquad (1)$$

Hierin staat x voor de nutriëntconcentratie en p voor de kans dat de biologische waterkwaliteit wordt behaald, b_0 en b_1 zijn de parameters van het model. Dit model beschrijft de kans op behalen van de vereiste biologische waterkwaliteit, die gelijk is aan de fractie van de monsterpunten die bij een bepaalde nutriëntconcentratie of in een nauwe range van nutriëntconcentraties, de vereiste biologische waterkwaliteit behalen. Deze curve, de logistische curve, heeft alle eigenschappen die we hierboven hebben beschreven: hij blijft tussen 0 en 1 en loopt aan de uiteinden van beide staarten bijna horizontaal (Figuur 3).

Formule 1 kan volgens de volgende stappen herschreven worden:

$$p*(1+e^{b_0+b_1 x}) = e^{b_0+b_1 x} \Rightarrow$$
$$p + p*e^{b_0+b_1 x} = e^{b_0+b_1 x} \Rightarrow$$
$$p = (1-p)e^{b_0+b_1 x} \Rightarrow \qquad (1A)$$
$$\frac{p}{1-p} = e^{b_0+b_1 x} \Rightarrow$$
$$\ln(\frac{p}{1-p}) = b_0 + b_1 x$$

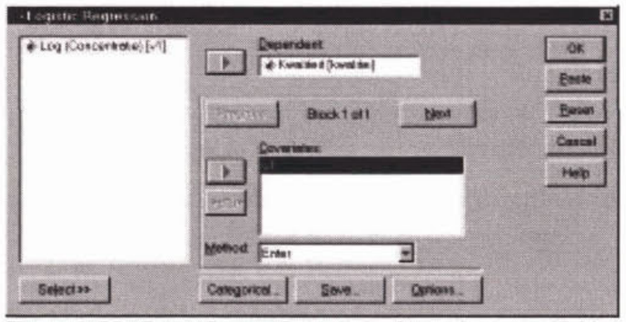
3.3 Logistische regressie m.b.v. SPSS

Hieronder volgt een stapsgewijze handleiding hoe de analyse in SPSS is uit te voeren gevolgd door enkele opmerkingen en de verwerking van de SPSS output tot een interpreteerbare grafiek en normconcentraties.

3.3.1 Stapsgewijze handleiding

Data invoer in SPSS
Neem als basis het werkblad gegeven van de file: "LogistischeRegressieExcel.xls". Kopieer dit werkblad naar een nieuwe file en sla deze op als "SSDinvoer.xls". Open SPSS en lees de Excel file in m.b.v. de commando's "File", "Open", "Data", File of type: Excel (*.xls)", "File name: SPSSinvoer.xls", "Open" en "opening Excel Data Source: OK". De data zijn nu ingelezen in twee kolommen: kolom 1 Log(concentratie), label: "v1" en kolom 2 Kwalitei, label: "Kwaliteit". In kolom 3 staan nog wat opmerkingen, deze spelen geen rol in de verdere analyse.

Data analyse in SPSS
Kies, als de data goed ingelezen zijn, voor "Analyse", "Regression" en "Binary Logistic" zodat het scherm als gegeven in Figuur 9 verschijnt.



Figuur 9: Het Binary Logistic Regression selectie scherm van SPSS

Kies net als in Figuur 9 Kwaliteit [kwalitei] als Dependent variable en Log(Concentratie) [v1] als Covariates (independent variable) en druk op OK. Er verschijnt nu een uitvoerscherm, waar men de gegevens als weergegeven in Tabel 2 onderaan (klik op "Variables in the Equation" regel in linker scherm).

Tabel 2: Samenvatting van uitvoer SPSS

Variables in the Equation			
	B	S.E.	Sig.
V1	-0.9049	0.0945	1.06E-21
Constant	2.9823	0.3167	4.62E-21

Hierin stelt "Constant" b_0 voor en V1, b_1. Het model wordt dus in dit geval:
logit {p(Kwaliteit=1)}= 2.98 -0.90* Log(Concentratie)
Tevens hebben we nu ook de betrouwbaarheid en significantie van b0 en b1 in de vorm van hun standaardfouten (SE) en p-waarden (Sign.).

3.3.2 Opmerkingen bij interpretatie van de uitvoer

1. De 2ln(likelihood) is hoger, naarmate meer waarnemingen in de analyse meedoen. Dit geldt zowel voor het 0-model (het model dat uitsluitend de gemiddelde kans schat; d.w.z. uitsluitend parameter b_0 bevat) als voor het regressiemodel, waarin zowel het intercept (parameter b_0) als de helling (parameter b_1) geschat is. Het verschil tussen de 2ln(likelihood) waarden van beide modellen (met en zonder hellingsparameter b_1) is de toetsingsgrootheid, die dient te worden vergeleken met de waarde van Chi-kwadraat bij 1 vrijheidsgraad (immers alleen de helling is extra geschat).
2. De fout in de schatting van intercept (b_0) en helling (b_1) neemt toe met afnemend aantal waarnemingen.
3. Grote afwijkingen tussen de gefitte en de werkelijke waarde (in de voorbeelden van dit rapport zijn de gegevens geconstrueerd, dus weten we wat het eindresultaat moet zijn) komen vooral voor bij lage aantallen waarnemingen.

In het kader van deze handleiding voert het te ver om een uitgebreider studie te laten zien naar de effecten van variatie van de parameters in het model en van variatie in het aantal nullen en enen. Voor het voorkomen van plantensoorten langs ecologische gradiënten heeft Bio (2000) daarnaar een uitvoerige studie gedaan, die voor geïnteresseerden een bron van inspiratie kan zijn.

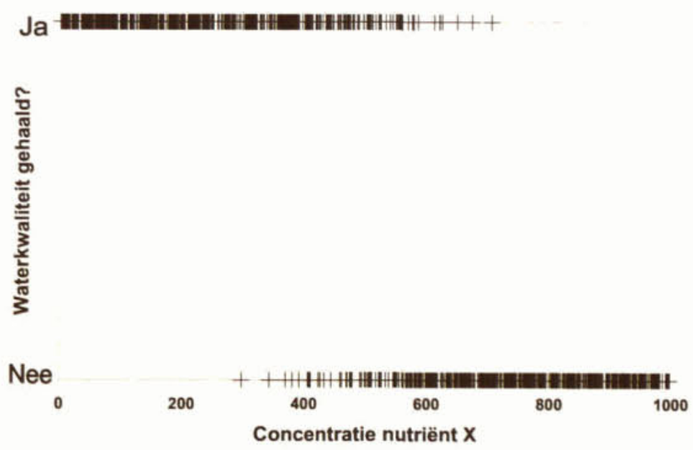
3.3.3 Verwerking van de resultaten m.b.v. Excel

Open de Excel-file "LogistischeRegressieSPSS.xls" en voer de waarden van de parameter b_0 en b_1 in als aangegeven in de file. Voer de handelingen uit als beschreven in het werkblad "Beschrijving". Figuur 10 laat het resultaat van de analyses grafisch zien, Tabel 3 geeft de cijfers weer.

2 Theorie, randvoorwaarden en uitvoering analyse

2.1 Theorie

Als basismodel voor de beschrijving van de relatie tussen het al of niet behalen van de beoogde biologische waterkwaliteit en de nutriëntenconcentraties worden de modellen II en III als gegeven in Figuur 1 gebruikt. Stel dat de waterbeheerder in een aantal meren zowel de concentraties van een aantal nutriënten heeft bepaald alsook het wel of niet aanwezig zijn van macrofyten. De beoogde biologische waterkwaliteit is het aanwezig zijn van macrofyten. Een grafische weergave van een geconstrueerde voorbeeld dataset is gegeven in Figuur 2. Bij lage concentraties wordt de beoogde biologische waterkwaliteit wel behaald en bij hoge concentraties niet. De concentratie range van 400 tot 600 is een overgangszone; in sommige gevallen wordt de biologische waterkwaliteit wel gehaald en in sommige niet (Figuur 2).



Figuur 2: Grafische weergave van een, voor gedifferentieerde normstelling verzamelde voorbeeld dataset. Het betreft data over het wel of niet halen van de beoogde biologische waterkwaliteit (bijvoorbeeld het aan- of afwezig zijn van macrofyten) en de concentratie van een nutriënt.

De relatie tussen biologische waterkwaliteit en nutriëntenconcentraties is in werkelijkheid meestal complexer. Het voorkomen van bijvoorbeeld macrofyten is beperkt tot een range van nutriëntenconcentraties. Zo zullen macrofyten bij zeer lage en zeer hoge concentraties niet voorkomen. In het eerste geval zullen de nutriënten de groei van de macrofyten beperken en in het laatste geval zullen de algen de macrofyten verdringen in de concurrentie om licht. In het algemeen zal de relatie dus tenminste een optimum hebben, dat wil zeggen dat bij toenemende concentraties van een nutriënt de kans op voorkomen van een levensgemeenschap eerst toe zal nemen en later weer af zal nemen (unimodaal model). In dit rapport gaan wij er voor de bespreking van de theorie van uit dat de in Nederland voorkomende concentraties van nutriënten niet limiterend zijn voor het halen van de beoogde biologische

In deze analyse worden echter alleen de gegevens gebruikt waarbij het kwaliteitsniveau gehaald is, en niet de data betreffende de locaties waarbij dit niet het geval is. De analyse beperkt zich dus alleen tot de data verzameld in de range van nutriëntenconcentraties waarbij het kwaliteitsniveau is gehaald (een van de horizontale lijnen in figuur 1) en houdt geen rekening met de overlap tussen kwaliteitsniveaus in een concentratiebereik van de nutriënten (zie model II en II in figuur 1). Van Tongeren en Gremmen (1999) hebben daarom voorgesteld logistische regressie te gebruiken in plaats van de methode Peeters-Gardeniers. Het voordeel van deze methodiek is dat zij alle data (dus zowel locaties waar wel als locaties waar niet de gewenste biologische waterkwaliteit is gehaald) gebruikt voor de beschrijving van de relatie en dat zij rekening houdt met het feit dat bij een bepaalde concentratie verschillende kwaliteitsniveaus kunnen optreden.

1.4 Leeswijzer

In Hoofdstuk 2 van dit rapport zullen de theorie en randvoorwaarden van de hier voorgestelde methodiek aan de orde worden gesteld. Hoofstuk 3 bevat theorie en een stapsgewijze beschrijving van de analyse, de uitwerking van een gesimuleerde dataset en de bewerking van een set gegevens afkomstig van het RIZA. Een deel van de tekst is met wijzigingen overgenomen uit Van Tongeren en Gremmen (1999). De data voor het praktijkvoorbeeld als beschreven in bijlage 1 zijn afkomstig van het RIZA. Het voorbeeld in bijlage 2 is weer afkomstig van een simulatie

Tabel 3: Voorbeeld van tabel met resultaten als aangemaakt door Excel-file die wordt gebruikt voor het visualiseren van de SPSS output regressie.

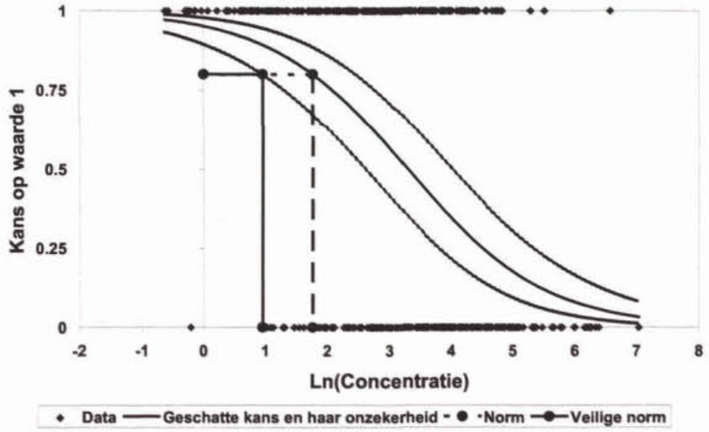
Aantal punten	500
Waarvan waarde 1	271
Waarvan waarde 0	229
Gemiddelde kans op waarde 1	0.542

Norm op basis van verwachte kans op slagen:

Kans op behalen gewenst resultaat	0.8
Norm concentratie (of Ln concentratie)	1.763770338

Veilige Norm op basis van ondergrens betrouwbaarheidsinterval:

Onbetrouwbaarheid	0.05
Kans op behalen gewenst resultaat	0.8
Norm concentratie (of Ln concentratie)	0.956030835



Figuur 8: Voorbeeld van grafiek als aangemaakt door Excel-file die wordt gebruikt voor het visualiseren van de SPSS output regressie. De norm geeft de nutriëntenconcentratie waarbij is geschat dat in 80% van de gevallen de biologische waterkwaliteit wordt gehaald. De veilige norm geeft de nutriëntenconcentratie waarbij het 95% zeker is dat in 80% van de gevallen de biologische waterkwaliteit wordt gehaald.

De gefitte geschatte kansen zijn berekend m.b.v. formule (2). Voor de verwachte kansen kan m.b.v. de standaardfouten van de parameters als berekend door SPSS, een betrouwbaarheidsinterval worden uitgerekend. De formule waarmee het 95% betrouwbaarheidsinterval voor de lineaire predictor wordt berekend is:

$$y = b_0 + b_1 x \pm t_{v,\alpha} * \sqrt{se_{b_0}^2 + (x - \bar{x}) * se_{b_1}^2} \quad (4)$$

waarin y de onder- respectievelijk bovengrens van het betrouwbaarheidsinterval, het eerste deel van de formule gelijk aan de gebruikelijke vergelijking voor de lineaire regressie (b_0 het intercept en b_1 de hellingshoek), $t_{v,a}$ de waarde van t bij het betreffende aantal vrijheidsgraden v en het onbetrouwbaarheidsniveau α (in ons geval 0.05), se_{b1} de standaardfout in het intercept en se_{b2} de fout in de helling. Met behulp van formule (4) worden uit de berekende waarden voor de onder- en bovengrens van de lineaire predictor de geschatte onder en bovengrens van het kansinterval berekend.

3.4 Een praktijkvoorbeeld

De dataset betreft het zomergemiddeld chlorofyl-a gehalte in sets van meren (criterium < 50 of 100 μg l^{-1}) versus totaal-fosfaat voor een aantal jaren (Portielje en van den Molen, 1998). De data per jaar zijn niet geheel onafhankelijk, met name de Friese boezemmeren lijken onderling sterk op elkaar. Voor dit voorbeeld is er toch van uitgegaan dat de data binnen een jaar onafhankelijk van elkaar zijn; de data van verschillende jaren zijn dat in ieder geval niet, daar in sommige jaren dezelfde meren bemonsterd zijn.

Eerst tekenen dan rekenen!
Figuur 9 geeft het histogram van de verdeling van alle waarnemingen weer. Deze figuur laat zien dat de meerderheid van de observaties een chlorofyl-a gehalte kleiner dan 40 μg l^{-1} betrof. Slechts weinig waarnemingen lieten een chlorofyl-a gehalte > 200 μg l^{-1} zien. De dataset lijkt dus zeer geschikt om normen te stellen voor de waterkwaliteitsniveaus "chlorofyl-a gehalten kleiner dan 50 of 100 μg l^{-1}", maar is minder geschikt om een nutriëntennorm te stellen gebaseerd op een gewenst chlorofylgehalte van kleiner dan 200 μg l^{-1}.
Bedenk dat voor een norm gebaseerd op de relatie tussen chlorofylgehalte en nutriëntenniveau ook een andere methode gevolgd kan worden: terugzoeken in de grafiek van chlorofyl tegen nutriëntengehalte. Deze methode is in het verleden gebruikt om de CUWVO-normen te bepalen.

Model IV geeft aan dat andere, hierboven al genoemde, factoren kunnen verhinderen dat de gewenste biologische waterkwaliteit überhaupt wordt gehaald, zelfs niet bij relatief lage nutriëntenconcentraties. In de modellen II en III zijn de strengst mogelijke normen de concentraties waarbij in alle gevallen de vereiste biologische kwaliteit wordt behaald (de linker pijl); de minst strenge norm is die concentratie waarbij soms de gewenste kwaliteit wordt behaald (de rechter pijl). De inspanning om het gewenste resultaat te bereiken is minimaal bij de hoogst in aanmerking komende normconcentratie en maximaal bij de laagst in aanmerking komende normconcentratie. De afweging tussen kosten (inspanning) en baten (kleiner risico dat de gewenste kwaliteit niet wordt gehaald) is beleidsmatig. Als de verzamelde data de relatie als aangegeven in model IV het best benaderen is de meest zinvolle norm de rechterpijl. Een verdere verlaging van de nutriëntenconcentraties heeft namelijk weinig effect op de biologische waterkwaliteit. Zinvoller is om in dat geval de tijd en moeite te steken in het achterhalen van de redenen waarom bij lage nutriëntenconcentraties de beoogde kwaliteit niet wordt gehaald.
Vanuit het bovenstaande is het duidelijk dat een nadere analyse van de relatie tussen nutriëntenconcentraties en biologische waterkwaliteit zich moet richten op het concentratiebereik tussen de minst en meest strenge norm als bepaald in model II en III. Deze nadere analyse kan worden gebaseerd op een statistische analyse van al beschikbare of te verzamelen data. Het uitvoeren van een statistische analyse heeft als voordeel dat getoetst kan worden of aan bepaalde voorwaarden wordt voldaan, bijvoorbeeld:

- De kans op het bereiken van het gewenste kwaliteitsniveau (bijvoorbeeld chlorofyl-a < 100 mg l^{-1}) of de gewenste levensgemeenschap (macrofyt gedomineerd) na voldoen aan de normen voor de nutriënten dient zo hoog mogelijk te zijn, d.w.z. dat de kans dat te nemen maatregelen geen succes zullen opleveren aanvaarbaar klein is (bijvoorbeeld 5 of 10%).

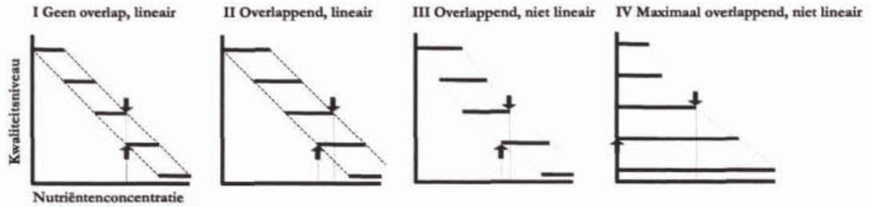
1.3 Statistische methodieken voor gedifferentieerde normstelling.

In het verleden is al op een aantal manieren geprobeerd gedifferentieerde normen op te stellen. Er is zowel gebruik gemaakt van de kennis aanwezig bij experts en of onderzoeken, het beschrijven van empirische relaties en het gebruik van simulatiemodellen. In dit rapport zullen we gebruik maken van het beschrijven van empirische relaties tussen nutriëntenconcentraties en waterkwaliteit. Peeters en Gardeniers (1998a; 1998b) hebben een empirische methode voor gedifferentieerde normstelling t.b.v. aquatische ecosystemen ontwikkeld. Zij hebben voor verschillende subtypen beken en sloten de relatie tussen de samenstelling van de aquatische levensgemeenschappen en de daaraan gerelateerde gegevens over abiotische omstandigheden (bijvoorbeeld stikstof, fosfaat) geanalyseerd. Hiertoe hebben zij uit de dataset verzameld door de waterbeheerders in STOWA-kader de locaties genomen waarbij aan een bepaald kwaliteitsniveau is voldaan (bijvoorbeeld middelste en hoogste niveau). Uit deze sub-datasets zijn verschillende nutriënten-percentielen bepaald waarbij wordt voldaan aan de kwaliteitsniveaus. Zij berekenden bijvoorbeeld dat in 90% van de monsters genomen in de bovenlopen van laagland stromende wateren, waarvoor het hoogste kwaliteitsniveau behaald was, de nitraat en nitriet stikstofconcentratie beneden de 0.47 mg l^{-1} was (Peeters en Gardeniers, 1998b).

kwetsbare andere gebieden, inclusief de Noordzee en de Waddenzee. Sommige provincies zijn daar al eerder mee begonnen (van Liere en Laane, 1993; Otte *et al.*, 1999), waarbij ecologische factoren meegenomen werden. Het nadeel van deze normering is dat er maar beperkt gebruik gemaakt wordt van beschikbare kennis en datasets over de relatie tussen nutriënten concentraties en ecologische kwaliteit; en de differentiatie wordt niet op uniforme wijze wordt uitgevoerd.

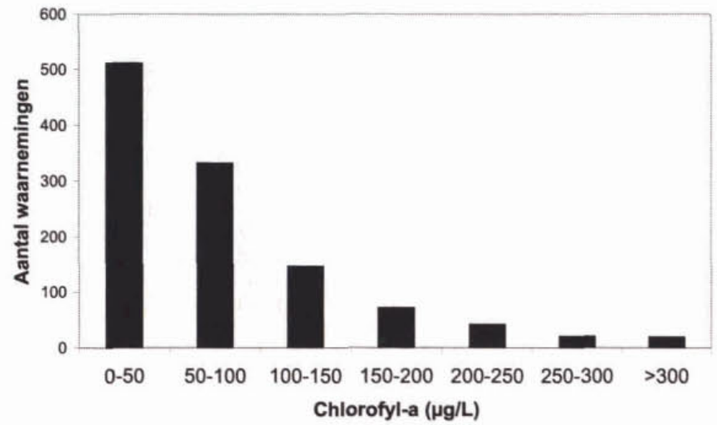
1.2 Gedifferentieerde normstelling naar watertypen en / of gebied

Een gebiedsgerichte en /of watertype gerichte norm voor nutriëntengehalten heeft als voordeel dat rekening kan worden gehouden met het gewenste doel, bijvoorbeeld de minimaal te behalen biotische kwaliteit (basiskwaliteit) of een gewenste levensgemeenschap (macrofyten). De basis van een gedifferentieerde norm is daarom de beschrijving van de specifieke relatie tussen nutriëntengehalten en de biotische kwaliteit voor een gebied of een watertype. Op voorhand is niet bekend hoe deze relatie is omdat deze afhangt van gebieds- en watertype-specifieke kenmerken (bijvoorbeeld waterhuishouding en landgebruik). We kunnen echter wel een aantal theoretische mogelijkheden op een rij zetten. Figuur 1 geeft vier van de vele mogelijke relaties.



Figuur 1: Vier mogelijke modellen voor de beschrijving van de relatie tussen nutriëntenconcentratie en het biologisch kwaliteitsniveau. Ondergrens en bovengrens voor mogelijke normen zijn weergegeven met pijltjes.

In het geval dat er een lineaire, eenduidige relatie is tussen kwaliteitsniveau en nutriëntenconcentratie (model I) is er geen probleem met het stellen van een norm voor nutriënten. Bij het onderschrijden van een bepaalde nutriëntenconcentratie springt het kwaliteitsniveau een niveau omhoog. Als norm kan de concentratie worden gekozen waarbij de biologische kwaliteit omslaat naar het gewenste niveau. In de werkelijkheid treedt echter het probleem op dat meer kwaliteitsniveaus bij een bepaalde concentratie kunnen voorkomen, zodat er niet een vaste concentratie is waarbij de gewenste kwaliteit wordt gehaald. Welk kwaliteitsniveau wordt gehaald is namelijk niet alleen afhankelijk van de concentraties aan nutriënten maar ook van de invloed van andere stoffen (bijvoorbeeld bestrijdingsmiddelen), inrichting en beheer van het gebied, meteorologische omstandigheden etc. Model II geeft een relatie aan waarbij het omslagpunt tussen kwaliteitsniveaus niet zo scherp is; er is een concentratiebereik waarbij twee (of eventueel meer) kwaliteitsniveaus kunnen optreden. Model III geeft aan dat er geen lineaire relatie hoeft te bestaan tussen het behaalde kwaliteitsniveau en de concentratie van de nutriënten; de overlap tussen kwaliteitsniveaus kan afnemen (of toenemen) met de nutriëntenconcentratie.



Figuur 9: Histogram van alle data van het praktijkvoorbeeld.

Met behulp van de Excel-file zijn normen voor fosfaatconcentraties berekend voor de afzonderlijke jaren voor twee gewenste waterkwaliteitsniveaus (chlorofyl-a gehalte is kleiner dan 50 en 100 µg l^{-1}, Tabel 4). Bij deze berekeningen is gebruik gemaakt van een kans op het behalen van de waterkwaliteit van 80% (kans op niet behalen is 0.2).

Tabel 4: Fosfaatnormen berekend met behulp van de Excel-file voor twee waterkwaliteitsniveaus voor alle jaren afzonderlijk gebruik makend van een kans op behalen van de waterkwaliteit van 0.8.

Jaar	Aantal waarnemingen	Fosfaatnorm (mg l^{-1}) chl-a < 50 µg l^{-1}	Fosfaatnorm (mg l^{-1}) chl-a < 100 µg l^{-1}
1990	169	0.049	0.122
1991	155	0.034	0.168
1992	186	0.051	0.148
1993	179	0.020	0.273
1994	163	0.039	0.192
1995	162	0.065	0.202
1996	135	0.073	0.240

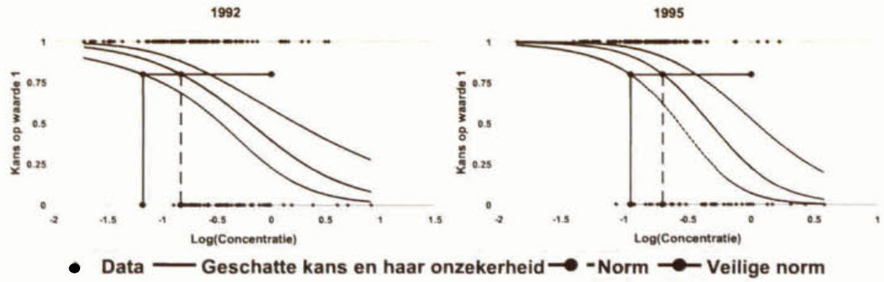
De normen verschillen maximaal een factor 3.5 tussen de jaren voor het waterkwaliteitsniveau "chlorofyl-a < 50 µg l^{-1}" en ongeveer een factor 2 voor het waterkwaliteitsniveau "chlorofyl-a < 100 µg l^{-1}". De verhouding tussen de normen voor de twee kwaliteitsniveaus varieert tussen 2 en 14. Het voorbeeld geeft aan dat de keuze van het gewenste kwaliteitsniveau en de selectie van te gebruiken data van grote invloed kunnen zijn op de resulterende norm. Het is dus van groot belang om voor het vaststellen van de norm goed na te denken over het kwaliteitsniveau dat gewenst is en de gegevens die voor het bepalen van de norm gebruikt worden. Als er geen statistiekprogramma beschikbaar is maar wel veel gegevens aanwezig zijn, kan door op basis van meer deelsets normen te berekenen een indruk verkregen worden van de betrouwbaarheid van de norm.

Het gebruik van alle gegevens tegelijk in dit voorbeeld is niet aan te raden, daar waarschijnlijk zowel in de tijd als in de ruimte te dicht bemonsterd is. Een gevolg hiervan is ook dat het geschatte betrouwbaarheidsinterval mogelijk veel te smal is.

Tabel 5: Fosfaat normen en veilige normen bij verschillende kansen op het niet behalen van de waterkwaliteit voor de datasets van 1992 en 1995. Het gewenste kwaliteitsniveau is chlorofyl-a < 100 µg l'.

	1992		1995	
p-waarde	Norm	Veilige norm	Norm	Veilige norm
0.05	0.029	0.005	0.078	0.033
0.10	0.063	0.019	0.123	0.061
0.15	0.102	0.039	0.164	0.087
0.20	0.148	0.066	0.202	0.112
0.25	0.200	0.099	0.241	0.137
0.30	0.261	0.137	0.281	0.161
0.35	0.331	0.180	0.324	0.186
0.40	0.415	0.228	0.369	0.212
0.45	0.515	0.282	0.418	0.239
0.50	0.636	0.344	0.472	0.268

Tabel 5 laat zien dat de fosfaatnormen tussen de jaren relatief veel verschillen wanneer een lage kans op niet behalen van de waterkwaliteit wordt gekozen ($p < 0.2$). De verschillen worden veel kleiner bij een kans op niet behalen van de waterkwaliteit van groter dan 0.2. Ook het verschil tussen de norm en de veilige norm wordt kleiner bij een hogere kans op niet behalen van de waterkwaliteit. Dit is het gevolg van het feit dat de sigmoide curve vlakker loopt nabij de extreme waarden 0 en 1 en zijn maximale steilheid bereikt bij p=0.5.



Figuur 10: Overzicht van de 1992 en 1995 datasets. Tevens zijn de normen en veilige normen voor een kans op niet behalen van de waterkwaliteit van 0.2 aangegeven.

1 Inleiding

1.1 Normering

Vanaf de Indicatieve Meerjaren Programma's Water (VROM, 1981) werden relaties tussen menselijk handelen en de oppervlaktewaterkwaliteit met name bezien vanuit het ecologisch perspectief. Een basiskwaliteit voor zoete water werd voorgesteld, met daaraan gekoppelde waarden voor fosfor en stikstof, daar waar het om eutrofiëring handelde. Om een indruk te geven van de toenmalige visie: voor fosfor werden waarden van 0.20 – 0.40 mg l^{-1} voorgesteld; in de normenlijst voor stagnante wateren werd 0.20 mg P l^{-1} als norm opgenomen. Voor het benaderen van meer natuurlijke fosforgehaltes werd een streefwaarde van < 0.10 mg P l^{-1} voorgesteld. Ook de IMP's wezen reeds op het belang van aanvullende maatregelen. Deze visie leidde tot meerdere eutrofiëringsprojecten, waarvan de belangrijkste m.b.t. normstelling wel de verschillende CUWVO (CIW) eutrofiëringsenquêtes waren (voor een overzicht zie Hosper, 1997 & Portielje en van der Molen, 1998).

Normstelling voor ondiepe stagnante wateren berust op analyses van een groot aantal ondiepe meren. In 1980 werden, bij de aanname dat bij 100 µg l^{-1} chlorofyl *a* het doel (helder water) bereikt zou zijn, MTR's van 0.15 mg P l^{-1} en 2.2 mg N l^{-1} (zomergemiddeld) afgeleid (CUWVO, 1980). De toenmalige meren in de database (<1980) werden vooral door groenwieren gedomineerd. In de derde eutrofiëringsenquête (CUWVO, 1987) waren inmiddels al een groot aantal meren met blauwwierdominantie en er werd bij 100 µg l^{-1} chlorofyl *a* voor deze meren een waarde van 0.07 mg P l^{-1} afgeleid; voor N was er geen reden tot wijziging. Echter, dit verschil had geen gevolgen voor de "officiële" normstelling. Voor andere wateren werd in de Derde Nota waterhuishouding ook de MTR voor P op 0.15 mg P l^{-1} (jaargemiddelde) gesteld. Wetenschappelijk onderbouwing daarvan ontbreekt tot nu toe. De IRC nam deze waarde voor de Rijn over. Voor stikstof werden voor andere wateren geen beslissingen genomen (al dan niet onderbouwd), noch door de Nederlandse overheid, noch door de IRC; maar de waarde 2.2 mg N l^{-1} is een eigen leven gaan leiden, en andere watertypen werden hieraan ook regelmatig (maar dus ten onrechte) getoetst. Schreurs (1992) leidde eveneens af dat dominantie van blauwwieren slechts voor kon komen bij zomergemiddelde waarden van fosfor <0.06 mg P l^{-1}. Ook na de Vierde eutrofiëringsenquête (Portielje & van der Molen, 1998) werd in de Vierde Nota Waterhuishouding de MTR van 0.15 mg P l^{-1} en 2.2 mg N l^{-1} (zomergemiddelden) gehandhaafd. Echter nu alléén voor die wateren waarvoor het bestemd en onderbouwd was (stagnante eutrofiëringsgevoelige wateren). Tevens werd eens streefwaarde (zomergemiddelde) ingevoerd wanneer men werkelijk de eutrofiëring wilde bestrijden: 0.05 mg P l^{-1} en 1 mg N l^{-1}. Deze waarden werden richtinggevend voor andere watertypen vanwege de afwentelingsproblematiek. Daarnaast stelt de Vierde Nota Waterhuishouding mogelijkheden voor om voor andere watertypen in verschillende gebieden gedifferentieerd te normeren, zowel hoger als lager moet mogelijk zijn, afhankelijk van de functie van het gebied, mits er maar geen afwenteling plaatsvindt naar

4 Literatuur

Bio, A.F.M. 2000. Does vegetation suit our models? : data and model assumptions and the assessment of species distribution in space. Proefschrift Universiteit Utrecht, Utrecht.

CIW/CUWVO. In druk. Relaties tussen concentraties en ecologische kwaliteit – methoden per watertype. Onderbouwing voor gedifferentieerde normstelling voor nutriënten.

CUWVO, 1980. Ontwikkeling van grenswaarden voor doorzicht, chlorofyl, fosfaat en stikstof. Coördinatiecommissie uitvoering wet verontreiniging oppervlaktewater. Resultaten van de tweede eutrofiëringsenquête.

CUWVO, 1987. Vergelijkend onderzoek naar de eutrofiëring in Nederlandse meren en plassen. Coördinatiecommissie uitvoering wet verontreiniging oppervlaktewater. Resultaten van de derde eutrofiëringsenquête.

CUWVO, 1988. Ecologische normdoelstellingen voor Nederlandse oppervlaktewateren. CUWVO, 's-Gravenhage.

Hosper, H., 1997. Clearing lakes, an ecosystem approach to the restoration and management of shallow lakes in the Netherlands. DSc. Thesis Agricultural University at Wageningen.

Liere, L. van en W. Laane. 1993. Streven naar (streef)waarden van het zoete oppervlaktewater in Nederland. In: Boers, P., W. Laane, L. van Liere, C. Peeters, S. Parma en J. van der Does. Eutrofiëring en beleid in Nederland, hoe verder? RIZA notitie 93056X, DGW nota 93.007, RIVM rapport 732404002.

Ministerie van Volkshuisvesting, Ruimtelijke Ordening en Milieubeheer. 1981. Indicatief Meerjaren Programma Water 1980-1984. SDU, Den Haag.

Ministerie van Verkeer en Waterstaat. 1998. Vierde nota waterhuishouding: regeringsbeslissing. Ministerie van Verkeer en Waterstaat, 's-Gravenhage.

Otte A.J., S.G, Vermeij en F. Heinis, 1999 Watertypegerichte normstelling voor nutriënten, en toepassing op meren en plassen. Aquasense rapportnummer 99.1221.

Oude Voshaar, J.H. 1994. Statistiek voor onderzoekers, met voorbeelden uit de landbouw- en milieuwetenschappen. Wageningen Pers.

Peeters, E.T.H.M. en J.J.P. Gardeniers, 1998a. Vereenvoudiging van de gedifferentieerde milieukwaliteit van oppervlaktewater in Fryslân. Leerstoelgroep

Aquatische ecologie en waterkwaliteitsbeheer, Wageningen Universiteit. Rapport M284.

Peeters, E.T.H.M. en J.J.P. Gardeniers. 1998b. Aanzet tot gedifferentieerde grens- en streefwaarden voor nutriënten in regionale wateren. H2O 2: 16-20.

Portielje, R. en D.T. van der Molen (1998). Relaties tussen eutrofiëringsvariabelen en systeemkenmerken van de Nederlandse meren en plassen. Deelrapport 2 van de Vierde Eutrofiëringsenquête. RIZA rapport 98.007.

Schreurs, H., 1992. Cyanobacterial dominance, relation to eutrophication and lake morphology. DSc thesis University of Amsterdam.

Tongeren, O.F.R. van en N.J.M. Gremmen. 1999. Gebruik van logistische regressie voor het vaststellen van de relatie tussen waterkwaliteit, nutriënten en beheer. Data-Analyse Ecologie, Westervoort.

Samenvatting

In het begin van de jaren 90 zijn provincies gestart met het opstellen van gebiedsgerichte en watertypengerichte normen. Deze gedifferentieerde normen werden tot op heden nog niet op een uniforme manier uitgevoerd en er werd nog maar beperkt gebruik gemaakt van de beschikbare data. Dit rapport is bedoeld als een inleiding tot het gebruik van logistische regressie voor gedifferentieerde normstelling. Logistische regressie heeft als voordeel boven veel andere methoden dat het gebruik kan maken van een veelheid aan beschikbare data en het verschillende keuzemogelijkheden wat betreft zekerheid van behalen van de kwaliteitsdoelstelling toelaat. De theorie van logistische regressie wordt in dit rapport uitgebreid besproken. Als uitgangspunt is genomen dat in het algemeen bij lage nutriëntenconcentraties aan de gewenste waterkwaliteit (bijvoorbeeld de aanwezigheid van hogere waterplanten) voldaan wordt en bij hoge nutriëntenconcentraties niet. Stapsgewijs wordt uitgelegd hoe de s-vormige relatie tussen behalen van de waterkwaliteit en de concentraties van nutriënten in het water kan worden beschreven. Tevens worden verschillende eisen waaraan de invoergegevens moeten voldoen, zoals onafhankelijkheid en representativiteit van de monsters, besproken, alsmede de gevolgen hiervan voor de verzameling van de gegevens en haar voorbewerking. Verschillende keuzes moeten a-priori bij de afleiding van de norm worden gemaakt. Ten eerste moet worden bepaald welke kans dat de waterkwaliteit wordt gehaald gewenst is. Tevens kan de zekerheid dat deze kans behaald wordt in de normstelling betrokken worden. Dit rapport bevat een volledige uitwerking van een theoretisch en praktijkvoorbeeld in Excel en SPSS. Tevens zijn twee bijlagen opgenomen met complexe voorbeelden.

Bijlage 1. Optimummodel

De gegevens voor dit voorbeeld zijn gegevens betreffende het voorkomen van blauwalgen in Nederlandse meren (Portielje en van der Molen, 1998).
Van een zestigtal meren is het zomergemiddelde bepaald van de dominantie van blauwalgen als percentage van het totaal aantal algen. De (arbitraire) kwaliteitseis die aan deze meren gesteld wordt is dat het aantal blauwalgen minder dan 30% bedraagt van het totaal aantal algen.
Deze kwaliteitseis wordt vergeleken met de zomergemiddelden van totaal P en totaal N. Daar zowel P als N scheef verdeeld waren, is allereerst het zomergemiddelde van totaal P en totaal N logaritmisch getransformeerd. Let op: doordat de gegevens aangeleverd zijn als zomergemiddelde was dit de enige mogelijkheid. Een nettere oplossing is om de oorspronkelijke gehalten logaritmisch te transformeren en vervolgens het gemiddelde te bepalen.
In een apart Excel spreadsheet voor complexe logistische regressie zijn vervolgens verscheidene modellen doorgerekend. De analyse had ook plaats kunnen vinden met behulp van een statistisch pakket, maar ons inziens is het ontbreken van betrouwbaarheidsintervallen van ondergeschikt belang, daar zeer veel arbitraire beslissingen genomen zijn, die van veel groter invloed zijn op het eindresultaat (bijvoorbeeld de kwaliteitseis van 30%). Tabel 1 geeft een overzicht van de resultaten van de analyses.

Tabel 6: Overzicht van de resultaten van de analyses. Zie tekst voor verklaring.

	Model	Deviance model	Verschil met het model	Extra verklaarde Deviance	Extra vrijheids-graden	p-waarde extra verklaarde deviance
0	Null	79,89	-	-	-	-
1	lineair P	78,79	0	1,10	1	0,30
2	lineair N	79,72	0	0,16	1	0,69
3	kwadratisch P	73,21	0	6,67	2	0,036
3	kwadratisch P	73,21	1	5,57	1	0,018
4	kwadratisch N	73,46	0	6,43	2	0,040
4	kwadratisch N	73,46	2	6,27	1	0,012
5	kwadratisch P en N	70,52	3	2,69	2	0,26
5	kwadratisch P en N	70,52	4	2,93	2	0,23
6	kwadratisch P en N met interactie	66,44	5	4,08	1	0,043

In de kolom 'model' is aangegeven welk model doorgerekend is. Het null-model is het model dat de afwezigheid van een relatie beschrijft (horizontale lijn). De lineaire modellen zijn de in de hoofdtekst beschreven modellen: logit(verwachte kans) tegen een nutriëntconcentratie. Bij de kwadratische modellen is het verband met het kwadraat van de (log getransformeerde) nutriëntconcentratie toegevoegd. Na terugtransformatie van de op deze wijze gefitte parabool van logit(verwachte kans)

tegen log(nutriëntconcentratie) naar kansen zien we een klokvormige curve, die sterk lijkt op die van de normale verdeling. De modellen met beide nutriënten zijn ingewikkelder. Model 6, met interactie) wordt in het onderstaande aan de hand van grafieken nader besproken.
Met opzet zijn in de tabel nog geen regressiecoëfficiënten vermeld, daar deze tabel voor een eerste indruk ruim voldoende informatie biedt.:

- De kolom "Deviance model" geeft de som van –2ln(likelihood) voor elk van de modellen. Dus hoe kleiner de Deviance, hoe beter het model de data beschrijft.
- De kolom "Verschil met het model" geeft aan met welk ander model het model in die rij vergeleken wordt ter toetsing.
- De kolom "Extra verklaarde deviance" geeft het verschil tussen de devianties van de modellen, en is dus een maat voor welk model de data beter beschrijft.
- De kolom "Extra vrijheidsgraden" geeft aan hoeveel extra parameters er voor het ingewikkelder model van de twee onderling vergeleken modellen extra geschat zijn
- De kolom "p-waarde extra verklaarde deviance" tenslotte geeft de kans dat de Chi-kwadraat verdeling met het desbetreffende aantal vrijheidsgraden groter of gelijk is aan de waarde van "Extra verklaarde deviance". Als deze kans klein is (bijvoorbeeld kleiner dan de gebruikelijke 0.05) is het model significant.

Het eerste dat opvalt, is dat er geen significant verband lijkt te zijn tussen de nutriëntenconcentraties en de dominantie van blauwalgen: zowel het lineaire model voor P (model 1) als het lineaire model voor N (model 2) zijn niet significant. Zodra we echter een kwadratische term toevoegen wordt het verband significant voor beide nutriënten (modellen 3 en 4, zowel vergeleken met model 0 als met 1 respectievelijk 2). Dit betekent dat er een verband gevonden is met een optimum (maximum) kans op voorkomen of met een minimum kans op voorkomen.
Toevoegen van N aan het model met P of van P aan het model met N levert geen significante verbetering op (model 5 vergeleken met 3 en 4). Dit wordt veroorzaakt door een duidelijke correlatie tussen de N en de P-gehalten. We kunnen het effect van N niet scheiden van dat van P, we weten dus niet of er een oorzakelijk verband is tussen de dominantie van blauwalgen en P-totaal of N-totaal of beiden. P-totaal en N-totaal kunnen elkaar in statistische zin vervangen. Het verstandigst is in zo'n geval de norm voor beiden te bepalen en te eisen dat aan beide normen voldaan is. Er is dan vrij grote zekerheid dat de kwaliteitsdoelstelling bereikt zal worden.
Het meest ingewikkelde model met de interactieterm komt niet in aanmerking, evenals het kwadratische model met P en N, omdat er geen significante daling is van de deviance ten opzichte van het kwadratische model met alleen P of N.
In het voorbeeld in bijlage 2 komt interactie nader aan de orde.

Tabel 7: Regressiecoëfficiënten voor de kwadratische modellen voor P en N

	Intercept	lineair	kwadratisch
P	2,05	2,74	0,90
N	3,26	-5,19	2,06

Tabel 2 geeft de regressiecoëfficiënten van de kwadratische modellen. Het geschatte N-model wordt bijvoorbeeld: logit(kans) = 3.26 - 5.19*ln(N-conc.) + 2.06*(ln(N-conc.))^2. Wanneer we van deze regressiemodellen de voorspelde kansen uitzetten

Ten geleide

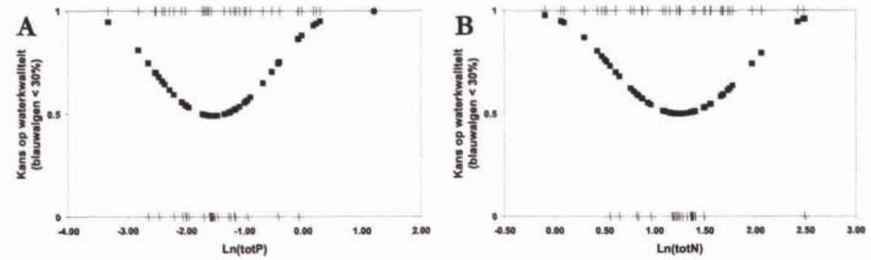
In de Vierde Nota Waterhuishouding wordt ruimte gegeven voor een meer flexibele omgang met normen (Ministerie van Verkeer en waterstaat, 1998). Voor niet-eutrofiëringsgevoelige wateren kan van de huidige MTR-waarden voor stikstof en fosfaat worden afgeweken. Een voorwaarde is wel dat in ieder geval aan het 'laagste ecologische niveau' zoals geformuleerd door de CUWVO (1988) voldaan dient te worden. Binnen de werkgroep V van de Commissie Integraal Waterbeheer is de sub-werkgroep "gedifferentieerde normstelling nutriënten in oppervlaktewater" verantwoordelijk voor de ontwikkeling van een methodiek voor het afleiden van gedifferentieerde normen op gebiedsniveau. Een van de methodieken voor gedifferentieerde normstelling die is voorgesteld door deze sub-werkgroep is logistische regressie (CIW, in druk). Dit rapport bevat zowel de theorie als een gebruikershandleiding over het toepassen van logistische regressie voor gedifferentieerde normstelling. Dit rapport bevat tevens een CD-rom waarop bestanden staan, waarmee de logistische regressie kan worden uitgevoerd binnen Excel en SPSS.
Het onderzoek is uitgevoerd door Onno F.R. van Tongeren (Data-analyse Ecologie) en Paul J. van den Brink van Alterra, Research Instituut voor de Groene Ruimte. Het rapport is mede tot stand gekomen dankzij de inbreng van Ary Roos (VROM), Floor Heinis (Heinis Waterbeheer en Ecologie), Rob Portielje (RIZA) en Niek Gremmen (Data-Analyse Ecologie). De STOWA dankt hen voor hun inzet.

Utrecht, juni 2001

De directeur van de STOWA,

Ir. J.M.J. Leenen

tegen de nutriëntconcentraties zien we in beide gevallen een minimumcurve (figuur 1 en 2). We verwachten dus op grond van deze regressiemodellen dat zowel bij hoge als bij lage nutriëntconcentratie aan de kwaliteitseis is voldaan. De geschatte curven zijn het complement van optimumcurven voor blauwalgen. Daar zeer hoge nutriëntconcentraties andere negatieve gevolgen zullen hebben, kiezen we de normen in het dalende deel van de minimumcurven.



Figuur 11: Het verband tussen de natuurlijke logaritme uit de concentratie totaal P (A) en totaal N (B) en het voldoen aan de kwaliteitseis dat blauwalgen minder dan 30% moeten uitmaken van het totaal aantal algen.

De teruggerekende normen vindt u in tabel 3.

Tabel 8: Normen voor totaal P en totaal N afgeleid uit de gegevens betreffende dominantie van blauwalgen. p(kwal) is de kans dat aan kwaliteitseis voldaan is, p(dom) is de kans dat blauwalgen dominant worden (1-p(kwal)). Normen zijn gegeven als natuurlijke logaritme van de concentratie en als concentratie.

		N		P	
p(kwal)	p(dom)	Ln(norm)	Normconc.	Ln(norm)	Normconc.
0.5	0.5	1.26	3.52	-1.73	0.18
0.8	0.2	0.44	1.55	-2.79	0.06

Het zal geen verbazing wekken dat deze niet veel afwijken van de thans geldende normen die immers gesteld zijn op basis van gegevens betreffende meren.

Bijlage 2. Interactie met beheer

Dit ingewikkelder voorbeeld is weer geconstrueerd, daar goede gegevens om dit te illustreren niet beschikbaar waren. Het genereren van de gegevens is in principe op dezelfde wijze gebeurd als voor het voorbeeld in paragraaf 3.1. Het model is echter wat ingewikkelder:
Logit(p(kwal) = 0 + 1 x log(conc) – 0,5 x $\log^2$(conc) + 4 x Beheer – 2 x Beheer x log(conc) + 0,5 x Beheer x $\log^2$(conc)
De concentraties zijn getrokken uit een lognormale verdeling, waarvan na logtransformatie het gemiddelde gelijk is aan 3 en de standaarddeviatie aan 1,4. Door middel van random getallen is aan 30% van de monsters de waarde 1 voor de beheersvariabele toegekend en aan 70% van de monsters de waarde 0.

Tabel 9 geeft de regressiecoëfficiënten van verschillende geschatte modellen. Het eerste deel (model 1 t/m 6) geeft de resultaten als stapsgewijs de parameters in de volgorde intercept, Nut1, Nut1^2, Beh, Beh*Nut1, Beh*Nut1^2 toegevoegd worden. De tweede rij geeft dezelfde output alleen zijn de parameters in een andere volgorde toegevoegd: intercept, Nut1, Beh, Beh*Nut1, Nut1^2, Beh*Nut1^2. Aangezien dezelfde parameters toegevoegd worden is het eindresultaat natuurlijk hetzelfde (vergelijk model 11 en 6). De nummers 12 t/m 17 zijn gebaseerd op de resultaten van de schattingen met weglating van steeds één van de variabelen. Het berekende verschil in deviantie met het volledige model geeft dus de vermindering in deviantie als de parameters als laatste aan het volledige model worden toegevoegd (bijvoorbeeld model 17 geeft de Deviance voor het model waar alle parameters inzitten behalve Beh*Nut1^2. De kolom "verschil" geeft de verlaging in deviance aan als deze parameter wel als laatste toegevoegd wordt. Uiteindelijk onstaat dus in alle gevallen het volledige model als gegeven bij nummer 6 en 11.

Inhoud

Tabel 9. Resultaten van de regressie van het verband tussen de waterkwaliteit en het uitvoeren van een beheersmaatregel en de concentratie van een nutriënt. In het eerste en het tweede deel van de tabel staan de regressiecoëfficiënten bij stapsgewijs toevoegen van variabelen, de door deze variabelen extra verklaarde deviantie en de significanties. In het derde deel van de tabel staan de significanties van de variabelen wanneer ze als laatste aan het volledige model worden toegevoegd.

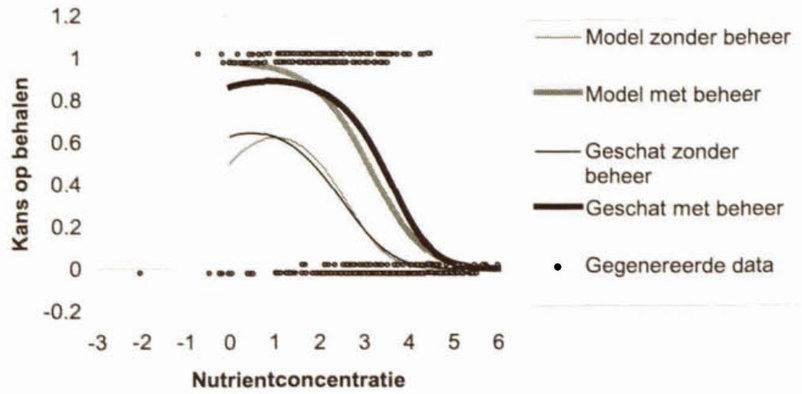
Model	Intercept	Nut1	Nut1^2	Beh	Beh*Nut1	Beh*Nut1^2	Deviance	Verschil	p-waarde
1	-0,69	0,00	0,00	0,00	0,00	0,00	637		
2	2,12	-1,01	0,00	0,00	0,00	0,00	493	144	4,49E-33
3	0,85	0,34	-0,29	0,00	0,00	0,00	473	21	5,54E-06
4	0,45	0,27	-0,30	1,96	0,00	0,00	414	59	1,36E-14
5	0,53	0,32	-0,33	1,33	0,25	0,00	413	1	3,72E-01
6	-0,08	1,28	-0,59	5,09	-3,16	0,68	401	11	7,85E-04

Model	Intercept	Nut1	Beh	Beh*Nut1	Nut1^2	Beh*Nut1^2	Deviance	Verschil	p-waarde
7	2,12	-1,01	0,00	0,00	0,00	0,00	493	144	4,49E-33
8	1,73	-1,12	1,88	0,00	0,00	0,00	434	59	1,41E-14
9	1,50	-1,02	2,69	-0,29	0,00	0,00	433	1	2,36E-01
10	0,53	0,32	1,33	0,25	-0,33	0,00	413	20	7,31E-06
11	-0,08	1,28	5,09	-3,16	-0,59	0,68	401	11	2,46E-08

Model	Zonder	Deviance	Verschil	p-waarde	Significant
12	Intercept	401	0,03	8,69E-01	n.s.
13	Nut1	409	7,22	7,20E-03	*
14	Beh	416	14,11	1,73E-04	*
15	Beh*Nut1	410	8,55	3,46E-03	*
16	Nut1^2	433	31,09	2,46E-08	*
17	Beh*Nut1^2	413	11,28	7,85E-04	*

In bovenstaande tabel zien we direct dat de verklaarde deviantie afhankelijk is van de volgorde waarin de variabelen in het model worden opgenomen (vergelijk model 1 t/m 6 met 7 t/m 11) . Het meest opvallend is dit voor de interactie Beheer*Nutriënt (Beh*Nut1), die in het eerste deel van de tabel (model 1 t/m 6) significant is en in het tweede deel (model 7 t/m 11) niet. Uitsluitsel over de significantie van de verschillende variabelen wordt pas gegeven in het derde deel van de tabel, waaruit blijkt dat alle variabelen en interacties een significante bijdrage leveren, wanneer ze als laatste aan het model worden toegevoegd.
Figuur 12 laat zien dat het geschatte regressiemodel voor 500 random waarnemingen (zwarte lijnen) erg lijkt op het geconstrueerde model (grijze lijnen). De betekenis van de term interactie in statistische zin kan aan de hand van deze figuur ook zeer duidelijk gemaakt worden: het effect van beheer en nutriëntconcentratie kan niet zomaar opgeteld worden. De beheersmaatregel zorgt niet alleen voor verhoging van de kans op behalen van de vereiste biologische kwaliteit, maar zorgt er tegelijk voor dat het effect van de nutriëntconcentratie verandert, het geschatte optimum verschuift als gevolg van de beheersmaatregel naar hogere nutriëntconcentraties en de vorm van de curve verandert ook.

Uit deze figuur blijkt tevens dat niet uitvoeren van de beheersmaatregel maximaal leidt tot ongeveer 65% slagingskans bij een concentratie van 1,1 eenheden op logaritmische schaal, de nutriëntconcentratie is dan 2,9. Bij uitvoeren van de beheersmaatregel kan de norm voor 80% slagingskans gezet worden bij Ln(nutriëntconcentratie) gelijk aan 2,14 eenheden. Dit betekent een nutriëntconcentratie van 8,5. Een norm zonder uitvoeren van de beheersmaatregel kan alleen gesteld worden met een betrekkelijk lage kans van slagen. Voor 50% ligt kan deze gesteld worden op concentratie 8 (2,08 op logaritmische schaal).



Figuur 12. Gesimuleerde (grijze lijnen) en geschatte (zwarte lijnen) relatie tussen Ln(nutriëntconcentratie) en de kans op behalen van de vereiste biologische kwaliteit. Monsterpunten waar de vereiste kwaliteit behaald is zijn weergegeven rond de waarde 1, monsterpunten waar deze niet behaald is rond de waarde 0. De iets hoger geplaatste punten zijn die waar de beheersmaatregel is uitgevoerd, bij de lager geplaatste punten is de beheersmaatregel niet uitgevoerd.

.

stowa

Stichting Toegepast Onderzoek Waterbeheer

Het gebruik van logistische regressie voor gedifferentieerde normstelling

Een analyse van de relatie tussen nutriënten, beheer en biologische waterkwaliteit

Onno F.R. van Tongeren
Paul J. van den Brink

2001 16

Arthur van Schendelstraat 816
Postbus 8090, 3503 RB Utrecht
Telefoon: 030 - 232 11 99
Fax: 030 - 232 17 66
E-mail: stowa@stowa.nl
http://www.stowa.nl

Publicaties en het publicatie-overzicht
van de STOWA kunt u uitsluitend bestellen bij:
Hageman Fulfilment
Postbus 1110
3300 CC Zwijndrecht
Telefoon: 078 - 629 33 32
fax: 078 - 610 42 87
E-mail: hff@wxs.nl
o.v.v. ISBN- of bestelnummer
en een duidelijk afleveradres.

ISBN 90-5773-128-2

stowa

ting Toegepast Onderzoek Waterbeheer

Het gebruik van logistische regressie voor gedifferentieerde normstelling

Een analyse van de relatie tussen nutriënten, beheer en biologische waterkwaliteit

2001 16