RIVM rapport 408657006/2002

MOVE nationaal Model voor de Vegetatie versie 3.2
Achtergronden en analyse van modelvarianten

M. Bakkenes, D. de Zwart en J.R.M. Alkemade

Dit onderzoek werd verricht in opdracht en ten laste van de directie van het RIVM, in het kader van project Ecologische Modellering, nr. S/408657/01/EM

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

Abstract

This report describes in two parts the process for producing optimal regression equations for vegetation response models for the third version of MOVE, a Dutch national model for MOdelling VEgetation responses using regression models. Part I presents an analysis of the available data and Part II the derivation of the optimal regression equations.

Values of all variables were found for the complete data range and the samples showed a good spatial distribution. Comparing a random sample of 13 species with records in the Dutch standard reference book, *'Heukels' Flora van Nederland'*, we found similar values and dispersion for these species. Optimal regression equations for each species were constructed using two methods. The first method compared up to 11 different regression models, starting with a simple model and creating new and more complex models by adding variables. The second method, a stepwise regression analysis, has, as the only control, the variation in start and end models of the stepwise regression. In this second method we experimented with six different variants of which three were used in the final selection procedure, bringing the total number of calculated models to 14 for each species.

Of the original 914 species, only species (models) meeting two criteria were selected. First, the measure of goodness of fit had to fall within the 5% ($\alpha = 0.05$) confidence interval, which means a possibility of 5% that a model will be wrongly rejected. When more than one model meets this criterion the model with the largest predictor will be selected. The second criterion is that at least one of the variables in the model must be able to change over a period of time. The use of these criteria led to models for 690 species. The goodness of fit criteria resulted in poor models for 15 species. Poor models are those with a relatively low predictor compared with the highest predictor in the set of the 14 different models. For these species a better model is selected, i.e. a model with a much higher predictor and a goodness of fit with a $p > 0.01$. All 14 different model runs are included in the final set. The variables with the fewest occurrences are those describing the impact of heavy metals, the presence of salt in the soil and vegetation type. Apparently, these variables are less important determining variables for many plant species.

Voorwoord

In het kader van het Milieu- en Natuurplanbureau werken het RIVM en Alterra samen met het RIZA en RIKZ aan de ontwikkeling van een gemeenschappelijk planbureauinstrumentarium waarvan de 'Natuurplanner' een voorbeeld is. De Natuurplanner bestaat uit de modellen SMART-SUMO, MOVE, LARCH en BIODIV. Met de Natuurplanner kunnen uitkomsten van speciefieke beleidsopties doorgerekend en geanalyseerd worden. Een belangrijk onderdeel is het doorrekenen van de natuurwaarde en –kwaliteit onder invloed van veranderingen in milieu, ruimtelijke inrichting en beheer.
Dit rapport behandelt één onderdeel van het Natuurplannerinstrumentarium, de plantenmodule. De eerste versie van de plantenmodule MOVE is in 1992 gemaakt. MOVE 1 beschrijft de ecologische relatie tussen drie milieufactoren en de kans op voorkomen. Bij de afleiding van deze relaties is een vegetatie opnamebestand van 15.000 records gebruikt (Schaminée *et al.*, 1995). De tweede versie van MOVE (1997) is gebaseerd op circa 30.000 opnamen. De derde versie van MOVE (De Heer *et al.*, 2000) is gebaseerd op een dataset van circa 100.000 opnamen. Naast de uitbreiding van het aantal opnamen is in de derde versie ook het aantal variabelen uitgebreid, zodat het nu ook mogelijk is naast de effecten van verzuring, verdroging en vermesting, effecten door te rekenen van verzilting en toxiciteit als gevolg van de aanwezigheid van zware metalen. Dit rapport behandelt de verbetering en optimalisatie van MOVE 3 (De Heer *et al.*, 2000) tot versie 3.2. In de eerste versie van MOVE 3 zijn er voor 914 plantensoorten responsiekrommen afgeleid, terwijl in de uiteindelijke versie die in dit rapport beschreven wordt er uiteindelijk nog maar 690 modellen (planten) overgebleven zijn. Alleen die soorten zijn geselecteerd waarvan het afgeleide model een goede overeenkomst heeft met de ruimtelijk verdeelde waarnemingspunten. Achtergrond-informatie en bijlagen bij dit rapport staan in RIVM rapport 408657006 (Bakkenes *et al.*, 2002).

Michel Bakkenes

Inhoud

Samenvatting

Dit rapport beschrijft het optimaliseren van de regressievergelijkingen voor MOVE 3.2. MOVE is een regressieMOdel voor de Nederlands VEgetatie. Het onderzoek is in twee fase uitgevoerd, allereerst is in deel I de geschiktheid van de dataset geanalyseerd en vervolgens (deel II) zijn de optimale regressievergelijkingen afgeleid.

Uit de analyse in deel I is geconcludeerd dat de dataset geschikt is om te gebruiken voor het afleiden van de regressievergelijkingen. Er zijn voor het complete bereik van de variabelen gegevens aanwezig. De ruimtelijke spreiding van de gegevens over Nederland is redelijk homogeen, dit is een belangrijke voorwaarde voor het kunnen gebruiken van deze gegevens voor regressie analyse. Bij het vergelijken van de waarden van de te gebruiken variabelen met 13 willekeurig getrokken plantensoorten blijkt dat de gemiddelde waarden en de spreiding van deze variabelen goed overeenkomen met vermeldingen in Heukels' Flora van Nederland voor deze soorten.
Er is op twee manieren geprobeerd om de optimale regressievergelijking (deel II) per plantensoort af te leiden. Bij de eerste methode is het aantal vrijheidsgraden van op voorhand opgelegd en worden specifieke modellen doorgerekend. In totaal zijn er op deze wijze elf verschillende modellen doorgerekend. Bij de tweede methode, de stapsgewijze regressie methode, bepaalt het rekenproces in grote mate welke modellen doorgerekend worden. Door het aangeven van het startmodel en het eindmodel is het mogelijk om enige sturing over het proces te houden. Bij de tweede methode zijn zes varianten doorgerekend. Drie varianten leverden nooit een beter model op, dus zijn er uiteindelijk drie varianten verder geanalyseerd. Vervolgens is per soort het beste model geselecteerd uit de in totaal veertien verschillende modellen.

De selectieprocedure bestaat uit drie stappen. Allereerst is gekeken naar de goodness of fit. Van de oorspronkelijk aanwezige 914 plantensoorten uit de complete dataset zijn alleen die modellen geselecteerd met een redelijke goodness of fit. Hiervoor is de Hosmer-Lemeshow test gebruikt met een α kleiner dan 0.05. Dit betekent dat de kans dat een model onterecht wordt afgewezen kleiner dan 5% is. Ten tweede is, wanneer er meerdere modellen aan deze eis voldoen, het model gekozen met de hoogste schatter. Ten derde is als extra eis gesteld dat er minimaal één variabele aanwezig moet zijn, die veranderlijk in de tijd is of kan zijn. Dit resulteerde in 'optimale' modellen voor 690 soorten. Voor 15 soorten leverde de goodness of fit maat geen biologisch betekenisvol model op. Voor deze soorten is het model geselecteerd dat net buiten de goodness of fit eis lag, $0.01 \leq \alpha \leq 0.05$, en een hogere maximale kans op voorkomen heeft.
In het uiteindelijke resultaat worden alle veertien modelvarianten minimaal éénmaal gekozen. De variabelen combipaf (toxiciteit als gevolg van de aanwezigheid van zware metalen), zout en vegetatietype komen het minst in de uiteindelijke modellen voor en zijn dus voor minder plantensoorten een belangrijke factor. De andere variabelen worden ongeveer even vaak gekozen.

1. Inleiding

Het doel van het in dit rapport beschreven onderzoek is het verbeteren van de responsie modellen die door De Heer *et al.* (2000) zijn uitgerekend voor het vegetatiemodel MOVE (*mo*del voor de *ve*getatie)
MOVE is een onderdeel van 'De Natuurplanner' (Latour *et al.*, 1997), een beslissings ondersteunend system (BOS). Dit statistische model voorspelt kansen op voorkomen van plantensoorten als functie van bepaalde invoerwaarden. MOVE wordt ingezet bij vraagstukken die door het Mlieu- en Natuurplanbureau van het RIVM beantwoord moeten worden en wordt onder andere gebruikt bij de nationale Milieu- en Natuurverkenningen en –Balansen (RIVM, 2000a; RIVM, 2000b; RIVM, 2001a; RIVM, 2001b; Van Hinsberg *et al.,* 1999; Van der Hoek et al., 2000; Van der Hoek *et al.*, 2002).

Dit rapport is één van een serie rapportages waarin alle op MOVE betrekking hebbende studies verschijnen. Eerder uitgebrachte rapporten zijn Wiertz *et al.* (1992) waarin de ontwikkeling van MOVE 1 staat beschreven, Alkemade *et al.* (1996) met een beschrijving van de calibratie van Ellenberg milieu-indicatiegetallen aan de werkelijk gemeten bodemfactoren en het rapport van De Heer *et al.* (2000) waarin een beschrijving staat van de afleiding van de kansfuncties van MOVE 3.

Bij het onderzoek door De Heer *et al.* (2000) hebben alle plantensoorten hetzelfde responsiemodel, waardoor de kans bestaat dat niet het meest optimale model per soort is afgeleid. Het doel van het in dit vervolgrapport beschreven onderzoek is het verbeteren van de afzonderlijke modellen per plant door per plant een specifiek 'optimaal' model af te leiden met een minimaal aantal variabelen waardoor het aantal eventueel aanwezige ruistermen in het afgeleide responsiemodel geminimaliseerd wordt.

Dit rapport beschrijft de procedures en afwegingen die zijn genomen om tot een 'geoptimaliseerd' vegetatie-regressiemodel voor 690 Nederlandse plantensoorten te komen. Het bestaat uit twee delen, deel I beschrijft de gebruikte dataset en deel II beschrijft de afleiding en achtergronden bij de verschillende onderzochte modelvarianten. In deel II wordt tevens uit de verkregen verzameling modelvarianten per plantensoort de meest optimale variant geselecteerd. De twee delen kunnen onafhankelijk van elkaar geraadpleegd worden. In het RIVM rapport 'Bijlagen bij: MOVE nationaal MOdel voor de VEgetatie versie 3.2, achtergronden en analyse van modelvarianten' (Bakkenes *et al.*, 2002) staan achtergrondinformatie, tabellen en de uiteindelijk 'optimaal' afgeleide modellen waarin in dit rapport verwezen wordt.

In deel I wordt de dataset geanalyseerd die bij de modelafleiding gebruikt is. Er wordt gebruik gemaakt van de database van het Alterra-project 'Plantengemeenschappen' (Schaminée *et al.*, 1995). Allereerst worden de afzonderlijke variabelen beschreven, dit is gedaan om te bepalen of de dataset wel genoeg variatie beschrijft en of de in Nederland voorkomende variatie in de dataset aanwezig is. Vervolgens worden de (a)biotische kenmerken van 13 willekeurig getrokken plantensoorten vergelijken met de dataset. Dit gebeurt door de locatie specifieke voorkomens van deze plantensoorten te vergelijken met de bijbehorende waarden in de dataset.

In deel II wordt de gevolgde procedure beschreven die gebruikt is om de verschillende modelvarianten te onderzoeken. Er worden twee verschillende methoden bekeken. Bij de eerste methode (hoofdstuk 3) worden afzonderlijk verschillende modelvarianten ten opzichte

van een basismodel één voor één doorgerekend, afhankelijk van de ‘goodness of fit’ van het model wordt één van die modellen als het nieuwe basismodel gekozen. Deze procedure herhaalt zich totdat er uiteindelijk geen variabele overblijft die bij een basismodel toegevoegd kan worden. Deze procedure resulteert in het doorrekenen van elf modelvarianten. De tweede methode (hoofdstuk 4) is de stapsgewijze regressiemethode. Bij een stapsgewijze regressiemethode kan er op verschillende manieren een ‘optimaal’ model afgeleid worden; dit rapport bescrhijft vier verschillende methoden.
Uiteindelijk wordt in deel II aan de hand van de verschillende uitgerekende modelvarianten het meest optimale model geselecteerd met behulp van een daarvoor geschikt selectiecriterium (hoofdstuk 5). Dit criterium stelt dat de kans dat een model onterecht wordt afgewezen slechts 5% mag zijn. Wanneer meerdere modellen aan dit criterium voldoen, wordt dat model gekozen waarvan de maximale kans op voorkomen het hoogst is. In de gevallen wanneer dit criterium niet een biologisch plausibel model oplevert is de eis van 5% kans op onterechte afwijzing verlaagd naar 1%. Uiteindelijk is er voor 690 van de oorspronkelijk 914 aanwezige plantensoorten een model afgeleid.

Deel I

Analyse van de dataset

2. De gegevens

2.1 Het opnamebestand

In deze studie wordt gebruik gemaakt van de database van het Alterra-project 'Plantengemeenschappen' (Schaminée *et al.*, 1995). Het geleverde bestand bevat in totaal 169.000 vegetatieopnamen uit de periode 1901 tot 1997 (Runhaar *et al.*, 2002). De opnamen komen uit terrestrische, semi-terrestrische en aquatische milieus. Het opnamebestand heeft met name betrekking op hogere planten. In de opnamen zitten niet alleen wilde soorten, maar ook cultuurgewassen. De planten zijn tot op het niveau van afzonderlijke soorten benoemd, in enkele gevallen tot op het niveau van ondersoorten. Moeilijk determineerbare soorten zijn soms opgenomen als soortgroep, bijv. het geslacht Agrostis. Deze studie beschouwt uitsluitend de inheemse, wilde, hogere planten op het niveau van de soort. In totaal zitten er 1599 soorten in de dataset, CBS-nummers tussen 1 en 6717 (CBS, 1993).
Het opnamebestand is tot stand gekomen door een niet-aselecte bemonstering. Bekend is dat er sprake is van een relatieve oververtegenwoordiging van opnamen op plekken met een relatief hoge botanische waarde ten koste van meer gewone plekken (zie ook Wiertz *et al.*, 1992; en figuur 2.8). Bij univariate modellering van responsfuncties kan de niet aselecte bemonstering verschuivingen in de responskrommen tot gevolg hebben (Runhaar *et al.*, 1994). Door Runhaar *et al.* (2002) is getracht om achteraf te corrigeren voor deze wijze van bemonstering. Een bezwaar tegen deze correctie achteraf is dat het niet goed mogelijk is hiervoor objectieve criteria vast te stellen. Wanneer, zoals in deze studie, responsfuncties worden afgeleid op basis van verschillende verklarende factoren tegelijkertijd, zullen de nadelen van een niet aselecte dataset voor een deel ondervangen worden. Door deze multivariate analyse vindt een vergaande opsplitsing van afzonderlijke milieus plaats, waardoor informatie over verschillende milieus minder wordt gemengd en de responsvergelijkingen minder verschuiven. Zo zullen bijvoorbeeld milieus met en zonder veel zeldzame soorten op basis van één of meer milieufactoren afzonderlijk onderscheiden worden. Verschuivingen die voorheen ontstonden door vermenging van informatie over deze milieus kunnen nu niet meer plaatsvinden. Om deze redenen is in deze studie geen correctie uitgevoerd op de onevenwichtige verdeling. Wel zijn opnamen die per definitie betrekking hebben op heterogene milieus (oevers en andere lijnvormige elementen) uitgesloten.

2.2 De omgevingsvariabelen in de dataset

Het overgrote deel van de opnamen in de dataset betreft uitsluitend een registratie van de plantensoorten. Een gelijktijdige bepaling van abiotische parameters ontbreekt meestal. De waarde van de meeste omgevingsvariabelen wordt daarom afgeleid uit de corresponderende Ellenberg indicatiewaarden (Ellenberg *et al.*, 1992) van de aanwezige plantensoorten voor de desbetreffende opname. Voor enkele variabelen zijn wel externe gegevens beschikbaar. Op deze wijze zijn de volgende omgevingsvariabelen afgeleid: op het gebied van de abiotiek zijn dit het stikstofgetal (Ellenberg-n), het vochtgetal (Ellenberg-f), het zuurgetal (Ellenberg-r) en het zoutgetal (Ellenberg-s); op het gebied van beheer zijn dit het maaigetal (Ellenberg-m), het lichtgetal (Ellenberg-l) en de vegetatiestructuur (veg). Hier zijn later nog 2 variabelen aan toegevoegd: de fysische geografische regio's (fgr) en de potentieel aangetaste fractie plantensoorten door zware metalen (cpaf2). Tabel 2.1 geeft een overzicht van alle variabelen. In onderstaande paragrafen wordt verder ingegaan op de betekenis van deze variabelen voor planten en de berekening hiervan.

Tabel 2.1 Omgevingsvariabelen uit de dataset

variabele	afkorting	type	range
vochtgetal	f	continue	1 – 12
lichtgetal	l	continue	1 – 9
maaigetal	m	continue	1 – 9
stikstofgetal	n	continue	1 – 9
zuurgetal	r	continue	1 – 9
vegetatiestructuur	veg	discreet	5 klassen
potentieel aangetaste fractie door zware metalen	cpaf2	continue	0 – 1
zoutgetal	s	continue	0 – 9
fysisch geografische regio	fgr	discreet	11 klassen

De Ellenberg indicatiewaarden worden bepaald als het rekenkundig gemiddelde van de indicatiegetallen van de aanwezige soorten. Er vindt geen weging naar het aantal maal dat een soort voorkomt (abundantie) plaats. De methode vereist dat elke opname voor elke variabele tenminste 2 soorten heeft met een indicatiegetal (Runhaar *et al.*, 2002). Opnamen waarin dit niet het geval is, krijgen een ‘missing value’ voor de betreffende variabele. Uit de dataset worden op voorhand alle opnamen met missing values voor de variabelen n, f, r, s, veg en fgr uit de dataset verwijderd. Het voordeel hiervan is dat omvang van de dataset zo beperkt mogelijk blijft (109065 opnamen) waardoor het rekenen sneller kan, maar het nadeel is dat er mogelijk onterecht soorten uit de dataset verwijderd zijn. Na toevoegen van het indicatiegetal voor de toxische druk als gevolg van zware metalen (cpaf2) blijven er nog 95529 opnamen over (met daarin 914 soorten), een afname van 12.5%. Paragraaf 2.2.4 beschrijft de gevolgde procedure om tot de afleiding van cpaf2 te komen.

Tabel 2.2 geeft de correlaties tussen de continue verklarende omgevingsvariabelen weer. Deze correlaties zijn berekend op de verkleinde dataset van 95529 opnamen. In bijna alle gevallen liggen de correlatiecoëfficiënten beneden de 0.3. De uitzondering hierop is de correlatie tussen r en n, deze heeft een waarde van 0.78. Deze correlatie vinden wij niet hoog genoeg, om of r of n niet in de analyse mee te nemen; nog ruim 39% blijft onverklaard. Een andere reden om toch zowel r en n mee te nemen is dat voor veel vraagstukken gewenst is om zowel iets over effecten van r (pH) als n (nutriënten, vermesting) te kunnen zeggen.

Tabel 2.2 Correlaties tussen de continue omgevingsvariabelen

	f	r	n	s	cpaf2
f	1	0.28	0.27	0.05	0.16
r		1	0.78	0.24	0.08
n			1	0.09	0.07
s				1	-0.03
cpaf2					1

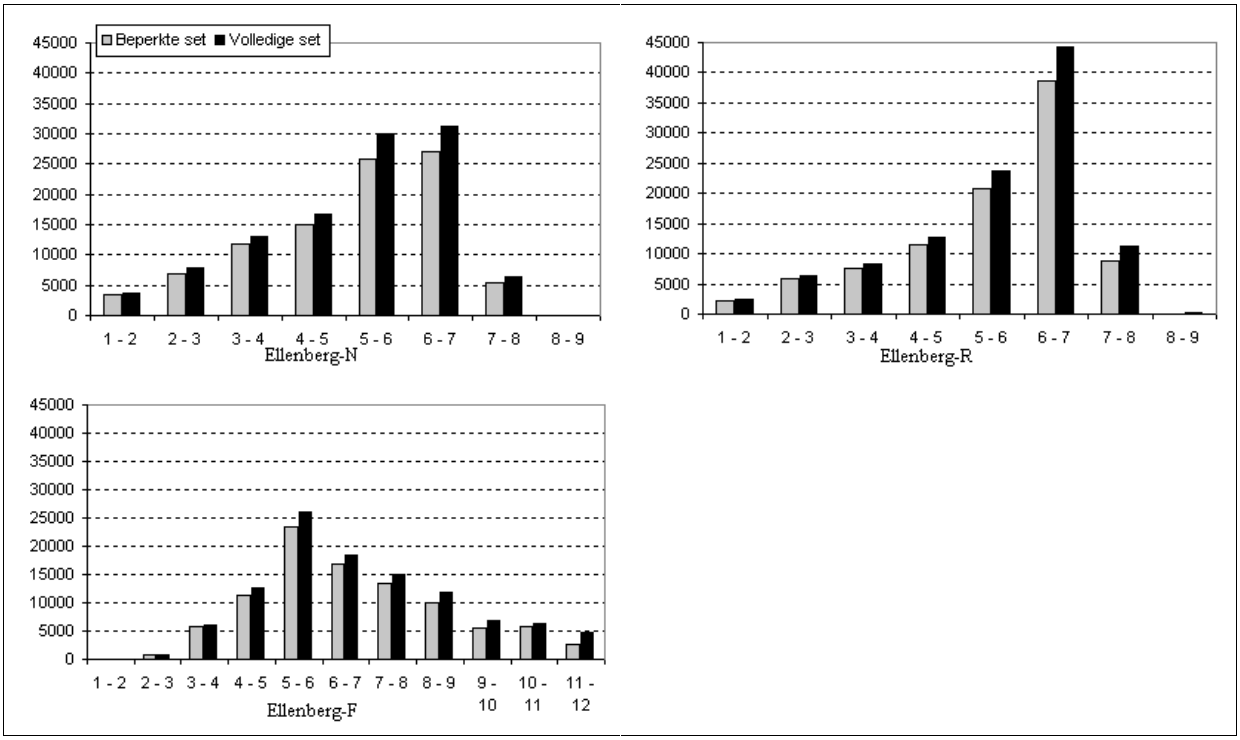
In tabel 2.3 staan voor de continue gegevens (f, -r, -n, s en de cpaf2) wat algemene statistische informatie, zoals het minimum, het gemiddelde, de mediaan. Uit de tabel en de bijbehorende figuren in de paragrafen 2.2.1 t/m 2.2.5 vallen met name de scheve verdeling van zowel Ellenberg zout en de combipaf op.

Tabel 2.3 Statistische informatie over de continue gegevens uit de dataset

	f	r	n	s	cpaf2
minimum	2.00	1.00	1.00	0.00	0.00
eerste kwartiel	5.42	4.94	4.23	0.12	0.00
gemiddelde	6.90	5.68	5.22	0.50	0.01
mediaan	6.500	6.11	5.56	0.28	0.00
derde kwartiel	8.250	6.75	6.33	0.50	0.00
maximum	12.00	8.67	8.67	8.67	0.72
totaal N	109065	109065	109065	109065	95529
variantie	4.38	2.07	2.17	1.07	0.0015
standaard deviantie	2.09	1.44	1.47	1.03	0.039

2.2.1 Abiotische variabelen die in MOVE 2 gebruikt worden

In de dataset zitten ook de zogenaamde MOVE 2 variabelen: de Ellenberg indicatiewaarden n, f en r voor respectievelijk voedselrijkdom, vochttoestand en zuurgraad. De indicatiegetallen voor stikstof hebben een bereik van voedselarm (1) naar voedselrijk (9). Voor zuurgraad loopt de schaal ook van 1 tot 9; van zuur naar basisch. De vochtgetallen hebben een schaal van 1 tot 12: van droogtetolerante soorten naar aquatische soorten. In figuur 2.1 staan de histogrammen van de verdeling van n, r en f over de dataset.

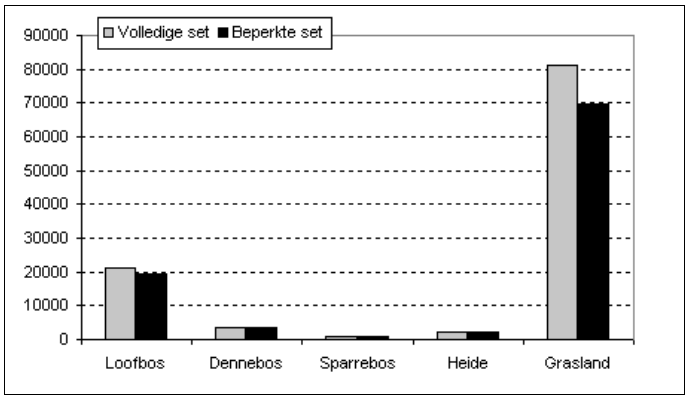


Figuur 2.1Verdeling van n-, r- en f-Ellenberg over de dataset ($n_{volledig}$=109065;$n_{beperkt}$=95529)

2.2.2 Variabelen die een indicatie zijn voor het gevoerde beheer

Planten concurreren met elkaar om nutriënten en licht. Beheer beïnvloedt de beschikbaarheid hiervan. Maaien en grazen zorgen voor afvoer van nutriënten uit het systeem. Door verwijdering van de bovengrondse biomassa verandert ook het lichtklimaat. Bovendien heeft beheer een directe invloed op planten. De groeivorm en het regeneratievermogen van een soort bepalen de gevoeligheid voor beheerschade (Oosterbeek *et al.*, 1997). Deze relaties geven het belang aan van het meenemen van beheervariabelen in MOVE 3.

Beheer kan in de modellering worden meegenomen door gebruik te maken van indicatiewaarden voor maaien (Briemle en Ellenberg, 1994; aangevuld en aangepast door Wamelink *et al.*, 1997) en licht (Ellenberg, 1992). Het maaigetal is met name van belang voor de directe effecten van het beheer. Het lichtgetal is, net als het stikstofgetal overigens, meer indirect gerelateerd aan menselijke ingrepen. Oosterbeek *et al.* (1997) geven een nadere analyse van de mate waarin het maai- en lichtgetal het effect van verschillende beheervormen en beheerintensiteit kunnen beschrijven. De getallen bleken slechts in beperkte mate de verschillende vormen te onderscheiden. Wel karakteriseren ze echter de beheerintensiteit, c.q. het aantal maal dat beheer per tijdseenheid wordt uitgevoerd. Het maaigetal onderscheidt intensief, extensief en zeer extensief beheer. Het lichtgetal is meer indirect gerelateerd aan menselijke beheermaatregelen. Tevens concluderen Oosterbeek *et al.* (1997) op grond van een verkennende analyse dat het gebruik van maai- en lichtgetallen de voorspelling van het voorkomen van soorten aanzienlijk kan verbeteren.
Net als bij de andere Ellenberg indicatiewaarden loopt de klassenindeling van 1 tot 9. De schaal van het maaigetal loopt daarbij op van ‘verdraagt geen maaien’ (1) tot ‘alleen concurrentiekrachtig bij regelmatig maaien resp. zware betreding’ (9). Het lichtgetal loopt van zeer schaduwrijk (1) naar zeer licht (9). De variabele ‘vegetatiestructuur’ (veg) geeft een globalere aanduiding van het beheer. Hierin komen zowel directe als indirecte effecten van beheer tot uiting alsmede de effecten van andere factoren waaronder natuurlijke successie. De variabele kan worden gebruikt als een pragmatisch alternatief voor de maai- en licht-indicatiewaarden. De variabele ‘vegetatiestructuur’ heeft 5 klassen: grasland (dwz. lage, kruidige vegetatie), heide, loofbos, dennenbos en sparrenbos. Deze klassen stemmen overeen met de vegetatieklassen zoals het bodemmodel SMART die hanteert (Kros, 1998). De klassen zijn toegekend op basis van de in de opname voorkomende soorten. Zie figuur 2.2 voor een overzicht van de verdeling van de verschillende vegetatiestructuurtypen over de dataset.



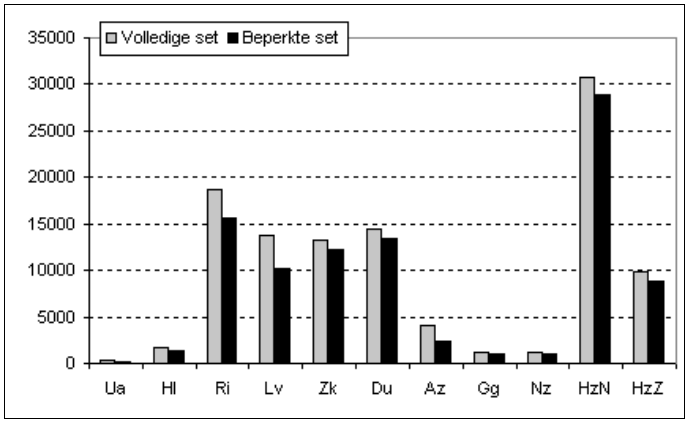
Figuur 2.2 Verdeling afzonderlijke vegetatietypen over de dataset($n_{volledig}=109065; n_{beperkt}=95529$)

Uit deze figuur blijkt dat de meeste opnamen in het vegetatietype grasland liggen op afstand gevolgd door loofbos.

2.2.3 De fysisch geografische regio’s

Om te voorkomen dat het regressiemodel planten met een specifieke geografische verspreiding op ‘onjuiste’ locaties voorspelt is er besloten om een variabele toe te voegen die informatie geeft over de ruimtelijke ligging. Hiervoor kunnen verschillende indelingen gebruikt worden, dit kunnen onder andere bodemtypen, flora districten of fysische geografische regio’s zijn. Omdat de fysische geografische regio’s in digitale kaartvorm beschikbaar was en omdat deze zowel een samenhang heeft met de flora districten en bodemtype is om praktische redenen voor deze indeling gekozen.

De fysisch geografische regio's bestaan uit 10 klassen: stedelijk gebied, heuvelland, hogere zandgronden, rivierengebied, laagveengebied, zeekleigebied, duingebied, afgesloten zeearmen, getijdengebied en noordzee. Omdat er wat de flora betreft een groot verschil aanwezig is tussen soorten die voorkomen in de fysisch geografische regio hogere zandgronden onder en boven de grote rivieren, is de regio hogere zandgronden opgedeeld in twee afzonderlijke regio's: hogere zandgronden boven de grote rivieren en hogere zandgronden onder de grote rivieren. In totaal worden er nu dus 11 fysisch geografische regio's onderscheiden (Van der Hoek *et al.*, 2000). In figuur 2.3 staat het aantal opnamen in de dataset per fysisch geografische regio weergegeven.



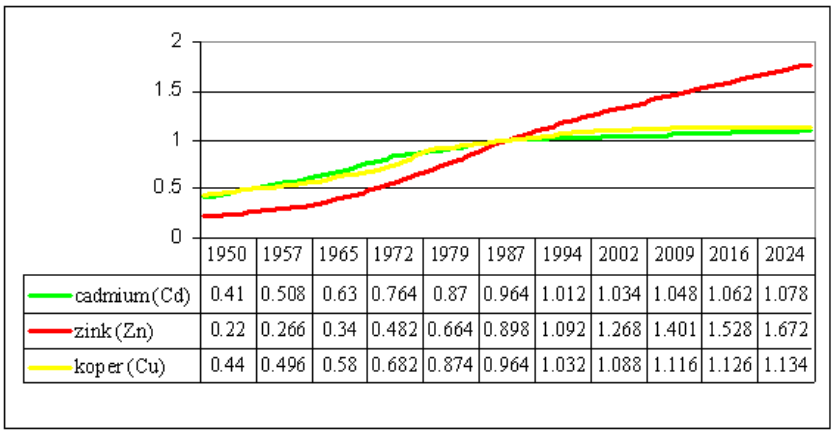
Figuur 2.3 Verdeling afzonderlijke(sub) fysich geografische regio's[1] over de dataset ($n_{volledig}$=109065;$n_{beperkt}$=95529)

2.2.4 Zware metalen

Het is bekend dat de aanwezigheid van zware metalen in het milieu aanleiding kan geven tot ernstige milieuproblemen (Kitagishi en Yamane, 1981). Hoewel sommige zware metalen (bijv. koper en zink) behoren tot de elementen die essentieel zijn voor enzymatische processen in levende organismen, zijn alle zware metalen reeds bij relatief lage concentraties giftig. Verschillende (groepen) organismen vertonen een duidelijk onderscheid in hun gevoeligheid voor de toxische werking van deze stoffen. Deze verschillen worden voornamelijk veroorzaakt door verschillen in blootstellingsroute en verschillen in biochemisch bouwplan (uitscheiding, conjugatie, etc.) (Walker, 1987; Janssen, 1991). De verschillen in soortspecifieke gevoeligheid wordt bijna spreekwoordelijk geïllustreerd door het uiterst regionaal voorkomen van een soort als het zinkviooltje.
Bij het berekenen van de toxische druk door zware metalen zijn de metalen cadmium (Cd), koper (Cu) en zink (Zn) betrokken. Deze berekening is in twee stappen uitgevoerd. Met behulp van een statistische opschalingsprocedure is het huidige totale gehalte zware metalen berekend (Tiktak, 1999). Vervolgens zijn met behulp van het model SOACAS (Tiktak *et al.*, 1998) de huidige metaalgehalten geëxtrapoleerd naar de periode 1950-2050 (figuur 2.4). De berekeningen zijn uitgevoerd voor de metalen Cd, Zn en Cu omdat voor deze metalen voldoende gegevens aanwezig waren. Als maat voor de toxische druk is het *reactief metaalgehalte* gebruikt (Tiktak *et al.*, 2000). Deze fractie kan gezien worden als een maat voor de potentieel biobeschikbare fractie. Er dient op gewezen worden dat effecten op bodemorganismen in het algemeen het best beschreven kunnen worden op basis van de concentratie in oplossing of op basis van de vrije metaalactiviteit (Gregor, 1999). Aangezien

[1] Fysisch geografische regio's: Ua: stedelijke gebied; Hl: heuvelland; Ri: rivierengebied; Lv: laagveengebied; Zk: zeekleigebied; Du: duingebied; Az: afgesloten zeearmen; Gg: getijdengebied; Nz: Noordzee; HzN: hogere zandgronden noord (boven de grote rivieren) en HzZ: hogere zandgronden zuid (onder de grote rivieren)

er te weinig gegevens beschikbaar zijn om de concentratie in oplossing of om de vrije metaalactiviteit betrouwbaar te voorspellen, is gekozen voor de reactieve fractie.



	1950	1957	1965	1972	1979	1987	1994	2002	2009	2016	2024
cadmium (Cd)	0.41	0.508	0.63	0.764	0.87	0.964	1.012	1.034	1.048	1.062	1.078
zink (Zn)	0.22	0.266	0.34	0.482	0.664	0.898	1.092	1.268	1.401	1.528	1.672
koper (Cu)	0.44	0.496	0.58	0.682	0.874	0.964	1.032	1.088	1.116	1.126	1.134

Figuur 2.4 De gebruikte trendlijnen binnen SOACAS om het reactief metaalgehalte af te leiden voor cadmium, zink en koper(als fracties 1990 = 1.0)

De toxische druk wordt in deze studie meegenomen in de vorm van de variabele ‘cpaf2’, de potentieel aangetaste fractie plantensoorten door zware metalen. Figuur 2.6 schematiseert de berekening van metaalconcentraties in de bodem naar cpaf2-waarde (zie ook bijlage I.a in Bakkenes *et al.* (2002) voor de gevolgde procedure).

De combipaf van cadmium, koper en zink wordt berekend volgens:

$$PAF_x = \frac{1}{1+e^{-\frac{(x-\alpha)}{\beta}}}$$

$$x = {}^{10}\log(M_{reactief}) \qquad (2.1)$$

$$PAF_{zm} = 1-(1-PAF_{Cd})*(1-PAF_{Cu})*(1-PAF_{Zn})$$

Binnen de bovenstaande formules worden de in de onderstaande tabel gedefinieerde alfa’s en bèta’s[1] gebruikt.

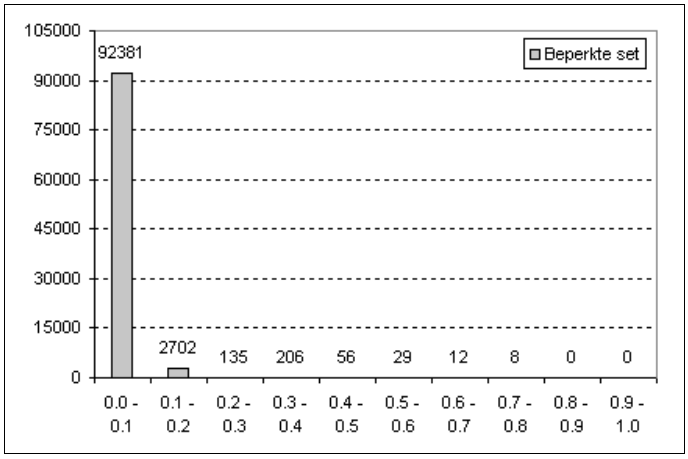
	Cd	Cu	Zn
alfa (α)	0.95	2.38	2.52
bèta (β)	0.22	0.22	0.19
aantal planten	19	12	17

[1]alfa’s en bèta’s op basis van NOEC-gegevens in mg/kg dw uit verschillende datasets (Crommentuijn, 1997; Klepper en Van de Meent, 1997; Will en Suter II, 1995; EEG; Janus / Bodar, 1999)

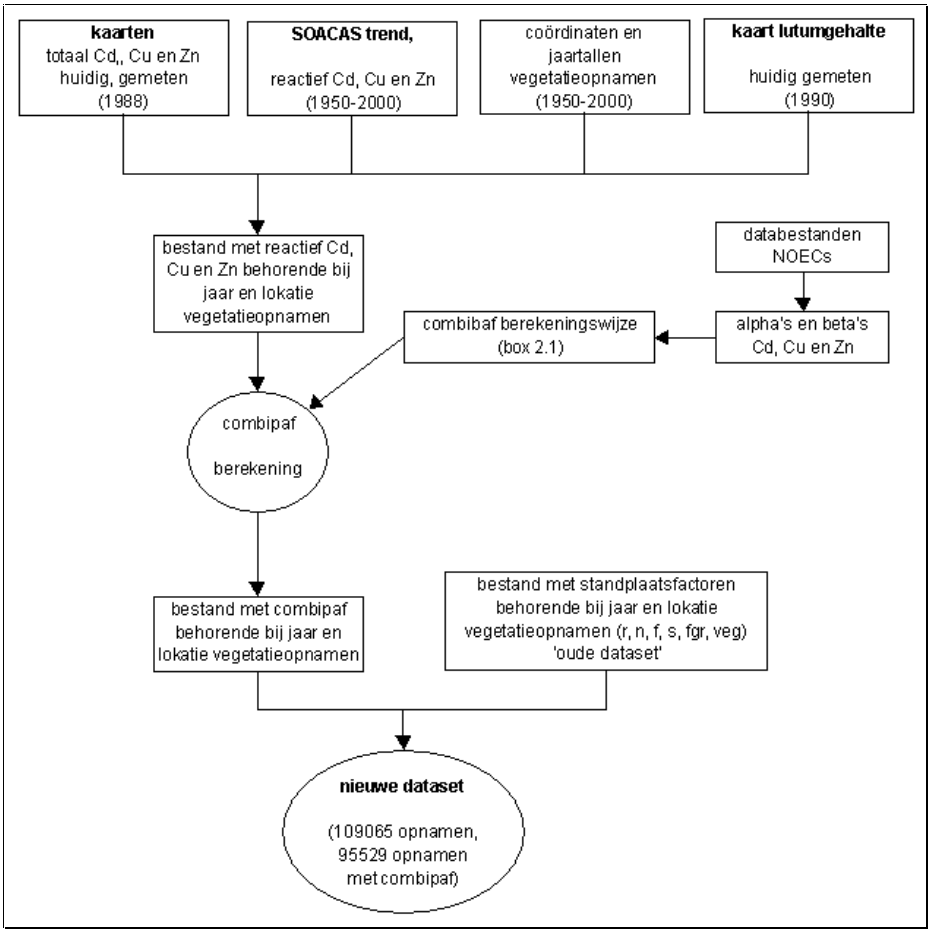
Box 2.1 Berekeningswijze combipaf (cpaf2)

De cpaf2-waarden (box 2.1) worden berekend op basis van reactief metaal; dit in tegenstelling tot een eerdere methode waarbij de cpaf2 voor zwaar metaal in oplossing is berekend. Deze verandering is aangebracht omdat bij de nu gebruikte methode (box 2.1) veel meer toxiciteitgegevens (no effect concentrations (NOECs)) beschikbaar zijn. Bij veel NOECs is geen pH of lutum percentage bekend. De onzekerheid in de toxiciteitparameters alfa en bèta is daardoor afgenomen, terwijl er wellicht nauwkeurigheid ingeleverd is bij de werkelijke beschikbaarheid van de metalen voor de planten. De cpaf2-waarden zijn in de uiteindelijke dataset afgerond op 2 cijfers achter de komma. Dit is gedaan omdat de gebruikte wijze van bereken zeer veel kleine waarden opleverde en deze ‘pseudo’ nullen in de uit te

voeren regressie als ruis meegenomen zouden worden. Figuur 2.5 laat de verdeling van de cpaf2-waarden over de dataset zien. Wat in deze figuur opvalt is dat er vooral zeer lage cpaf2-waarden aanwezig zijn.



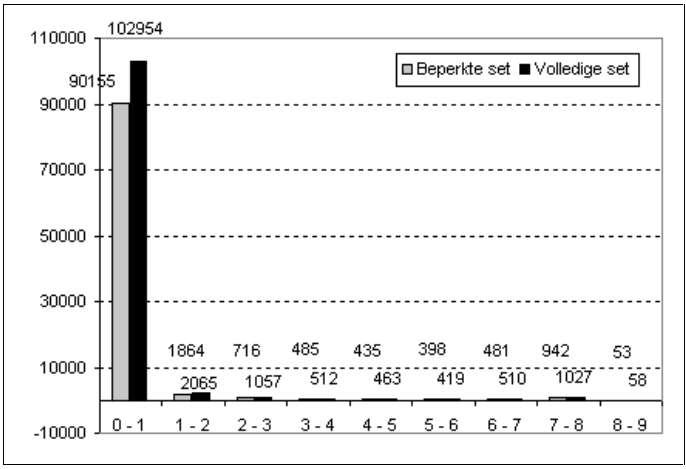
Figuur 2.5 Verdeling van cpaf2 over de dataset (n=95529)



Figuur 2.6 Schematische weergave van de afleiding van cpaf2 en toevoeging aan de dataset

2.2.5 Het zoutgehalte

Het zoutgehalte van de bodem is een belangrijke factor voor het voorkomen van planten. Net als bij zuurgraad, nutriëntenbeschikbaarheid en vochtgetal is ook hierbij gebruik gemaakt van de Ellenberg indicatiewaarden voor zout (s). De schaal loopt van zout- intolerant (0) naar sterk zoutbehoevend (9).



Figuur 2.7 Verdeling van Ellenberg-s waarden over de dataset ($n_{volledig}$=109065;$n_{beperkt}$=95529)

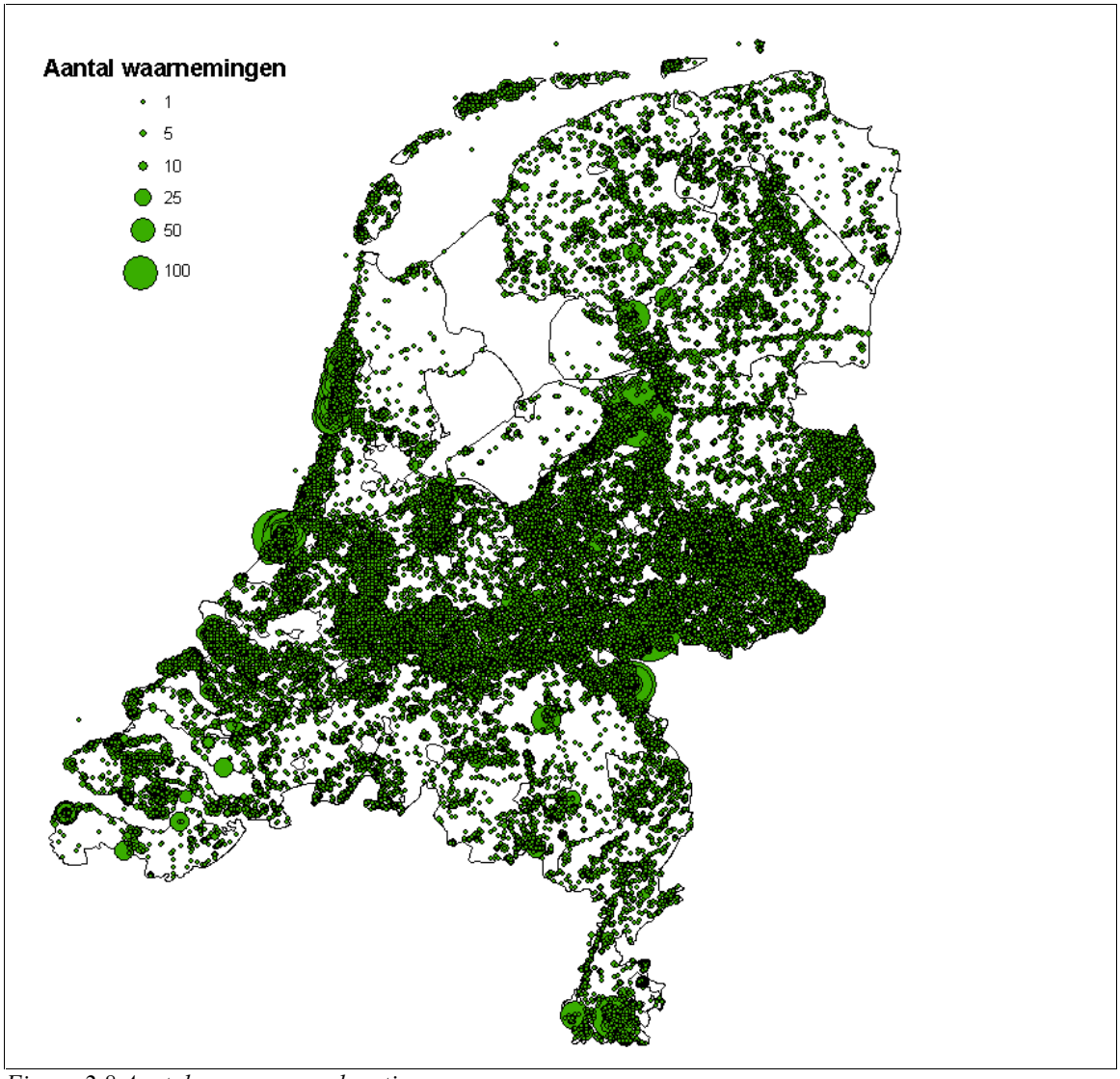
2.3 Verdeling van de opnamepunten

Om na te gaan hoe de opnamepunten verdeeld zijn over de dataset is er allereerst gekeken naar de verdeling van de opnamen per fysisch geografische regio (fgr). In tabel 2.4 staat per fgr aangegeven wat de relatieve oppervlakte natuur t.o.v. het totale oppervlak natuur in Nederland (eerste kolom) is en tevens staat per fgr het relatief aantal opnamen (tweede kolom). De correlatie coëfficiënt (r) tussen deze twee kolommen is gelijk aan 0.80. Dit lijkt erop te wijzen dat het aantal opnamen redelijk gelijk verdeeld is over Nederland en dat deze verdeling in overeenstemming is met het areaal natuur per fgr. Wanneer er alleen naar de fracties onderling gekeken zou worden, dan blijkt dat er een slecht verband is en dat in alle gebieden het aantal opnamen afwijkt van wat er op basis van het oppervlakte natuur in dat gebied gebied verwacht zou mogen worden. Er is een χ^2 analyse uitgevoerd om te analyseren of het aantal opnamen per regio overeenkomt met het aantal dat verwacht zou mogen worden op basis van de hoeveelheid aanwezige natuur. Uit de laatste kolom van tabel 2.4, $(f_o - f_e)^2$, blijkt dat dit voor geen van de regio's het geval is. Er bestaat een significante afwijking van het aantal ten opzichte van de verwachting. Het aantal waarnemingen in het rivierengebied wijkt bijvoorbeeld sterk af van wat er verwacht mag worden, wanneer de opnamen gelijkelijk verdeeld zouden zijn over natuurgebieden in de fysisch geografische regio's. Maar 4.5% van de hoeveelheid natuur in Nederland ligt in het rivierengebied, maar meer dan 17% van het totaal aantal opnamen liggen in deze regio.

Tabel 2.4 Relatieve oppervlakte natuur per fgr en relatief aantal opnamen per fgr in %

	fractie (%)		aantal		
	opp. natuur	opnamen	f_o	f_e	$(f_o - f_e)^2$
Heuvelland	0.59	1.54	1670	645	1628.9
Rivierengebied	4.49	17.39	18914	4879	40373.3
Laagveengebied	7.97	12.73	13846	8666	3096.3
Zeekleigebied	6.30	12.35	13430	6855	6306.4
Duingebied	10.87	13.38	14549	11823	628.5
Afgesloten zeearmen	2.88	3.77	4095	3133	295.4
Getijdengebied	2.56	1.10	1196	2779	901.7
Hogere zandgronden boven de grote rivieren	45.26	28.64	31146	49212	6632.1
Hogere zandgronden onder de grote rivieren	19.08	9.10	9894	20747	5677.3
				χ^2	65540.0

Wanneer er nu gekeken wordt naar de verdeling van deze opnamen over het totale areaal valt op dat er een relatieve oververtegenwoordiging van opnamen is op plekken die een relatief hoge natuurwaarde hebben ten koste van meer algemene gewone plekken (figuur 2.8). Het gaat hier om plekken waar een hoge dynamiek aanwezig is, zoals stijlranden, kwelgebieden, kustgebieden.



Figuur 2.8 Aantal opnamen per locatie

2.4 Soortwaarnemingen in de dataset

Op voorhand zijn van de oorspronkelijke 1599 soortwaarnemingen uit de dataset bijna 700 soorten met minder dan 50 positieve waarnemingen verwijderd. De kans op een significant model is kleiner bij een laag aantal positieve waarnemingen. Om te vermijden dat er veel soorten doorgerekend worden die geen significant model opleveren zijn deze soorten van tevoren uit de dataset gehaald. Een nadeel hiervan is dat soorten die ondanks weinig positieve waarnemingen toch een significant model hebben niet doorgerekend worden. Ondanks dit bezwaar is er toch voor gekozen om bijna 700 soorten te verwijderen. In totaal zijn er na deze stap voor 914 soorten gegevens beschikbaar die gekoppeld zijn aan opnamepunten. In de oorspronkelijk aangeleverde gegevens stonden alle plantvoorkomens in één groot bestand, maar om het doorrekenen van de individuele soorten te vereenvoudigen zijn deze in afzonderlijke bestanden gezet. Ieder bestand bestaat uit 109065 regels, dit is gelijk aan het totale aantal opnamepunten. Met nullen en enen wordt aangegeven of die soort waargenomen is bij die specifieke opnamen. Met deze één op één relatie kan nu voor iedere soort die opnamen geselecteerd worden waar deze soort is waargenomen.
In de eerste plaats is voor iedere soort het bereik van de continue gegevens (f, r, n, s en cpaf2) waarbinnen deze soort voorkomt bepaald. De minimale, maximale en gemiddelde waarden, de standaard deviantie, de variantie en de som zijn er berekend. Voor de geclassificeerde gegevens (fgr, veg) is bepaald wat de minimale en maximale waarde en het aantal waarnemingen per klassen is. Bijlage I.b t/m I.d (Bakkenes *et al.*, 2002) laat voor alle 914 soorten deze ranges zien.

Voor dertien willekeurig getrokken soorten worden de waarden uit het opnamebestand vergeleken met de beschreven preferenties in Heukels' Flora van Nederland (Van der Meijden, 1990). In onderstaand overzicht (box 2.2) staan voor de willekeurig getrokken soorten de bijbehorende omschrijving volgens Heukels'. Per soort is overgenomen de tekst die betrekking heeft op het voorkomen van die soort en per soort is de indeling per ecologische groep overgenomen.

Alliaria petiolata (Look-zonder-look)
Heukels: Deze soort komt voor op half beschaduwde plaatsen op voedselrijke, bij voorkeur zandige grond. Algemeen, vooral in Pleistocene districten zich uitbreidend; in Drents district en IJsselmeerpolders vrij zeldzaam. Eco: H47, H48, H69
Arctium lappa (Grote klit)
Heukels: In bermen en op ruderale plaatsen, ook in lichte loofbossen en hooggelegen grienden. Plaatselijk vrij algemeen in Fluviatiel district, elders zeldzaam Eco: R48
Dianthus deltoides (Steenanjer)
Heukels: In droge, zandige graslanden en bermen. Plaatselijk vrij algemeen in het stroomgebied van Overijsselse Vecht en Dinkel; zeldzaam elders in Pleistocene districten en in Renodunaal district; elders zeer zeldzaam. Ten dele wel verwilderd of adventief plant Eco: G62.
Elymus athericus (Strandkweek)
Heukels: Aan zeeduiken, in aanspoelselgordels en ruigten aan de rand van schorren, voorts in de duinen en mogelijk ook in het binnenland. Algemeen in Estuariëndistrict, Renodunaal district, Waddendistrict en aan de kust van Noordelijk kleidistrict, Gelders district en IJsselmeerpolders; voorts zeldzaam in Laagveen district en Fluviatiel district. Eco: bP60st, bR40, R64.
Hypericum humifusum (Liggend hertshooi)
Heukels: Op vochtige zand- en leemgrond. Plaatselijk vrij algemeen in Pleistocene districten en zuid Limburgs district; zeldzaam in het Zeeuwse en Zuid-Hollandse deel van Renodunaal district. Eco: P42.
Juncus acutiflorus (Veldrus)
Heukels: In natte, onbemeste hooilanden, vooral langs beken, ook op moerassige heiden, langs vennen, aan kwelsloten; soms in natte duinvalleien. Vrij algemeen in Pleistocene districten, elders zeldzaam. Eco: G22, G27.
Lotus corniculatus subsp. corniculatus (Gewone rolklaver)
Heukels: Op allerlei zonnige, grazige, droge tot iets vochtige, weinig of niet bemeste plaatsen. Algemeen. Ook uitgezaaid. Eco: G43, G47, G62, G63, G67.
Mercurialis perennis (Bosbingelkruid)

Heukels: In loofbossen, meestal op kalkrijke grond. Vrij zeldzaam in zuid Limburgs district; daarbuiten op enige verspreide vindplaatsen. Eco: H43.
Myosotis palustris (Moerasvergeet-mij-nietje)
Heukels: Aan en in zoet water, in drassige gras- en rietlanden. In lichte broekbossen en grienden. Algemeen. Eco: G28, R28.
Rubus fruticosus (Gewone braam)
Heukels: Aan bosranden, in ruigten en in bossen. Algemeen, vooral in Pleistocene districten, zuid Limburgs en Estuariëndistrict. Ook in cultuur om de eetbare vruchten. Eco: R44, R47, R64, R67, H41, H42, H47, H61, H62, H69.
Rumex conglomeratus (Kluwenzuring)
Heukels: Aan waterkanten, in natte graslanden, soms aan vochtige bospaden. Algemeen; in Drents, Gelders, Waddendistrict en IJsselmeerpolders minder algemeen. Eco: G28, H28.
Salicornia procumbens (Langarige zeekraal)
Heukels: In Estuariën, Waddendistrict en aan de kust van Noordelijk kleidistrict, vroeger ook langs de Zuiderzee. Buitendijks naar de zeekant algemeen voorkomend op kaal slik en in slijkgras- en kweldergras vegetaties, voornamelijk beneden de gemiddeld hoogwaterlijn. Binnendijks slechts op sterk zilte, vochthoudende plaatsen, meestal onder invloed van zoute kwel. Eco: zP20.
Trifolium pratense (Rode klaver)
Heukels: In graslanden en bermen op vochthoudende grond. Algemeen; ook vaak uitgezaaid. Eco: G47, G48.

Box 2.2 Overzicht en beschrijving in Heukels van de 13 willekeurig getrokken *soorten*

Regionale verspreiding

In Heukels' (Van der Meijden, 1990) wordt voor de regionale verspreiding van plantensoorten de indeling in Flora Districten van Nederland[2] gebruikt. In de dataset is de indeling naar fysisch geografische eenheden gebruikt. Om de beschreven regionale verspreiding te kunnen vergelijken met de dataset is aan de hand van de beschrijving van de ligging van deze districten een vertaaltabel gemaakt waarin per fysisch geografische regio de bijbehorende floradistricten staan (tabel 2.5).

Tabel 2.5 Vertaaltabel van fysisch geografische regio's naar Flora Districten

	afkorting	floradistricten[2]
Heuvelland	Hl	Z
Rivierengebied	Ri	F
Laagveengebied	Lv	L
Zeekleigebied	Zk	N, E, L, IJ
Duingebied	Du	R, V
Afgesloten zeearmen	Az	W
Getijdengebied	Gg	E
Hogere zandgronden boven de grote rivieren	HzN	G, S, D
Hogere zandgronden onder de grote rivieren	HzZ	K

In tabel 2.5 staat per soort het aantal waarnemingen per fgr. De meeste soorten komen in bijna alle fgr's voor, maar vaak zijn er enkele fgr's waar die soort het meest voorkomt. Deze dominante regio's zijn vergeleken met de omschrijving volgens Heukels'. De resultaten van deze vergelijking staan in tabel 2.6. Over het algemeen komt de regionale verspreiding zeer goed overeen met de waarnemingen in de dataset, er zijn een paar soorten waar wat opmerkingen bij gemaakt kunnen worden. Zo moet *Alliaria petiolata* algemeen voorkomen en zich met name uitbreiden over de Hogere zandgronden (HzN en HzZ), maar deze wordt voornamelijk aangetroffen in het Rivierengebied (Ri), de Duinen (Du) en de Hogere zandgronden boven de grote rivieren (HzN) (tabel 2.6). *Salicornia procumbens* moet

[2] Flora districten: D (Drents district), E (Estuariëndistrict), F (Fluviatiel district), G (Gelders district), H (Hafdistricten E, L, N), K (Kempens district), L (Laagveen district), N (Noordelijk kleidistrict), P (Pleistocene districten D, G, K, S, V), R (Renodunaal district), S (Subcentreuroop district), V (Vlaams district), W (Waddendistrict), IJ (IJsselmeerpolders), Z (Zuidlimburgs district)

voornamelijk voorkomen in Estuariën, het Waddendistrict en aan de kust van het noordelijk Kleidistrict, deze klopt redelijk. Er zijn alleen weinig waarnemingen, in totaal 73, waarvan er relatief veel in de fysische geografische regio 'Laagveen (Lv)' liggen.

Tabel 2.6 Aantal waarnemingen per fgr voor 13 willekeurig geselecteerde soorten

	aantal malen waargenomen in fysisch geografische regio											
wetenschappelijke naam	totaal	stad	Hl	Ri	Zk	Lv	Du	Az	Gg	Nz	HzN	HzZ
Alliaria petiolata	1780	9	101	802	12	57	479			3	260	57
Arctium lappa	128	1		58	3	30	17	4			12	3
Dianthus deltoides	173	1		112		1					59	
Elymus athericus	1935	2		19	1	609	552	286	368	85	7	6
Hypericum humifusum	92		1	4			3				65	19
Juncus acutiflorus	2319	2	11	83	72	27	5	3			979	1137
Lotus corniculatus subsp. corniculatus	5268	15	544	1300	29	288	2116	71	38	324	397	146
Mercurialis perennis	110		102	4							4	
Myosotis palustris	6019	11	18	2120	1165	740	76	76	2	2	1381	428
Rubus fruticosus	15626	28	270	1380	783	225	276	23	1	4	11137	1499
Rumex conglomeratus	1307	3	12	588	64	241	78	65			238	18
Salicornia procumbens	73					19	15	23	15	1		
Trifolium pratense	7595	41	370	3085	680	1294	393	150	38	13	1317	214

Tabel 2.7 Opmerkingen over overeenkomst beschrijving voorkomen Heukels' in Flora districten tov waargenomen voorkomens in de dataset per fysisch geografische regio's. Zie tabel 2.3 voor vertaling fgr naar flora district.

wetenschappelijke naam	opmerkingen[3]
Alliaria petiolata	Moet algemeen voorkomen en zich uitbreiden over de hogere zandgronden, maar komt vooral voor in Ri (F), Du (R,W) en HzN (G,S,D)
Arctium lappa	Met name Ri (F) klopt, maar toch ook aanzienlijk aantal waarnemingen in Lv (L)
Dianthus deltoides	Is met name fluviatiel (Ri -> F), stroomgebieden van Overijsselse Vecht en de Dinkel, zou zeldzaam moeten zijn in het Pleistocene district (D,G,K,S,V) maar toch liggen in Hz N(G,S,D) 34% van de waarnemingen
Elymus athericus	Klopt, alleen is niet echt zeldzaam in Lv (L)
Hypericum humifusum	Is soort van Pleistocene districten, maar niet zoals aangegeven voor het heuvelland (maar 1 waarneming)
Juncus acutiflorus	
Lotus corniculatus subsp. corniculatus	
Mercurialis perennis	
Myosotis palustris	Klopt, moet vochtig zijn dus waarnemingen liggen vooral in laag Nederland
Rubus fruticosus	Klopt zeer goed voor het Pleistocene district (D,G,K,S,V), maar klopt niet echt voor Hl (Z) en Zk (E)
Rumex conglomeratus	
Salicornia procumbens	Klopt ten dele, er liggen alleen relatief veel waarnemingen in Lv (L)
Trifolium pratense	

Verspreiding vergeleken met de ecologische groepsindeling

Aan de hand van de indeling van de geselecteerde soorten in ecologische groepen (volgens Heukel) is de indeling van tabel 2.8 gemaakt. In de ecologische groepen worden soorten gekenmerkt naar zoutgehalte (saliniteit), vegetatiestructuur en successiestadium, vochttoestand, trofiegraad en zuurgraad, en eventueel additionele informatie die iets zegt over bijvoorbeeld de dynamiek. Aan de hand van deze indeling kunnen de soortwaarnemingen uit

[3] De codering tussen haakjes achter de codering voor de fysisch geografische regio's is de codering zoals gebruikt in de beschrijving van de Flora Districten (Van der Meijden, 1990)

de dataset vergeleken worden met de algemene omschrijving volgens Heukels'. In de tabellen 2.9 en 2.10 staan de resultaten uit de dataset naast de algemene afleiding uit de ecologische groepen. Per continue variabelen is de minimale, maximale en gemiddelde waarden afgeleid, voor de nominale variabele 'vegetatietype' is per nominale waarde de fractie van het aantal waarnemingen ten opzichte van het totaal aantal waarnemingen per soort bepaald.

Tabel2.8 Ecologische groepsindeling volgens Van der Meijden (1990) van de 13 willekeurig getrokken soorten

wetenschappelijke naam	eco-grp	saliniteit	vegetatiestructuur	vochttoestand	trofie en zuurgraad	additioneel
Alliaria petiolata	H47		bos en struweel	vochtig	matig voedselrijk	
	H48		bos en struweel	vochtig	zeer voedselrijk	
	H69		bos en struweel	droog	matig tot zeer voedselrijk	
Arctium lappa	R48		ruigte	vochtig	zeer voedselrijk	
Dianthus deltoides	G62		grasland	droog	voedselarm zwak zuur	
Elymus athericus	bP60st	brak	pioniervegetatie	droog		stuivend
	bR40	brak	ruigte	vochtig		
	R64		ruigte	droog	voedselarm	
Hypericum humifusum	P42		pioniervegetatie	vochtig	voedselarm zwak zuur	
Juncus acutiflorus	G22		grasland	nat	voedselarm zwak zuur	
	G27		grasland	nat	matig voedselrijk	
Lotus corniculatus subsp. corniculatus	G43		grasland	vochtig	voedselarm basisich	
	G47		grasland	vochtig	matig voedselrijk	
	G62		grasland	droog	voedselarm zwak zuur	
	G63		grasland	droog	voedselarm basisich	
	G67		grasland	droog	matig voedselrijk	
Mercurialis perennis	H43		bos en struweel	vochtig	voedselarm basisich	
Myosotis palustris	G28		grasland	nat	zeer voedselrijk	
	R28		ruigte	nat	zeer voedselrijk	
Rubus fruticosus	R44		ruigte	vochtig	voedselarm	
	R47		ruigte	vochtig	matig voedselrijk	
	R64		ruigte	droog	voedselarm	
	R67		ruigte	droog	matig voedselrijk	
	H41		bos en struweel	vochtig	voedselarm zuur	
	H42		bos en struweel	vochtig	voedselarm zwak zuur	
	H47		bos en struweel	vochtig	matig voedselrijk	
	H61		bos en struweel	droog	voedselarm zuur	
	H62		bos en struweel	droog	voedselarm zwak zuur	
	H69		bos en struweel	droog	matig tot zeer voedselrijk	
Rumex conglomeratus	G28		grasland	nat	zeer voedselrijk	
	H28		bos en struweel	nat	zeer voedselrijk	
Salicornia procumbens	zP20	zilt	pioniervegetatie	nat		
Trifolium pratense	G47		grasland	vochtig	matig voedselrijk	
	G48		grasland	vochtig	zeer voedselrijk	

Voor de variabele ***zoutgehalte*** wordt het gemiddelde Ellenberg zoutgetal vergeleken met de saliniteit beschrijving uit de ecologische groepen (tabel 2.9). *Elymus athericus* is een soort die in zowel brakke (P60stb, R40b) als in wat zoetere systemen (R64) voor moet kunnen komen. De waarnemingen liggen tussen zoutgetallen van 0.2 en 7.8 met een gemiddelde van

2.8. Dus gemiddeld liggen de waarde in een brak tot zout gebied met enkele waarnemingen in zoete (0.2) gebieden en enkele waarnemingen in een zeer zout gebied (7.8). Voor *Salicornia procumbens* een soort van het zilte gebied, verwacht je een gemiddelde waarde die hoger zou zijn dan voor *Elymus athericus*. Dit klopt, de gemiddelde waarde is 7.4 en de waarnemingen liggen tussen 2.6 en 8.5. De overige soorten zijn, volgens de ecologische groepsindeling, soorten uit een zoet milieu en de gemiddelde waarden voor het zoutgetal van deze soorten liggen allemaal onder de 0.4 en zijn dus 'zoete' soorten.

De verdeling over de ***vegetatietypen*** wordt vergeleken met de vegetatiestructuur zoals omschreven in de ecologische groepsindeling (tabel 2.9). Omdat sommige soorten kenmerkend zijn voor meer dan één vegetatiestructuurtype is het niet altijd zo dat de berekende fractie van de waarnemingen per vegetatietype altijd één op één is met de vegetatiestructuurtype, maar over het algemeen komen de waarnemingen met de indeling goed overeen. Bijvoorbeeld *Lotus corniculatus subsp. corniculatus*, een grasland type, komt in 90% van de waarnemingen ook in het vegetatietype grasland voor. *Rubus fruticosus* een soort kenmerkend voor zowel ruigte als bos en struweel wordt voor 70% en 18% in de waarnemingen toegewezen aan bos (loofbos) en grasland. De indeling 'ruigte' komt waarschijnlijk overeen met grasland en de indeling 'bos en struweel' is waarschijnlijk hoofdzakelijk bos met misschien wat heide en/of gras.

Tabel 2.9 Vergelijking van Ellenberg zoutgetal (minimum, maximum en gemiddelde) t.o.v. saliniteit uit ecologische groepsindeling (Van der Meijden, 1990) en fracties voorkomen in vegetatietypen t.o.v. vegetatiestructuur en successiestadium (Van der Meijden, 1990).

		zoutgehalte				vegetatietype[4]					
wetenschappelijke naam	eco-grp	min	max	gem	saliniteit	dec	pin	spr	hea	grp	vegetatiestructuur
Alliaria petiolata	H47; H48; H69	0.00	0.93	0.09		0.71	0.07	0.00	0.00	0.22	bos en struweel
Arctium lappa	R48	0.00	0.71	0.18		0.32	0.05	0.00	0.00	0.63	ruigte
Dianthus deltoides	G62	0.06	1.17	0.40		0.00	0.01	0.00	0.01	0.99	grasland
Elymus athericus	bP60st	0.19	7.75	2.84	brak	0.02	0.07	0.00	0.00	0.91	pioniervegetatie
	bR40	0.19	7.75	2.84	brak	0.02	0.07	0.00	0.00	0.91	ruigte
	R64	0.19	7.75	2.84		0.02	0.07	0.00	0.00	0.91	ruigte
Hypericum humifusum	P42	0.00	0.89	0.30		0.17	0.01	0.02	0.01	0.78	pioniervegetatie
Juncus acutiflorus	G22; G27	0.00	1.36	0.27		0.04	0.00	0.00	0.01	0.95	grasland
Lotus corniculatus subsp. corniculatus	G43; G47; G62; G63; G67	0.00	4.86	0.33		0.04	0.05	0.00	0.02	0.90	grasland
Mercurialis perennis	H43	0.00	0.21	0.01		0.98	0.00	0.00	0.00	0.02	bos en struweel
Myosotis palustris	G28	0.00	1.69	0.28		0.07	0.00	0.00	0.00	0.93	grasland
	R28	0.00	1.69	0.28		0.07	0.00	0.00	0.00	0.93	ruigte
Rubus fruticosus	R44; R47; R64; R67	0.00	4.88	0.14		0.70	0.06	0.03	0.01	0.18	ruigte
	H41; H42; H47; H61; H62; H69	0.00	4.88	0.14		0.70	0.06	0.03	0.01	0.18	bos en struweel
Rumex conglomeratus	G28	0.00	4.19	0.29		0.23	0.01	0.00	0.00	0.77	grasland
	H28	0.00	4.19	0.29		0.23	0.01	0.00	0.00	0.77	bos en struweel
Salicornia procumbens	zP20	2.59	8.50	7.42	zilt	0.00	0.00	0.00	0.00	1.00	pioniervegetatie
Trifolium pratense	G47; G48	0.00	5.25	0.38		0.02	0.00	0.00	0.00	0.97	grasland

[4] Vegetatietypen in de dataset: dec – loofbos, pin – dennenbos, spr – sparrenbos, hea – heide, grp – arm grasland.

De ecologische groepsindeling maakt onderscheid in vier ***vochttoestanden*** (tabel 2.9): aquatisch, nat, vochtig en droog. Het bereik van de Ellenberg waarden ligt tussen 1 en 12, waarbij 1 kurkdroog is en 12 overeenkomt met continue of bijna continue onder water staande planten. Gebruikmakend van de indeling van Ellenberg komen we tot de volgende indeling: droog: 1 – 3, vochtig: 4 – 7, nat: 8 – 10 en aquatisch: 11 – 12. Door op deze manier naar de vochtgetallen te kijken, waarbij zowel de minimale, maximale en gemiddelde waarden worden vergelijken met de vochttoestand blijkt voor de meeste soorten de indeling goed overeen te komen. Eén soort is op deze wijze wat lastiger in een groep te plaatsen. *Dianthus deltoides* hoort in een droog milieu thuis, maar ligt iets meer bij de overgang droog-vochtig; het bereik ligt tussen 3.1 en 6.0 met een gemiddelde van 4.2. De overige soorten vallen allemaal zeer goed in de beschreven klassen, sommige soorten, zoals *Rubus fruticosus* zijn zeer vochttolerant en kunnen in ‘ieder’ milieu voorkomen. Dit blijkt ook uit de waarden in de dataset, tussen de 2.6 en 10.0 met een gemiddelde van 6.2.

De ***trofiegraad*** (tabel 2.10) beschrijft eigenlijk twee variabelen, het **zuur-** en **nutriënten**getal. Er worden 7 gecombineerde klassen onderscheiden van ‘voedselarm zuur’ tot ‘matig tot zeer voedselrijk’. Wanneer alleen naar de beschikbare hoeveelheid aanwezige nutriënten gekeken wordt, worden er 4 klassen onderscheiden: ‘voedselarm’, ‘matig voedselrijk’, ‘zeer voedselrijk’ en de combinatieklasse ‘matig tot zeer voedselrijk’. Voor het nutriëntengetal gebruikt Ellenberg (Ellenberg *et al.*, 1992) de volgende indeling: 1: Zeer nutriëntenarm, 3: Nutriëntenarm, 5: Matig nutriëntenrijk, 7: Nutriëntenrijk, 8: Uitgesproken nutriënten indicator en 9: Zeer uitgesproken nutriënten indicator. Binnen de trofiegraad worden de drie verschillende klassen voor zuurgraad onderscheiden, dit zijn de klassen: ‘zuur’, ‘zwak zuur’ en ‘basisch’. Het zuurgetal naar Ellenberg (Ellenberg *et al.*, 1992) is in 9 klassen opgedeeld: 1: Sterkzuur, 3: Zuur, 5: Matig zuur, 7: Zwakzuur tot zwak basisch en 9: Basisch en kalkrijk. De klassen 2, 4, 6, en 8 zijn tussenliggende klassen.
Wanneer de drie zuurklassen van de trofiegraad binnen deze Ellenbergindeling ingepast worden kan dit gedaan worden door de Ellenberg range (1-9) in drie gelijke porties te verdelen. De drie klassen ‘zuur’, ‘zwak zuur’ en ‘basisch’ komen dan overeen met Ellenberg ranges van 1.0 – 3.7, 3.7 – 6.3 en 6.3 – 9.0.
Met deze gebruikte indeling/vertaling komen voor de meeste soorten het bereik van de Ellenberg zuurgraad en nutriënten goed overeen met de ‘geclassificeerde’ grenzen. Alleen de planten die volgens de trofiegraad-indeling alleen in het voedselrijke traject voor moeten komen geven minder goed resultaat. *Rumex conglomeratus* heeft bijvoorbeeld een gemiddeld N-getal van 6.25, maar ook waarnemingen met een minimum van 3.85 die ver buiten het minimale bereik volgens de trofiegraad (6.3) liggen.
Er zijn ook soorten die zeer goed in de trofiegraad indeling passen. Een voorbeeld hiervan is *Rubus fruticosus* (r: bereik 1.0 - 8.0 en gemiddelde 4.7; n: bereik 1.0 – 8.5 en gemiddelde 5.0) met een trofiegraad indeling voor zuurgraad van 1 tot 6.3 en nutriënten van 1 tot 6.3. Een ander voorbeeld is *Lotus corniculatus subsp. corniculatus* (r: bereik 2.3 – 7.6 en gemiddelde 5.9; n: bereik 1.9 – 6.8 en gemiddelde 4.1) met een indeling naar zuurgraad en nutriënten volgens de trofiegraad van respectievelijk 3.7 tot 9.0 en 1.0 tot 9.0.
Het vergelijken van de zuurgraad is lastig, omdat niet bij iedere ecologische groepsindeling een indicatie voor zuurgraad beschreven staat. Maar kijkend naar de soorten waarbij wel iets over de zuurgraad vermeld staat klopt ook deze beschrijving goed met de waarnemingen uit de steekproef.

Tabel 2.10 Vergelijking van Ellenberg vocht-, zuur- en stikstofgetal (minimum, maximum en gemiddelde) t.o.v. de vochttoestand en trofietoestand uit ecologische groepsindeling (Van der Meijden, 1990). Bovendien staat ook nog de additionele kenmerken in deze tabel vermeld.

		vocht				zuurgraad			nutriënten				
wetenschappelijke naam	eco-grp	min	max	gem	vocht	min	max	gem	min	max	gem	trofie	additioneel
Alliaria petiolata	H47	3.33	10.06	5.76	vochtig	3.43	8.00	6.56	3.11	8.50	6.94	matig voedselrijk	
	H48	3.33	10.06	5.76	vochtig	3.43	8.00	6.56	3.11	8.50	6.94	zeer voedselrijk	
	H69	3.33	10.06	5.76	droog	3.43	8.00	6.56	3.11	8.50	6.94	matig tot zeer voedselrijk	
Arctium lappa	R48	4.21	7.88	6.01	vochtig	3.60	7.52	6.69	4.82	8.33	6.86	zeer voedselrijk	
Dianthus deltoides	G62	3.14	6.00	4.16	droog	3.25	6.44	4.48	2.13	5.04	3.23	voedselarm zwak zuur	
Elymus athericus	bP60st	3.00	8.00	5.46	droog	3.40	8.00	6.71	2.29	8.00	5.34		stuivend
	bR40	3.00	8.00	5.46	vochtig	3.40	8.00	6.71	2.29	8.00	5.34		
	R64	3.00	8.00	5.46	droog	3.40	8.00	6.71	2.29	8.00	5.34	voedselarm	
Hypericum humifusum	P42	4.35	7.72	5.98	vochtig	2.40	6.25	4.33	2.20	6.76	4.29	voedselarm zwak zuur	
Juncus acutiflorus	G22	4.74	10.42	7.73	nat	2.17	7.17	4.88	1.57	7.08	3.87	voedselarm zwak zuur	
	G27	4.74	10.42	7.73	nat	2.17	7.17	4.88	1.57	7.08	3.87	matig voedselrijk	
Lotus corniculatus subsp. corniculatus	G43	2.90	9.80	4.73	vochtig	2.33	7.64	5.92	1.90	6.83	4.12	voedselarm basisch	
	G47	2.90	9.80	4.73	vochtig	2.33	7.64	5.92	1.90	6.83	4.12	matig voedselrijk	
	G62	2.90	9.80	4.73	droog	2.33	7.64	5.92	1.90	6.83	4.12	voedselarm zwak zuur	
	G63	2.90	9.80	4.73	droog	2.33	7.64	5.92	1.90	6.83	4.12	voedselarm basisch	
	G67	2.90	9.80	4.73	droog	2.33	7.64	5.92	1.90	6.83	4.12	matig voedselrijk	
Mercurialis perennis	H43	4.75	8.67	5.57	vochtig	5.27	7.50	6.54	4.69	7.67	6.36	voedselarm basisch	
Myosotis palustris	G28	4.88	11.25	8.73	nat	3.00	8.00	6.41	2.79	7.75	5.78	zeer voedselrijk	
	R28	4.88	11.25	8.73	nat	3.00	8.00	6.41	2.79	7.75	5.78	zeer voedselrijk	
Rubus fruticosus	R44	2.60	10.00	6.20	vochtig	1.00	8.00	4.67	1.00	8.50	5.00	voedselarm	
	R47	2.60	10.00	6.20	vochtig	1.00	8.00	4.67	1.00	8.50	5.00	matig voedselrijk	
	R64	2.60	10.00	6.20	droog	1.00	8.00	4.67	1.00	8.50	5.00	voedselarm	
	R67	2.60	10.00	6.20	droog	1.00	8.00	4.67	1.00	8.50	5.00	matig voedselrijk	
	H41	2.60	10.00	6.20	vochtig	1.00	8.00	4.67	1.00	8.50	5.00	voedselarm zuur	
	H42	2.60	10.00	6.20	vochtig	1.00	8.00	4.67	1.00	8.50	5.00	voedselarm zwak zuur	
	H47	2.60	10.00	6.20	vochtig	1.00	8.00	4.67	1.00	8.50	5.00	matig voedselrijk	
	H61	2.60	10.00	6.20	droog	1.00	8.00	4.67	1.00	8.50	5.00	voedselarm zuur	
	H62	2.60	10.00	6.20	droog	1.00	8.00	4.67	1.00	8.50	5.00	voedselarm zwak zuur	
	H69	2.60	10.00	6.20	droog	1.00	8.00	4.67	1.00	8.50	5.00	matig tot zeer voedselrijk	
Rumex conglomeratus	G28	4.15	10.62	7.20	nat	3.40	8.00	6.47	3.85	8.22	6.25	zeer voedselrijk	
	H28	4.15	10.62	7.20	nat	3.40	8.00	6.47	3.85	8.22	6.25	zeer voedselrijk	
Salicornia procumbens	zP20	6.32	10.00	8.06	nat	6.89	7.67	7.26	3.50	6.88	5.17		
Trifolium pratense	G47	3.29	9.08	5.77	vochtig	2.60	8.00	6.07	2.29	7.56	5.26	matig voedselrijk	
	G48	3.29	9.08	5.77	vochtig	2.60	8.00	6.07	2.29	7.56	5.26	zeer voedselrijk	

2.5 Conclusie

Uit het korte onderzoek uit dit hoofdstuk kan geconcludeerd worden dat de dataset de verspreiding en verdeling van de plantensoorten in Nederland redelijk goed beschrijft. De dataset is niet volkomen onafhankelijk opgezet, op plaatsen waar een grote variatie en dus dynamiek aanwezig is, is de opnamedichtheid hoger dan op plaatsen met een lage dynamiek. Ook is het zo dat er binnen Nederland plekken zijn die minder goed onderzocht zijn, maar waarvan op basis van de variatie wel meer opnamepunten verwacht zouden worden.
Omdat er heel veel opnamepunten zijn en deze een goed beeld geven van de plekken waar soorten voorkomen, zowel de onderwaarden als de bovenwaarden zijn aanwezig, is de verwachting dat deze verzameling gegevens een goede afspiegeling geeft van de ruimtelijke variatie in Nederland. Ook zal door de multiple lineaire regressiemethode die gebruikt gaat worden om de modellen af te leiden de eventueel aanwezige bias waarschijnlijk uitgemiddeld worden.
Concluderend kan gesteld worden dat de onderzochte dataset van Schaminée *et al.* (1995) geschikt is om regressiemodellen mee af te leiden.
Voor extra informatie over de relatie tussen responsies en de dataverzameling biedt het rapport ‘Afstemming biotische responsmodules DEMNAT-SMART/MOVE’ van Runhaar *et al.* (2002) een goede aanvulling. In dit rapport zijn de MOVE regressies van De Heer *et al.* (2000) gebruikt.

Deel II

Analyse van de modellen

3. Regressiemodellen

3.1 Inleiding regressiemodellen[5]

Het centrale idee achter regressiemodellen is afhankelijkheid: een variabele die op één of andere manier van één of meerdere andere variabelen afhangt (Webster en Oliver, 1990). Uit de regressie op zich is niet af te leiden of er een direct oorzakelijk verband bestaat tussen variabelen. Bovendien maakt het uit het oogpunt van de regressieanalyse niet uit welke variabele oorzaak of gevolg is. Regressiemodellen worden gebruikt voor het beschrijven van data, schatten van parameters, voorspellen en schatten van waarden en regelen van instellingen (Montgomery en Peck, 1992). In dit rapport worden regressiemodellen gebruikt om het bestaan van een relatie tussen verschillende variabelen te schatten en uiteindelijk voorspellingen te kunnen maken. Regressieanalyse is een statistische methode die gebruikt kan worden om relaties tussen soorten en hun omgeving te onderzoeken op basis van soort waarnemingen en omgevingsvariabelen op een serie meetpunten (Jongman *et al.*, 1987). Er wordt bijvoorbeeld een relatie verondersteld tussen het wel of niet voorkomen van planten en de zuurgraad van de bodem. De zuurgraad van de bodem heeft een directe invloed op het voorkomen van planten, maar ook een indirecte invloed omdat de zuurgraad van invloed is op de in het bodemvocht opgeloste stoffen. De toepassing van regressieanalyse gaat uit van twee type variabelen, de afhankelijke (de variabele die geschat moet worden) en de onafhankelijke variabele(n). De onafhankelijke variabele(n) worden ook wel covariaten genoemd (Hosmer en Lemeshow, 1989). Regressieanalyse kan behalve voor het schatten ook gebruikt worden voor interpolatie, kalibratie en in bepaalde gevallen voor beschrijving (Webster en Oliver, 1990).
Het doel van regressieanalyse is het beschrijven van de respons variabele als een functie van één of meerdere verklarende variabelen. Regressieanalyse is gebaseerd op een respons model dat uit twee delen bestaat: een deel dat beschrijft hoe de verwachte respons afhangt van de verklarende variabelen (het systematische deel) en een deel dat beschrijft hoe de waargenomen respons afwijkt van de verwachte respons (de fout). Het systematische deel kan beschreven worden met een regressievergelijking. De fout kan beschreven worden door de statistische verdeling van de fout. Wanneer een rechte lijn gefit wordt door data heeft het respons model de volgende vorm:

$$y = b_0 + b_1 x + \varepsilon \qquad (3.1)$$

met y de response variabele
x de verklarende variabele
ε de fout
b_0 en b_1 vaste maar onbekende coëfficiënten; de intercept en helling parameters

De verwachte respons **(**E(Y|x)**)** is gelijk aan $b_0 + b_1$ x. De waarde E(Y|x) moet gelezen worden als ‘de verwachte waarde van Y, gegeven de waarde x’. Het systematische deel van het model is dus een rechte lijn en wordt gespecificeerd door de regressievergelijking:

$$E(Y \mid x) = b_0 + b_1 x. \qquad (3.2)$$

[5] Veel van de in deze paragraaf beschreven tekst is overgenomen uit Jongman *et al.* (1987) en Webster en Oliver (1990)

De fout is de verdeling van ε, de willekeurige variatie van de waargenomen response om de verwachte respons. Het doel van regressieanalyse is dus het schatten van het systematische deel uit de data daarbij rekening houdend met de fout van het model.
Een regressie techniek die geschikt is voor aan- en afwezigheid (binaire) gegevens is de logistische regressie. Bij logistische regressie wordt geprobeerd om de waarschijnlijkheid van voorkomen van een soort uit te drukken als een functie van de verklarende variabelen.

3.1.1 De logistische regressie

Er zijn twee type verklarende variabelen: de nominale en de continue. Wanneer het wel of niet voorkomen afhangt van een nominale variabelen dan kan met behulp van de chi-kwadraat test nagegaan worden of het voorkomen van een specifieke soort bepaald wordt door één of meerdere nominale waarde(n). Bij continue gegevens geldt net als bij nominale gegevens dat de verwachte respons gelijk is aan de waarschijnlijkheid van voorkomen van de soort op een plek met een specifieke waarde van de omgevingsvariabele. Met behulp van een kromme kan deze waarschijnlijkheid beschreven worden, dus vergelijking 3.2 voldoet niet omdat deze ook negatieve waarden kan aannemen en kansen per definitie tussen de 0 en 1 liggen. Door de exponent van vergelijking 3.2 te nemen kan dit probleem opgelost worden:

$$E(Y \mid x) = \exp(b_0 + b_1 x) \tag{3.3}$$

Alleen is het nu mogelijk dat de rechterkant van vergelijking 3.3 groter wordt dan 1, dus moet de functie nog eenmaal aangepast worden:

$$E(Y \mid x) = p = \frac{\left[\exp(b_0 + b_1 x)\right]}{\left[1 + \exp(b_0 + b_1 x)\right]} \tag{3.4}$$

Vergelijking 3.4 beschrijft een zogenaamde sigmoide (S) kromme. Alle drie de functies (3.2 t/m 3.4) zijn monotoon stijgend en hebben twee parameters. Het deel $b_0 + b_1 x$ wordt de lineaire schatter genoemd.
In het voorgaande is het systematische deel van de respons functie behandeld. Er rest nu alleen nog de fout. Omdat de respons maar twee waarden aan kan nemen (0 en 1) heeft de fout een binominale verdeling met som 1. De variantie van y is $p(1 - p)$. Het complete model is nu beschreven.
Om de parameters uit de gegevens te kunnen schatten kunnen we geen normale least-squares (minimale kwadranten) regressie gebruiken, omdat de fouten niet normaal verdeeld zijn en zij geen constante variantie hebben. In plaats daarvan wordt de logistische regressie gebruikt. De logistische regressie is een speciale vorm van het gegeneraliseerde lineaire model (GLM, McCullagh & Nelder 1983). De term logistisch komt van de gebruikte log transformatie, de transformatie van p:

$$g(x) = \ln\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 x \tag{3.5}$$

Vergelijking 3.5 is een andere manier om vergelijking 3.4 te schrijven.
Het belang van deze transformatie is dat g(x) veel van de gunstige eigenschappen van een lineair model heeft. De logit, g(x), is lineair in haar parameters, kan continue zijn en kan liggen tussen $-\infty$ en $+\infty$, afhankelijk van het bereik van x.

Bij regressie analyse met een binair afhankelijke variabele geldt (Hosmer en Lemeshow, 1989):

1) de conditionele gemiddelde van de responsievergelijking moet zo geformuleerd zijn dat deze tussen 0 en 1 ligt. Het logistische model voldoet hieraan.
2) de binomiale, en niet de normale, verdeling beschrijft de verdeling van de fouten en is de statistische verdeling waarop de analyse gebaseerd wordt.
3) het analyseprincipe zoals gebruikt bij de lineaire regressie worden ook gebruikt bij de logistische regressie.

3.1.2 Multipele regressie

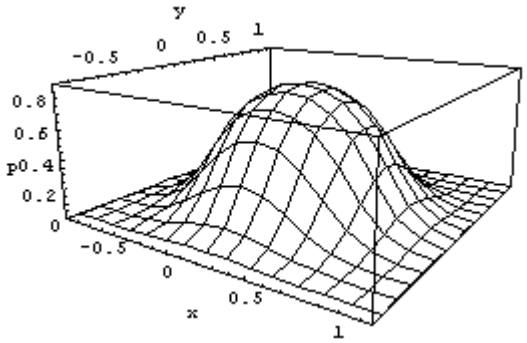
In de vorige paragraaf werd de responsvariabele uitgedrukt als een functie van één omgevingsvariabele. Een soort kan echter reageren op meer dan één omgevingsvariabelen. Dit kan met behulp van multipele regressie gemodelleerd worden. Wanneer er enige correlatie bestaat tussen de omgevingsvariabelen of wanneer er sprake is van interacties dan kunnen afzonderlijke analyses van de respons van elke omgevingsvariabelen afzonderlijk niet de multiple regressie vervangen. Interactie effecten treden op wanneer het effect van één variabele afhangt van de waarde van een andere variabele.
De vergelijking 3.5 in paragraaf 3.1.1 kan eenvoudig uitgebreid worden naar een multiple regressie waarbij de beschrijving van een lijn vervangen wordt door de beschrijving van een vlak:

$$\log_e\left[p/(1-p)\right] = b_0 + b_1 x_1 + b_2 x_2 \qquad (3.6)$$

Bovenstaande vergelijking kan eenvoudig uitgebreid met meer dan twee verklarende variabelen.
Kwadratische modellen ontstaan door in de lineaire schatter van vergelijking 3.6 twee kwadratische termen mee te nemen ($b_2x_1^2$ en $b_4x_2^2$).

$$E(Y \mid x) = b_0 + b_1 x_1 + b_2 x_1^{\,2} + b_3 x_2 + b_4 x_2^{\,2} \qquad (3.7)$$

Dit resulteert in een bivariaat Guassisch logit vlak, wanneer zowel b1 en b3 positief en b_2 en b_4 negatief zijn (figuur 3.1). Dit oppervlak heeft ellipsen als contourlijnen (lijnen van gelijke waarschijnlijkheid) en hoofdassen parallel aan de x_1 en x_2 assen.



Figuur 3.1 3D view van een bivariaat Gaussisch logit oppervlak, waarbij de waarschijnlijkheid van voorkomen (p) op de z-as staat en de twee verklarende variabelen x en y op het horizontale vlak staan.

Interactie tussen verklarende variabelen

Er is sprake van interactie tussen twee verklarende variabelen wanneer de invloed van de ene variabele afhankelijk is van de waarde van de andere variabele. Er kan getest worden of er interactie optreedt tussen twee variabelen door de regressievergelijking uit te breiden met producttermen, zoals $x_1 \cdot x_2$:

$$E(Y \mid x) = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2 = (b_0 + b_2 x_2) + (b_1 + b_3 x_2) x_1 \qquad (3.8)$$

Uit vergelijking 3.8 blijkt dat de relatie in dit model tussen E(Y|x) en x_1 nog steeds een rechte lijn is, maar dat de intercept en de helling en dus het effect van x_1 afhangen van de waarde van x_2. Omgekeerd geldt dat het effect van x_2 afhangt van de waarde van x_1. De interactie kan getest worden met behulp van een *t*-test met als nul hypothese dat b_3 gelijk is aan 0.
Door het Gaussische model van vergelijking 3.7 uit te breiden met een productterm krijgen we in het logit geval:

$$\log_e [p/(1-p)] = b_0 + b_1 x_1 + b_2 x_1^{\,2} + b_3 x_2 + b_4 x_2^{\,2} + b_5 x_1 x_2 \qquad (3.9)$$

Wanneer $b_2 + b_4 < 0$ en $4 b_2 b_4 - b_5^{\,2} > 0$, dan beschrijft vergelijking 3.9 een unimodaal oppervlak met ellipsvormige contourlijnen. Maar in tegenstelling tot figuur 3.1 geldt nu niet de restrictie dat de hoofdassen horizontaal of vertikaal lopen. Wanneer aan één van deze voorwaarden niet voldaan wordt, dan beschrijft vergelijking 3.9 een oppervlak met een enkel minimum of een oppervlak met een zogenaamd zadelpunt (bijv. Carroll 1972). Wanneer het oppervlak unimodaal is dan kan het globale optimum (u_1, u_2) berekend worden uit de coëfficiënten van vergelijking 3.9 door middel van:

$$\begin{aligned} u_1 &= (b_5 b_3 - 2\, b_1 b_4)/d \\ u_2 &= (b_5 b_1 - 2\, b_3 b_2)/d \end{aligned} \qquad (3.10)$$

waarbij $d = 4 b_2 b_4 - b_5^{\,2}$.

Het optimum met betrekking tot x_1 voor een specifieke waarde van x_2 is $-(b_1 + b_5 x_2)/(2\, b_2)$ en hangt dus af van de waarde van x_2 wanneer $b_5 \neq 0$.

Nominaal verklarende variabelen

Het is ook mogelijk om multiple regressies te gebruiken om het gelijktijdig optredende effect te onderzoeken van nominale omgevingsvariabelen of van zowel kwantitatieve en nominale omgevingsvariabelen. Afhankelijk van de gebruikte statistische software worden nominale variabelen omgezet in dummy of design (Hosmer en Lemeshow,1989) variabelen ook wel contrasten genoemd. Binnen S-Plus wordt standaard met Helmert contrasten gewerkt, een hele simpele methode is de ‘treatment’ contrast, waarbij elke klasse een dummy variabele is die de overige klassen uitsluit. Bijvoorbeeld de nominale variabele bodemtype met drie klassen: klei, veen en zand. Bij de ‘treatment’ contrast wordt bijvoorbeeld klei als referentie klassen genomen en definiëren we voor veen en zand twee dummy variabelen x_1 en x_2 wiens waarden of 0 of 1 zijn. De klasse veen wordt nu beschreven door voor variabele x_1 een 1 in te vullen en voor x_2 een 0 in te vullen. Dit resulteert in onderstaand schema:

	x_1	x_2
klei	0	0
veen	1	0
zand	0	1

Bij de Helmert contrasten wordt er van een ander schema gebruik gemaakt om de nominale variabele om te zetten in dummy variabelen. Het bovenstaande schema ziet er bij de opdeling volgens Helmert als volgt uit:

	x_1	x_2
klei	-1	-1
veen	1	-1
zand	0	2

Het systematische deel van het model kan gebruik makend van de introductie van dummy variabelen als volgt geschreven worden:

$$E(Y \mid x) = b_1 + b_2 x_2 + b_3 x_3 \qquad (3.11)$$

In dit voorbeeld geeft de coëfficiënt b_1 de verwachte respons van de referentieklas klei, de coëfficiënt b_2 geeft het verschil in verwachte respons tussen veen en klei en coëfficiënt b_3 geeft het verschil tussen zand en klei.

3.1.3 Modelkeuze en regressie diagnostiek

Vroege studies door Gause (1930), Curtis en McIntosh (1951) en Whittaker (1956) laten zien dat monotone respons curven te eenvoudig zijn als ecologisch respons model en dat een unimodaal model beter geschikt is. Door eenvoudig ecologisch te redeneren kan duidelijk gemaakt worden dat ook bimodale curven een realistische optie kunnen zijn. Een soort kan weggeconcurreerd zijn op haar fysiologisch optimum door beter concurrerende soorten, terwijl de soort zich wel weet te handhaven bij minder geschikte omgevingsvoorwaarden wanneer de competitie minder groot is. De resulterende responsie curve die op deze veldomstandigheden gefit is, is dan het resultaat van de fysiologische responsie curve en de concurrentie tussen soorten (Fresco, 1982). Hill (1977) oppert daarentegen dat een goed gekozen ecologische variabele, het voorkomen van bimodale verdelingen minimaliseert. Responsie curven hebben enkele voordelen ten opzichte van frequentie profielen voor kwantitatieve omgevingsvariabelen:

- eenvoudige kromme geven door middel van hun parameters een compactere beschrijving dan frequentieprofielen
- er hoeven geen arbitraire klassengrenzen gekozen te worden
- er is geen verschil van gegevens omdat de omgevingsvariabelen niet in klassen ingedeeld hoeven te worden
- wanneer het Gaussische model toepasbaar is, dan kunnen statistische testen die gebaseerd zijn op de curven beter analyseren of de omgevingsvariabelen de soort beïnvloedt dan de chi-kwadraat test die gebaseerd is op het frequentieprofiel.

Een duidelijk nadeel van krommen is dat er een model gekozen kan worden dat niet geschikt is voor de beschikbare gegevens.
Datapunten die de regressie overmatig beïnvloeden moeten met extra grote nauwkeurigheid bekijken worden. Wanneer er twee of meer verklarende variabelen zijn, dan is het verstandig om de variabelen twee aan twee tegen elkaar uit te zetten, zodat eventuele geïsoleerde

uitlopers gedetecteerd kunnen worden. Wanneer zulke uitlopers aanwezig zijn moet nagegaan worden of dit een opnamefout is of dat de opnameplek atypisch was. Hierna moet besloten worden of de uitloper in de verzameling gegevens kan blijven. Een extra controle is om de laagste en hoogste *x* uit de plekken waar een soort voorkomt te verwijderen om te controleren of de gefitte respons min of meer gelijk blijft (vergelijk de 'Jackknife' techniek, Efron 1982).

3.2 De verschillende uitgerekende modellen

In hoofdstuk 2.2 van deel I staat beschreven welke variabelen een rol kunnen spelen bij het verklaren en uiteindelijk voorspellen van het voorkomen van specifieke plantensoorten op bepaalde locaties. Als vertrekpunt van dit onderzoek is het MOVE 2 model genomen. In dit multiple lineaire regressiemodel wordt gebruik gemaakt van variabelen die de zuurgraad (r), nutriëntenbeschikbaarheid (n) en vochtgehalte (f) en de onderlinge interacties (interactietermen) daartussen beschrijven. Deze studie onderzoekt wat het toevoegen van nieuwe omgevingsvariabelen op het gebied van beheer, toxiciteit, saliniteit en ruimtelijke ligging bijdraagt aan het verbeteren van het oorspronkelijke model. Het betreft de variabelen vegetatiestructuur (veg), potentieel aangetaste fractie plantensoorten door zware metalen (cpaf2), de saliniteit (Ellenberg-s) en fysisch geografische regio's (fgr).
In de vegetatiestructuur komen gevolgen van beheer, zowel direct als indirect, historische factoren en bijvoorbeeld natuurlijke successie naar voren. De fysisch geografische regio's worden meegenomen om de ruimtelijke afwisseling en preferenties van soorten mee te nemen. De indeling naar fgr's beschrijft onder andere verschillen in substraat en verschillen in processen (rivier of kust). Deze variaties verwachten wij ook terug te zien in de lokaties waarop bepaalde plantensoorten voorkomen. Het zoutgetal (Ellenberg-s) beschrijft de tolerantie van planten ten opzichte van zout. Het zware metalengetal (cpaf2) is een maat voor de toxiciteit van zware metalen voor planten, met deze variabele willen wij corrigeren voor eventueel te positief ingeschatte waarden.

Deze vier extra variabelen kunnen in 16 (2^4) verschillende combinaties samengevoegd worden. Dit zou betekenen dat er in totaal minimaal 14624 modellen (914 soorten * 16 modellen) doorgerekend moeten worden. Hierbij zijn nog niet de modellen meegenomen die opnieuw uitgerekend moeten worden om de niet unimodalen oplossingen (in het vervolg bimodale modellen genoemd) eruit te filteren. Bimodale modellen zijn modellen die als artefact een minimum hebben, omdat er 'gedwongen' een kwadratische functie gebruikt is om het model te fitten, terwijl een rechte lijn voor deze variabele waarschijnlijk beter recht doet aan het te modelleren gedrag. Deze modellen worden vervangen door één voor één de kwadratische termen uit de regressie te verwijderen (zie §3.3).
Wanneer niet het MOVE 2-model als uitgangspunt genomen zou worden maar alle mogelijke combinaties doorgerekend zouden worden dan moeten er 128 (2^7) (wanneer er bij een lineaire term altijd een kwadratische term hoort), of 1024 (2^{10}) (wanneer er kwadratische termen mogen bestaan zonder lineaire term) of 8192 (2^{13}) (wanneer er ook nog gevarieerd kan worden in het wel of niet meenemen van de interactietermen) combinaties doorgerekend worden. Het doorrekenen van alle mogelijke modellen is niet mogelijk, zonder gebruik te maken van stapsgewijze regressiemethoden (zie hoofdstuk 4).

In dit hoofdstuk wordt het afzonderlijk doorrekenen van de verschillende modellen beschreven die het MOVE 2-model als uitgangspunt hebben. Om het aantal door te rekenen modellen enigszins te beperken zijn er 11 verschillende modellen doorgerekend (zie tabel 3.1). Aan het basismodel (MOVE 2-model) zijn alle variabelen afzonderlijk toegevoegd (5 modellen, inclusief het basismodel). Hierna is op basis van de gemiddelde verklaarde

deviantie het meest optimale model geselecteerd (model d) en zijn op basis van dit model (basismodel 2) 3 nieuwe afgeleide modellen doorgerekend die allen één extra variabelen toevoegen. Dit is herhaald totdat alle variabelen toegevoegd zijn. Voor de verschillende extra variabelen zijn in tegenstelling tot de variabelen die in het MOVE 2-model zijn doorgerekend geen kwadratische termen doorgerekend. Dit is met name om praktische redenen gedaan en min of meer gedwongen door de dataset; van beide variabelen liggen bijna alle waarnemingen in de rand van hun bereik (zie de figuren 2.5 en 2.7). Door alleen een lineair verband te modelleren wordt er alleen nog maar onderscheid gemaakt in wel of niet zouttolerant en in meer of mindere mate ongevoelig voor de toxiciteit als gevolg van zware metalen. Door gebruik te maken van geklassificeerde variabelen voor zouttolerantie en zware metalen tolerantie, wel of niet tolerant, is dit effect ook te bereiken. Maar omdat dan een zekere arbitraire grens getrokken zou moeten worden is ervoor gekozen om toch de continue variabelen te gebruiken en deze lineair in het regressiemodel mee te nemen.
Het MOVE 2-model is gebruikt om een startpunt te krijgen waarmee de verkregen resultaten vergeleken kunnen worden en om inzicht te krijgen in de toegevoegde waarden van deze extra variabelen.

Tabel 3.1 Overzicht van de verschillende onderzochte modelvarianten

model	functie	opmerkingen
model a	y = f(r, n, f)	basismodel 1 (§3.2.1)
model b	y = f(r, n, f, cpaf2)	basismodel 1 plus toxiciteit agv zware metalen (§3.2.2)
model c	y = f(r, n, f, veg)	basismodel 1 plus vegetatietype (§3.2.2)
model d	y = f(r, n, f, fgr)	basismodel 1 plus fysisch geografische regio's → Basismodel 2 (§3.2.2)
model e	y = f(r, n, f, s)	basismodel 1 plus zoutgehalte (§3.2.2)
model f	y = f(r, n, f, fgr, cpaf2)	basismodel 2 plus toxiciteit agv zware metalen (§3.2.3)
model g	y = f(r, n, f, fgr, veg)	basismodel 2 plus vegetatietype (§3.2.3) → Basismodel 3
model h	y = f(r, n, f, fgr, s)	basismodel 2 plus zoutgehalte (§3.2.3)
model i	y = f(r, n, f, fgr, veg, cpaf2)	basismodel 3 plus toxiciteit agv zware metalen (§3.2.4)
model j	y = f(r, n, f, fgr, veg, s)	basismodel 3 plus zoutgehalte (§3.2.4) → Basismodel 4
model z	y = f(r, n, f, fgr, veg, s, cpaf2)	volledig model (§3.2.5)

In de paragrafen 3.2.2.1 t/m 3.2.2.4 worden alle modellen kort beschreven die maar één extra variabele toevoegen aan het basismodel dat in paragraaf 3.2.1 staat beschreven. Op basis van deze analyse wordt het meest optimale model in termen van extra toegevoegde deviantie geselecteerd als 'basismodel 2'. In de paragrafen 3.2.3.x wordt na afzonderlijk toevoegen van de overige variabele aan 'basismodel 2' het meest optimale 'basismodel 3' bepaald. In de paragrafen 3.2.4.x wordt op basis van 'basismodel 3' na afzonderlijk toevoegen van de twee overgebleven variabelen 'basismodel 4' bepaald. En tenslotte staat in paragraaf 3.2.5 het complete model beschreven. In hoofdstuk 5 wordt een methode beschreven die gebruikt wordt om van alle beschreven modelvarianten per soort het optimale model te selecteren.

3.2.1 Het basismodel

Dit model heeft dezelfde functie als het MOVE 2-model: de trofiegraad (stikstof(n)-getal), de bodemvochtigheid (vocht(f)-getal) en de zuurgraad (zuur(r)-getal).
In het oorspronkelijke complete model zijn van de 914 soorten er 86 bimodaal[6] (§3.3). Dit betekent dus dat 86 soorten een minimum of twee optima (zadelrug) zullen hebben. In

[6] Bimodale modellen zijn modellen waarvan de coëfficiëntendeterminant ***D*** (Kaldewaij & Van Tiel, 1986) van het stelsel van afgeleide vergelijkingen positief is, of anders gezegd waarvan de regressievergelijking een minimum heeft. Wanneer één van de kwadratische termen een negatieve coëfficiënt heeft, dan is het zeer waarschijnlijk dat deze vergelijking een minimum heeft.

paragraaf 3.3 staat voor alle modellen uitgelegd hoe en waarom deze bimodale modellen opnieuw zijn uitgerekend en de bimodaliteit wordt opgeheven.
Dit model heeft de volgende lineaire schatter:

$$y = \alpha + c\_r * E\_r + c\_r^2 * (E\_r * E\_r) + c\_n * E\_n + c\_n^2 * (E\_n * E\_n) + c\_f * E\_f + c\_f^2 * (E\_f * E\_f) + c\_rn * (E\_r * E\_n) + c\_rf * (E\_r * E\_f) + c\_nf * (E\_n * E\_f) \quad (3.12)$$

Waarbij: α de intercept is
c_x de coëfficiënt is voor Ellenberg waarde x
c_x^2 de kwadratische coëfficiënt is voor Ellenberg waarde x
c_xy is de interactieterm coëfficiënt voor Ellenberg waarden x en y
E_x is de Ellenberg waarde voor x

Bij de bimodale modellen zullen één of meerdere kwadratische coëfficiënten (c_x^2) ontbreken of gelijk aan 0 gesteld zijn.

Het MOVE 2-model is opnieuw uitgerekend omdat in het oorspronkelijke model van een andere dataset is uitgegaan. Om toch de nieuwe modellen met het basismodel te kunnen vergelijken wordt opnieuw dit model afgeleid. Het aantal vrijheidsgraden van dit model is maximaal 9 en minimaal 7 (maar één kwadratische term). De minimale, maximale en gemiddelde verklaarde deviantie zijn respectievelijk gelijk aan 5.48, 77.63 en 31.27. De mediaan van de verklaarde deviantie is gelijk aan 29.11,

Tabel 3.2 Verklaarde percentage deviantie voor model a

	gem.	std dev.	std fout	Mediaan	min	max	sample variantie
model a	31.27	13.06	0.43	29.11	5.48	77.63	170.62

3.2.2 Modellen afgeleid van het basismodel

In deze paragraaf staan alle modellen beschreven die maar op één variabele verschillen van het basismodel (§3.2.1). In de paragrafen 3.2.2.1 t/m 3.2.2.4 staan de modellen waarbij respectievelijk de variabelen ‘cpaf2’, ‘veg’, ‘fgr’ en ‘s’ aan het basismodel worden toegevoegd.

Tabel 3.3 Verklaarde deviantie voor modellen afgeleid van het basismodel

	gem.	std dev.	std fout	Mediaan	min	max	sample variantie
model b	31.79	13.05	0.43	29.79	6.78	77.50	170.36
model c	34.86	11.98	0.40	33.40	7.71	77.66	143.43
model d	37.71	13.45	0.44	36.03	9.76	85.70	180.86
model e	34.28	13.05	0.43	32.61	8.04	79.08	170.35

3.2.2.1 Model b

In model b wordt naast de variabelen uit ‘model a’ ook nog de toxiciteit indicatiewaarde voor zware metalen (cpaf2) meegenomen. Omdat er een aantal ‘NoData’ waarden voor cpaf2 in de dataset zitten neemt het aantal waarnemingen af van 109065 naar 95529. Het aantal vrijheidsgraden van dit model is maximaal 10 en minimaal 8, er is ten opzichte van ‘model a’ één extra verklarende variabele toegevoegd. Het aantal bimodale modellen bedraagt 70, bij het verwijderen van de bimodalen bij dit model is hetzelfde sub-model schema gebruikt als

bij het basismodel (§3.3), omdat ook dit model alleen maar kwadratische termen heeft voor het zuurgetal, stikstofgetal en het vochtgetal.

De gemiddelde verklaarde deviantie (%) (tabel 3.3) is door het toevoegen van deze extra variabele toegenomen ten opzichte van ‘model a’ met 0.52 tot 31.79. De minimum verklaarde deviantie is ook toegenomen tot 6.78, maar heel opvallend is dat de maximum verklaarde deviantie is afgenomen tot 77.50. Voor 346 soorten heeft het toevoegen van de variabele cpaf2 een negatief effect in de verklaarde deviantie, dus voor deze soorten is het toevoegen van deze variabele geen verbetering van het model. De maximale afname van de verklaarde deviantie is 26.7%, de maximale toename is 21.5%.

3.2.2.2 Model c

In model c wordt als extra variabele het vegetatietype toegevoegd. Dit is een zogenaamde factor variabele[7]. Dit is een nominale variabele die vijf verschillende waarden kan hebben (loofbos: ‘DEC’, grasland: ‘GRP’, heide: ‘HEA’, dennenbos: ‘PIN’ en sparrenbos: ‘SPR’). Door het toevoegen van deze ene variabele neemt het aantal vrijheidsgraden niet met één maar met vier (aantal klassen min één) toe. Het aantal vrijheidsgraden van dit model ligt tussen de 11 en 13. Het totaal aantal bimodale modellen bedraagt 91. Bij het vervangen van de kwadratische termen in de functie door lineaire termen worden alle bimodale modellen weer unimodaal.

De gemiddelde verklaarde deviantie (tabel 3.3) is door het toevoegen van deze extra variabele(n) toegenomen tot 34.86 (meer dan 3% extra), zowel het minimum als het maximum is toegenomen. In tegenstelling tot ‘model b’ neemt nu maar voor één soort de verklaarde deviantie af na toevoegen van deze variabele. De afname is daarbij ook maar zeer gering, maar 0.005.

3.2.2.3 Model d

In dit model wordt als extra variabele ten opzichte van ‘model a’ de factor variabele[7] fysisch geografische regio’s toegevoegd. Deze variabele heeft 11 verschillende klassen (zie §2.2) waardoor het aantal vrijheidsgraden na het verwijderen van de 88 bimodale modellen tussen de 17 en 19 ligt.

De gemiddelde verklaarde deviantie (tabel 3.3) is door het toevoegen van deze extra variabele(n) toegenomen tot 37.71, een toename van 6.49%. Ook nu is zowel de minimale als de maximale verklaarde deviantie sterk toegenomen. Na het toevoegen van deze variabele zijn voor alle modellen de modellen verbeterd, de minimale verbetering is 0.25% en de maximale toename is meer dan 41%. Dus het toevoegen van deze ‘regionale’ variabele voegt een zeer belangrijke hoeveelheid verklaarde deviantie toe aan het basismodel.

3.2.2.4 Model e

In dit model wordt als extra verklarende variabele het zoutgetal (Ellenberg-s) meegenomen. Deze variabele heeft een waarde tussen 0 (zout-intolerant) en 9 (sterk zoutbehoevend). Het aantal vrijheidsgraden van dit model na het vervangen van de 68 bimodale modellen ligt tussen de 8 en 10.

[7] Een factor variabele binnen een regressiemodel is een variabele die een ‘eindig’ nominaal bereik heeft. Dit betekent dus dat een factor variabele in een eindig aantal klassen is op te delen.

Door het toevoegen van het zoutgetal is de gemiddelde verklaarde deviantie (tabel 3.3) ten opzichte van 'model a' toegenomen met 3% tot 34.28%. Dit geldt zowel voor de minimale en maximale verklaarde deviantie. In totaal krijgen twee modellen een zeer kleine, verwaarloosbare afname in de hoeveelheid verklaarde deviantie, de grootste afname is slechts 0.00005. De overige modellen hebben een hogere verklaarde deviantie waarbij de maximale toename gelijk is aan 26.9%.

3.2.3 Modellen afgeleid van 'Basismodel 2'

Uit de analyses in paragraaf 3.2.2 blijkt dat na het toevoegen van de variabele die de ruimtelijke verdeling (de fysisch geografische regio's) beschrijft de meeste verklaarde deviantie toegevoegd wordt. Deze variabele wordt dan ook beschouwd als de belangrijkste variabele en dient in het vervolg als de basis waarop extra verklarende variabelen toegevoegd gaan worden. Dit model wordt 'basismodel 2' genoemd. In de paragrafen 3.2.3.1 t/m 3.2.3.3 worden achtereenvolgens de variabelen 'cpaf2', 'veg' en 's' toegevoegd.

Tabel 3.4 Verklaarde deviantie voor modellen afgeleid van 'basismodel 2'

	gem.	std dev.	std fout	mediaan	min	max	sample variantie
model f	37.73	13.50	0.45	36.09	8.96	85.13	182.13
model g	40.28	12.71	0.42	38.90	11.31	85.73	161.49
model h	39.59	13.61	0.45	38.19	11.18	85.72	185.22

3.2.3.1 Model f

In dit model wordt als extra variabele aan 'basismodel 2' de toxiciteit indicatiewaarde voor zware metalen (cpaf2) meegenomen. Het aantal vrijheidsgraden na het vervangen van de 75 bimodale modellen ligt tussen de 18 en 20.

Door het toevoegen van de toxiciteitwaarde voor zware metalen aan het 'basismodel 2' is de gemiddelde verklaarde deviantie (tabel 3.4) toegenomen met slechts 0.02% tot 37.73%. Voor 437 modellen neemt door het toevoegen van de extra variabele cpaf2 de verklaarde deviantie af; de maximale afname is 32.9%. Voor de rest van de soorten levert cpaf2 een positieve bijdrage aan de verklaarde deviantie, het maximum hierbij is 7.2%. Hierbij moet opgemerkt worden dat het totaal aantal waarnemingen bij dit model minder is dan bij 'basismodel 2'. Er zijn 95529 waarnemingen gebruikt tegenover 109065 in basismodel 2, een verschil van 13536 waarnemingen.

3.2.3.2 Model g

In dit model wordt als extra (factor)variabele aan 'basismodel 2' het vegetatietype (veg) meegenomen. Het aantal vrijheidsgraden van dit model ligt tussen de 21 en 23 na het vervangen van de 83 bimodale modellen.

Door het toevoegen van het vegetatietype aan het 'basismodel 2' is de gemiddelde verklaarde deviantie (tabel 3.4) toegenomen met 2.6% tot 40.3%. Bij twee modellen neemt de verklaarde deviantie af, de maximale afname is daarentegen zeer klein, namelijk maar 0.005%. De maximale bijdrage aan de verklaarde deviantie is 20%.

3.2.3.3 Model h

In dit model wordt als extra variabele aan ‘basismodel 2’ het zoutgetal (s) meegenomen. Na het vervangen van de 73 bimodale modellen ligt het aantal vrijheidsgraden tussen de 18 en 20.

Het toevoegen van het zoutgetal aan ‘basismodel 2’ heeft de gemiddelde deviantie (tabel 3.4) verhoogd met 1.88% tot gemiddeld 39.59%. Bij alle modellen neemt de verklaarde deviantie toe, de minimale toename is 0.0001 en de maximale toename is 14.1%.

3.2.4 Modellen afgeleid van ‘Basismodel 3’

Uit de analyses in paragraaf 3.2.3 blijkt dat het toevoegen van de variabele ‘vegetatietype’ de meeste hoeveelheid extra verklaarde deviantie toevoegd aan ‘basismodel 2’. Deze variabele wordt dan ook als de belangrijkste extra variabele beschouwd in de uitbreiding van ‘basismodel 2’. ‘Basismodel 3’ bestaat nu uit de variabelen n, r, f, fgr en veg. In deze paragraaf wordt het effect op het model van het toevoegen van de variabelen voor de toxiciteit als gevolg van zware metalen (cpaf2) en de variabele voor het zoutgetal bekeken.

Tabel 3.5 Verklaarde deviantie voor modellen afgeleid van ‘basismodel 3’

	gem.	std dev.	std fout	mediaan	min	max	sample variantie
model i	40.19	12.72	0.42	38.71	11.41	85.15	161.77
model j	41.38	13.03	0.43	40.19	11.59	85.77	169.90

3.2.4.1 Model i

In dit model wordt als extra variabele aan ‘basismodel 3’ de toxiciteitwaarde zware metalen (cpaf2) toegevoegd. Na het vervangen van de 71 bimodale modellen ligt het aantal vrijheidsgraden tussen de 22 en 24.

Na het toevoegen van de variabele cpaf2 is de gemiddelde verklaarde deviantie (tabel 3.5) afgenomen met 0.09%. Na het toevoegen van deze extra variabelen neemt voor 463 modellen de verklaarde deviantie af, de maximale afname bedraagt 32.8%. De maximale toename van de overige modellen is 7.2%. Bij dit model moet ook weer opgemerkt worden dat door het toevoegen van cpaf2 het totaal aantal waarnemingen is afgenomen met 13536, waardoor deze afname misschien toch in een ander licht gezien moet worden.

3.2.4.2 Model j

Bij dit model wordt ten opzichte van ‘basismodel 3’ de variabele zoutgetal (s) toegevoegd. Na het vervangen van de 79 bimodal modellen ligt het aantal vrijheidsgraden tussen de 22 en 24.

Door het toevoegen van het zoutgetal (s) aan ‘basismodel 3’ is de gemiddelde verklaarde deviantie (tabel 3.5) met 1.10% toegenomen. De maximale toegevoegde verklaarde deviantie is 13.0% en de minimale toevoeging is 0%.

3.2.5 Het volledige model

Uit de analyse van paragraaf 3.2.4 kan geconcludeerd kunnen worden dat het toevoegen van het zoutgetal aan ‘basismodel 3’ belangrijker is dan het toevoegen van de toxiciteitwaarde van zware metalen. Om de analyse af te maken is tenslotte het volledige model doorgerekend. In dit model zitten alle variabelen die op voorhand van belang geacht werden.

Ten opzichte van het tot nu toe meest verklarende model (model j) voegt dit model wanneer alleen gekeken wordt naar de gemiddelde verklaarde deviantie (tabel 3.6) niks toe. De gemiddelde verklaarde deviantie neemt met 0.09% af ten opzichte van model j. Voor 472 modellen neemt de deviantie af en voor 442 modellen (soorten) is dit model beter dan model j. De maximale afname van de verklaarde deviantie is 32.9%, terwijl de maximale toename gelijk is aan 8.1%.

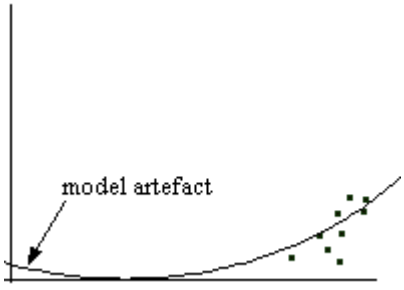
Tabel 3.6 Verklaarde deviantie voor model z

	gem.	std fout	mediaan	std dev.	sample variantie	min	max
model z	41.29	0.43	40.12	13.03	169.80	11.52	85.20

In hoofdstuk 4 wordt met behulp van stapsgewijze regressie (vanuit het basismodel) bepaald wat het beste stapsgewijze model is. In hoofdstuk 5 wordt beschreven hoe uit bovenstaande modellen en het optimale stapsgewijze regressiemodel het meest optimale model geselecteerd wordt.

3.3 Bimodale modellen

Bimodale modellen zijn modellen met twee toppen. In onze modellen betekend het hebben van twee toppen dat er een minimum aanwezig zal zijn. Bimodale modellen treden bijvoorbeeld op wanneer lineair gedrag door middel van een kwadratische term gemodelleerd gaat worden (zie figuur 3.2). Het voorkomen van een verschijnsel (punten) aan de rand van het modelgebied wordt in figuur 3.2 door middel van een parabolische functie gemodelleerd. Het rechterbeen van deze parabool beschrijft het gedrag goed, maar het linkerbeen is een modelartefact. Beter zou het zijn om dit gedrag door middel van een lineair model te beschrijven. Met de door ons gebruikte regressievergelijkingen zijn ‘echte’ bimodale modellen, dus modellen met twee toppen, niet goed te modelleren. Om deze wel goed te modelleren moet er minimaal een derde macht in de regressiefunctie staan.



Figuur 3.2 Bimodaal modelartefact

Of een model bimodaal is kan kan bepaald worden door het berekenen van de coëfficiëntendeterminant ***D*** (Kaldewaij en Van Tiel, 1986) van het stelsel van afgeleide vergelijkingen. Modellen met een positieve ***D***, of anders gezegd modellen met een regressievergelijking met een minimum zijn hoogstwaarschijnlijk bimodaal. De bimodaliteit is waarschijnlijk een model artefact. In het uitgevoerde onderzoek zijn dié modellen als bimodaal aangeduid wanneer de kwadratische termen een positieve coëfficiënt hebben.

De gevolgde procedure voor het verbeteren van het model door het verwijderen van één of meer van de kwadratische termen uit de regressievergelijking wordt in deze paragraaf beschreven.

Omdat er maar drie variabelen (n, r en f) een kwadratische term hebben en er geen model doorgerekend wordt zonder kwadratische termen moet er voor ieder gevonden bimodaal model zes submodellen doorgerekend worden. In totaal zijn er 8 (2^3) modelvarianten die doorgerekend zouden kunnen worden, maar er vallen er twee af: het model met 3 kwadratische termen is al uitgerekend en het model helemaal zonder kwadratische termen wordt niet uitgerekend. In tabel 3.7 staan voor alle submodellen aangegeven wat de variaties zijn in de oorspronkelijke modellen. Tabel 3.8 laat per model, per submodel het aantal maal zien dat door het doorrekenen van een submodel de bimodaliteit van een model opgelost werd en hoe vaak een bepaald submodel de hoogste verklaarde deviantie had en dus 'optimaal' was. De modellen die tot een oplossing zijn gekomen en waarvan het percentage verklaarde deviantie het hoogst was ('optimale modellen') zijn uiteindelijk in de berekeningen in paragraaf 3.2 meegenomen.

Tabel 3.7 Meegenomen variabelen bij de submodellen voor het elimineren van bimodale modellen.

	kwadratische termen		
	r-kwadraat	n-kwadraat	f-kwadraat
submodel 1		x	x
submodel 2	x		x
submodel 3	x	x	
submodel 4	x		
submodel 5		x	
submodel 6			x

Tabel 3.8 Overzicht van de uitgerekende sub-modellen, waarin per sub-model het aantal niet bimodale oplossingen staat aangegeven en hoeveel er optimaal zijn (hoogste verklaarde deviantie)

		sub-model						
		1	2	3	4	5	6	aantal
model a	# oplossingen	36	29	41	49	61	64	
	# optimaal	29	23	27	1	2	4	86
model b	# oplossingen	21	22	32	44	47	48	
	# optimaal	12	19	24	1	1	13	70
model c	# oplossingen	44	27	48	49	73	66	
	# optimaal	37	21	26	2	2	3	91
model d	# oplossingen	38	34	48	55	67	65	
	# optimaal	28	21	31	1	3	4	88
model e	# oplossingen	34	19	39	36	56	51	
	# optimaal	29	13	22	1	2	1	68
model f	# oplossingen	41	27	47	52	64	57	
	# optimaal	31	18	29	1	2	2	83
model g	# oplossingen	38	23	39	38	65	59	
	# optimaal	36	15	24	0	2	2	79
model h	# oplossingen	25	21	39	41	59	45	
	# optimaal	24	14	27	2	2	2	71
model z	# oplossingen	29	18	31	36	57	47	
	# optimaal	29	11	24	1	2	1	68

4. Stapsgewijze regressiemodellen

In hoofdstuk 3 zijn 11 verschillende modellen onderzocht. Omdat het MOVE 2-model als uitgangspunt wordt genomen en er alleen gevarieerd mocht worden in de overige vier toegevoegde variabelen was er maar een keuze tussen 16 verschillende modellen mogelijk waarvan er 11 zijn uitgerekend. Wanneer deze restrictie niet geldt dan loopt het aantal mogelijke combinaties snel op. Als alle variabelen vrijelijk gekozen kunnen worden zijn er in totaal 8192 (2^{13}) combinaties mogelijk. Deze kunnen niet meer met de hand één voor één uitgerekend worden, hiervoor zijn geautomatiseerde procedures ontwikkeld, de stapsgewijze regressies. Binnen de geautomatiseerde stapsgewijze regressie procedures wordt er gebruik gemaakt van criteria (zoals het informatiecriterium AIC (§4.1)) die bepalen of het toevoegen of verwijderen van een variabele bijdraagt aan het verbeteren van het model. Niet alle combinaties worden in zo'n stapsgewijze procedure doorgerekend, afhankelijk van het bereiken van het stopcriterium wordt besloten of het zinvol is om verder te zoeken of dat de 'optimale' verzameling variabelen bereikt is.
Er bestaan verschillende criteria op basis waarvan besloten wordt wat het meest optimale model is, deze staan beschreven in paragraaf 4.1. De bekendste en meest gebruikte zijn de likelihood ratio (§4.1.1) en de informatiecriteria (§4.1.2): AIC en BIC. In deze studie zijn beide informatiecriteria gebruikt. Binnen een stapsgewijze regressie analyse zijn naast de keuze van het selectiecriterium nog vele variaties mogelijk. Zo is het mogelijk om te variëren in het start- en eindpunt van de analyse. Dit betekent dat er gekozen kan worden tussen een topdown, bottom-up benadering en iedere mogelijke variatie daartussen. Een topdown benadering levert als nulmodel de complete verzameling variabelen en een bottom-up benadering levert als nulmodel het lege model (geen variabelen) en geeft als bovengrens van de analyse de complete verzameling van variabelen. Alle mogelijke variaties op de bovenstaande benaderingen zijn dus ook mogelijk.
In dit onderzoek is besloten om 6 variaties door te rekenen. Er wordt gevarieerd in het selectiecriterium (AIC en BIC, zie §4.1.2) en er wordt gevarieerd in de benadering topdown, bottom-up en een optie daartussen, het MOVE 2-model. Resultaten van deze stapsgewijze procedures staan in de paragrafen 4.1 en 4.2.

4.1 De selectiecriteria

Er zijn verschillende methode binnen de stapsgewijze regressie analyse waarmee het meest optimale logistische model geselecteerd kan worden (Bio, 2000). De twee meest gebruikte aanpakken zijn de 'likelihood ratio' en de informatiecriteria aanpak.

4.1.1 De waarschijnlijkheids 'likelihood' ratio aanpak

Wanneer wordt uitgegaan van een serie logistische regressie modellen met een verzameling geneste schatters, dan is het mogelijk om de waarschijnlijkheid 'likelihood' ratio test (LRT) te gebruiken om de verschillende modellen onderling te vergelijken. De vergelijking gebeurt op basis van de geschiktheid of significantie van de extra parameter(s). Deze vergelijking (ΔD) is gebaseerd op de chi-kwadraat verdeling van het verschil in deviance D (afwijkend gedrag) (Hosmer en Lemeshow, 1989):

$$\Delta D = D(\text{voor minder complex model}) - D(\text{voor complexer model}) \qquad (4.1)$$

ΔD kan uitgedrukt worden als een waarschijnlijkheidverhouding:

$$\Delta D = -2\ln\left[\frac{waarschijnlijkheid\ zonder\ extra\ term}{waarschijnlijkheid\ met\ extra\ term}\right] \qquad (4.2)$$

ΔD benaderd een χ^2 verdeling met een aantal vrijheidsgraden die gelijk is aan het verschil in vrijheidsgraden tussen de twee modellen (McCullagh & Nelder, 1989, Hosmer en Lemeshow, 1989). De LRT wordt uitgerekend onder de nul hypothese dat de additionele term(en) in het complexere model niet significant is (of zijn) bij een bepaalde α. Wanneer ΔD groter is dan de respectievelijke χ^2 wordt de nul hypothese verworpen en wordt de extra term dus geaccepteerd.
De keuze van de α waarde waarbij een term opgenomen moet worden hangt af van het doel van de modellering; dat wil zeggen of een fout van type I (afwijzing van een significante model term) of een fout van type II (acceptatie van een niet-significante model term) het meest kritisch is. In de literatuur worden niveaus tussen de 0.05 en 0.25 (b.v. Bendel en Afifi, 1977; Costanza en Afifi, 1979; White en Bennets, 1996) gevonden, maar ook extreme niveaus, zoals 0.9 die toegepast is in een medische studie door Kay en Little (1986) waarbij zij zich richten op het meenemen van alle termen die maar enigszins de moeite lonen.

4.1.2 De informatiecriteria aanpak

Er zijn twee veelgebruikte informatiecriteria methoden, Akaike's en Bayesian (Bio, 2000). Akaike's informatie criteria (AIC) selecteert het model dat een minimale verliesfunctie oplevert. De verliesfunctie bestaat uit een maat voor de schattingsfout en een strafwaarde voor toegenomen modelcomplexiteit (Akaike, 1973 en 1974). Deze maat wordt als volgt uitgerekend:

$$\text{AIC} = \text{-2 ln(gemaximaliseerde waarschijnlijkheid)} + 2 \cdot \text{(aantal parameters)} \qquad (4.3)$$

De AIC was oorspronkelijk ontwikkeld voor modellen die door middel van 'least squares' gefit zijn. De AIC is een speciaal geval van Mallows' C_p (Mallows, 1973) waar de gemiddelde kwadratische fout van de gefitte modellen gebruikt wordt om σ^2 te schatten. De AIC schat de gemiddelde kwadratische fout van een modelschatting voor toekomstige observaties als een eenvoudige functie van de restsom van de kwadraten wanneer een lineaire 'least squares' procedure gebruikt wordt om het model te fitten. Wanneer een andere methode gebruikt is om een model te fitten dan de 'least squares', dan moet een maat voor de schattingsfout gebruikt worden die bij de toegepaste methode past (bv. Burman en Nolan, 1995). Modellen zoals de logistische modellen die met behulp van maximale waarschijnlijkheid gefit worden kunnen eenvoudig onderling vergeleken worden. Voor de logistische regressie modellen wordt de AIC als volgt geschat (Webster en McBratney, 1989):

$$\hat{AIC} = \left[n\ln\left(\frac{2\pi}{n}\right) + n + 2\right] + D + 2p \qquad (4.4)$$

Modellen worden geselecteerd die een minimale verlies functie hebben. De verliesfunctie bestaat uit een maat voor de schattingsfout en een boeteterm voor de toegenomen modelcomplexiteit. Het model met de laagste geschatte AIC wordt gekozen. Omdat het getal tussen vierkante haken gelijk is voor één en dezelfde dataset, is het mogelijk om modellen die alleen verschillen in het aantal parameters (p) te vergelijken door alleen $D + 2p$ uit te rekenen.

Akaike (1978) en Schwarz (1978) hebben met behulp van een Bayesiaanse benadering het selectie vraagstuk benaderd. Het Bayesiaanse Informatie Criterium (BIC) is ook wel bekend als het Schwarz' Information Criterion (SIC). Analoog aan vergelijking 4.3 wordt de BIC gedefinieerd door middel van:

BIC = -2 ln(gemaximaliseerde waarschijnlijkheid) + (aantal parameters) * ln(n)

$$BIC = \left[n \ln\left(\frac{2\pi}{n}\right) + n + 2 \right] + D + p \ln(n) \qquad (4.5)$$

Logistische modellen die alleen verschillen in het aantal parameters p kunnen direct vergeleken worden door $D + p \ln(n)$ uit te rekenen. Dit criterium selecteert dat model waarin de verliesfunctie is geminimaliseerd. De verliesfunctie bestaat uit een maat voor de schattingsfout – hier de restterm van de model deviantie (D) – en een boeteterm voor de toegenomen model complexiteit en monstergrootte ($p \cdot ln(n)$).
Er is een verschil in de filosofie tussen AIC en BIC. AIC neemt aan dat het 'ware model' bestaat uit een hoge (eventueel zelfs een oneindige) dimensionaliteit. Uitgaande van de data wordt het best mogelijke model geschat. Daarom zal de dimensie van het gekozen model laag zijn bij weinig data en toenemen wanneer er meer informatie beschikbaar komt. BIC neemt aan dat er een laag-dimensionaal 'true model' bestaat, dit wordt op consequente wijze door het criterium geschat (Shibata, 1989; Buckland *et al.*, 1997). Numeriek is het verschil in BIC voor twee modellen gelijk aan het verschil in AIC vermenigvuldigd met ½ ln(n), waardoor de BIC (wanneer $n \geq 8$) die modellen selecteert waarin minder parameters zitten.

4.2 Modellen met stopcriterium 'AIC'

Het informatiecriterium AIC is voor drie verschillende modelvarianten gebruikt als selectiecriterium voor het afleiden van het meest optimale model. De drie doorgerekende varianten verschillen alleen in het startpunt van waaruit de regressie analyse uitgevoerd wordt. Allereerste is de variant doorgerekend met als startmodel het lege model en is de ruimte waarbinnen gezocht kan worden begrensd tussen geen variabelen (ondergrens) en alle mogelijke variabelen (het complete model inclusief de kwadratische termen: 'model z' zie §3.2 en §3.2.5). In de tweede variant is het startmodel het complete model en tevens de bovengrens; de ondergrens is niet vastgelegd en dus het nulmodel. De derde en laatste variant heeft als startmodel het MOVE 2-model (zie §3.2.1) en is begrensd door het nulmodel (geen variabelen) en het complete model 'model z' (zie §3.2 en §3.2.5). Resultaten van deze analyse staan in tabel 4.1

Tabel 4.1a Overzicht van de resultaten van de stapsgewijze regressie modellen met als informatiecriterium AIC

model		df	verklaarde deviantie	Hosmer-Lemeshow test[8]
AIC leeg	minimum	5	10.94	1.63
	maximum	25	85.08	761.61
	gemiddelde	21.04	40.48	29.77
	# significant			406
	% significant			44.42%
AIC compleet	minimum	7	11.20	1.45
	maximum	25	85.13	504.60
	gemiddelde	22.93	41.21	28.38
	# significant			420
	% significant			45.95%
AIC move 2	minimum	6	11.06	2.39
	maximum	25	85.13	504.60
	gemiddelde	22.09	41.06	28.45
	# significant			426
	% significant			46.61%

Tabel 4.1b Overzicht van de resultaten van de stapsgewijze regressie modellen met als informatiecriterium AIC

model		omissies													
		r	r^2	n	n^2	f	f^2	s	cpaf2	fgr	veg	rn	rf	nf	bi-modaal
AIC leeg	#	166	128	133	106	44	83	178	478	22	209	381	442	423	105
	%	18.16	14.00	14.55	11.60	4.81	9.08	19.47	52.30	2.41	22.87	41.68	48.36	46.28	11.49
AIC compleet	#	4	52	5	35	9	32	83	287	23	142	125	239	224	42
	%	0.44	5.69	0.55	3.83	0.98	3.50	9.08	31.40	2.52	15.54	13.68	26.15	24.51	4.60
AIC move 2	#	4	55	6	37	9	32	176	480	34	220	138	256	243	44
	%	0.44	6.02	0.66	4.05	0.98	3.50	19.26	52.52	3.72	24.07	15.10	28.01	26.59	4.81

Wat opvalt in bovenstaande tabellen is de invloed van het startmodel dat als eerste geanalyseerd wordt in de stapsgewijze regressieprocedure en waarmee de overige varianten vergeleken worden. Wanneer als startmodel het complete model of het MOVE 2-model wordt aangeboden dan blijft het uiteindelijke optimale model in de buurt van het startmodel liggen (gemiddeld aantal vrijheidsgraden groter dan 22 met een maximum aantal vrijheidsgraden van 25). Bijvoorbeeld wanneer het complete model als startmodel wordt aangeboden blijven vaak alle aangeboden variabelen in het uiteindelijke optimale model (tabel 4.1b). Alleen de variabelen *cpaf2*, *veg* en de interactietermen *rn*, *rf* en *nf* worden nog redelijk vaak uit het optimale model weggelaten. Ditzelfde geldt bij het aanbieden van het MOVE 2-model als startpunt, ook nu blijft het oorspronkelijke model (bestaande uit *r*, r^2, *n*, n^2, *f* en f^2) intact en zijn de interactietermen (*rn*, *rf* en *nf*) vaak verwijderd. De variabele *fgr* is erg belangrijk, in 96.3% van de modellen aanwezig, en met name de variabele *cpaf2* wordt in nog geen 50% van de modellen toegevoegd en is blijkbaar de minst belangrijke variabele. Dit komt overeen met de resultaten uit hoofdstuk 3.2, waarin als laatste variabelen de toxiciteitindicatiewaarde (*cpaf2*) wordt toegevoegd aan het complete model (§3.2.5).
De drie modellen verschillen niet noemenswaardig in het percentage verklaarde deviantie en de Hosmer-Lemeshow goodness of fit testwaarde, het percentage significante modellen volgens deze test schommelt rond 45 procent.

[8] De gebruikte Hosmer-Lemeshow goodness-of-fit test (Hosmer en Lemeshow, 1989; Bio, 2000) is χ^2 verdeeld met 8 vrijheidsgraden (10 klassen – 2). Als criterium is het 5% betrouwbaarheidsinterval gebruikt en dit betekent dat de Hosmer-Lemeshow-waarde kleiner of gelijk moet zijn aan 15.507 om significant te zijn.

4.3 Modellen met stopcriterium 'BIC'

Net als bij het gebruiken van de AIC als informatiecriterium is bij de BIC ook de invloed van het startmodel erg belangrijk. Het uiteindelijke optimale model lijkt wat de uiteindelijk mee te nemen variabelen erg sterk op het startmodel, zie met name BIC compleet en BIC MOVE 2. Dit geldt niet wanneer er met een leeg model gestart wordt. Het lijkt erop dat het starten met een volledige model in combinatie met BIC geen goede resultaten oplevert. In bijna alle gevallen blijft het aangeboden model (het complete model) onveranderd, het gemiddeld aantal vrijheidsgraden is 24.5 ten opzichte van een maximum van 25 wanneer alle variabelen meegenomen worden. Volgens Bio (2000) moeten er bij het gebruik van het BIC selectiecriterium, modellen afgeleid worden die gemiddeld minder variabelen (en dus minder vrijheidsgraden) hebben, maar wanneer het volledige model als startmodel gebruikt wordt levert de BIC methode modellen op met gemiddeld meer vrijheidsgraden dan de methode die van AIC gebruik maakt in deze dataset.

Tabel 4.2a Overzicht van de resultaten van de stapsgewijze regressie modellen met als informatiecriterium BIC

model		df	verklaarde deviantie	Hosmer-Lemeshow test[8]
BIC leeg	minimum	0	0.00	0.99
	maximum	25	84.64	777.59
	gemiddelde	16.08	38.09	33.06
	# significant			344
	% significant			37.64%
BIC compleet	minimum	11	11.52	2.58
	maximum	25	85.21	497.32
	gemiddelde	24.48	41.21	28.36
	# significant			432
	% significant			47.26%
BIC move 2	minimum	8	7.81	1.89
	maximum	25	85.14	522.85
	gemiddelde	18.59	39.55	29.40
	# significant			415
	% significant			45.40%

Tabel 4.2b Overzicht van de resultaten van de stapsgewijze regressie modellen met als informatiecriterium BIC

model		omissies													
		r	r^2	n	n^2	f	f^2	s	cpaf2	fgr	veg	rn	rf	nf	bi-modaal
BIC leeg	#	331	252	271	224	120	197	374	745	181	455	631	692	688	159
	%	36.21	27.57	29.65	24.51	13.13	21.55	40.92	81.51	19.80	49.78	69.04	75.71	75.27	17.40
BIC compleet	#	0	5	0	2	0	0	3	16	14	69	11	9	12	67
	%	0.00	0.55	0.00	0.22	0.00	0.00	0.33	1.75	1.53	7.55	1.20	0.98	1.31	7.33
BIC move 2	#	0	1	0	0	0	0	367	744	278	484	12	8	13	67
	%	0.00	0.11	0.00	0.00	0.00	0.00	40.15	81.40	30.42	52.95	1.31	0.88	1.42	7.33

Wanneer met een leeg model gestart wordt, wordt er voor drie soorten, *Berteroa incana* (grijskruid), *Epilobium obscurum* (donkergroene basterdwederik) en *Sagina apetala* (tengere vetmuur) geen model gevonden, terwijl met MOVE 2 als startmodel er voor deze soorten wel significante modellen gevonden worden met Hosmer-Lemeshow testwaarden van respectievelijk 10.7, 4.9 en 7.2 en een verklaarde deviantie van respectievelijk 41%, 12% en 14%. Dus ook het starten met een leeg model levert niet altijd goede modellen op. Wat opvalt is dat de BIC methode in vergelijking met de AIC methode een grotere variatie in optimale modellen oplevert, vergelijk het aantal vrijheidsgraden (df) in tabel 4.1a met 4.2a.

Bij drie van de zes uitgerekende stapsgewijze regressiemethoden komt als beste model het volledige model, model z, uit de analyse (zie tabel 4.1a en 4.2a). Omdat model z al afgeleid is besloten om met de overige drie resultaten uit de stapsgewijze analyse verder te gaan. Dit zijn de resultaten uit AIC compleet, BIC leeg en BIC MOVE 2. In hoofdstuk 5 wordt beschreven hoe deze resultaten en de resultaten uit de handmatige analyse (hoofdstuk 3) gebruikt zijn om per soort het meest optimale model te selecteren.

5. Selecteren van het 'optimale' model

5.1 Goodness of fit maten

Goodness of fit maten (Bio, 2000) geven een beschrijving, indicatie hoe goed een model het te modelleren gedrag benaderd, dit wil zeggen hoe goed de overeenkomst is tussen waarnemingen en modelvoorspellingen. Goodness of fit wordt ook wel aanpassingsgraad genoemd. Tijdens deze studie zijn er twee verschillende maten bekeken: de Hosmer-Lemeshow test (Hosmer en Lemeshow, 1989) en het percentage verklaarde deviantie (*%D*). De Hosmer-Lemeshow test vergelijkt groepen waargenomen en voorspelde responsies met elkaar, waarbij de geschatte waarden van ieder model oplopend geordend zijn voordat deze in één van de x aantal klassen ingedeeld werd. De som van de geschatte waarden per klasse is voor iedere afzonderlijke klasse gelijk. In Hosmer *et al.* (1988) worden deze en enkele andere classificatiestrategieën beschreven. Het Hosmer-Lemeshow goodness of fit criterium ($\hat{C}$) wordt min of meer op dezelfde wijze berekend als de Pearson χ^2 statistiek:

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n_k \overline{\pi})^2}{n_k \overline{\pi}_k (1 - \overline{\pi}_{k)}} = \sum_{k-1}^{g} \frac{(o_k - e_k)^2}{e_k (1 - \overline{\pi}_k)} \qquad (5.1)$$

waar n_k het aantal waarnemingen, o_k het aantal voorkomens is (Occurrences), $\overline{\pi}_k$ de gemiddelde geschatte waarschijnlijkheid in de k-de klasse is en e_k het aantal verwachte of geschatte voorkomens in de k-de klasse is (Hosmer en Lemeshow, 1989). $\hat{C}$ wordt dan vergeleken met een χ^2-verdeling met k - 2 vrijheidsgraden. Het verdelen van de paren waargenomen en geschatte waarden over de klassen kan op verschillende manieren gedaan worden: klassen met een gelijk aantal opnamen, klassen van gelijke breedte, klassen waarin de som van de geschatte waarden gelijk is, enzovoort. Binnen dit onderzoek is gebruik gemaakt van tien klassen, waarbij elke klasse bestaat uit een frequentie van geschatte voorkomens die gelijk is aan eentiende van de totale som van geschatte voorkomens (zie §6.3 in Bio, 2000). Bij deze methode ligt de nadruk op de geschatte voorkomens, terwijl de fit op de geschatte afwezigheid minder goed geschat wordt.
Een andere veel gebruikte maat is het percentage verklaarde deviantie (*%D*). De deviantie hangt af van het aantal opnamen en het aantal positieve waarnemingen daarin. Het percentage verklaarde deviantie is als maat eigenlijk geen goodness of fit, maar wordt hiervoor wel vaak gebruikt (Swartzman en Huang, 1989; Heikkinen, 1996; Van de Rijt *et al.*, 1996). Wanneer het aantal opnamen toeneemt, neemt het percentage verklaarde deviantie af. Er bestaat een sterk verband tussen de samenstelling van de opnamen: opnamen met extreme aantallen 0-en of 1-en hebben hogere waarde voor %*D* dan opnamen met een meer uniforme verdeling. Dus moet er tijdens het evalueren en vergelijken van verschillende modellen met behulp van de %*D* rekening gehouden worden dat een deel van de verklaarde deviantie spontaan optreedt en dat er een verband bestaat tussen de opnamegrootte en de samenstelling van de opnameset. Hoewel de %*D* geen rechtstreekse maat voor het goed fitten van een model is (Yee en Mitchell, 1991), laat deze test het wel toe om verschillende modellen te vergelijken die op dezelfde verzameling gegevens zijn gebaseerd.

5.2 De gebruikte selectiemethode

Het proces van het selecteren van een optimaal model is geen eenduidige vaststaande procedure. Afhankelijk van het doel of gebruik van het model is het mogelijk dat andere selectiecriteria gekozen worden en dat een ander model geselecteerd wordt. Het doel waarvoor de plantenmodellen (zie deel I) afgeleid zijn is om het mogelijk te maken om gevolgen van omgevingsveranderingen op afzonderlijke plantensoorten te modelleren en of te voorspellen. Dus per soort moet minimaal één van de meegenomen variabelen temporeel kunnen variëren. Wanneer een model alleen bestaat uit de variabele *fgr* dan zal dit model geen variatie in de tijd laten zien omdat deze ruimtelijke indeling stationair is. Een andere eis is dat het gekozen model een 'goed' model moet zijn. Er zijn verschillende maten die een indicatie geven van de goodness of fit (zie §5.1), zoals bijvoorbeeld de Hosmer-Lemeshow test. Maar ook bijvoorbeeld de hoeveelheid verklaarde deviantie en de maximale schatter. Tenslotte is het wenselijk om het aantal vrijheidsgraden per model zo laag mogelijk te houden, dus maximaal resultaat met een minimaal aantal variabelen. Dit is omdat het model toegepast zal gaan worden voor schattingen voor heel Nederland en de invoer voor het model afkomstig is van andere modellen, met daarin onzekerheden. Omdat er veel verschillende modellen zijn doorgerekend zijn er verschillende variaties mogelijk. Het selecteren van een optimaal model wordt hierdoor bemoeilijkt. De belangrijkste eis is dat het afgeleide model een goede fit heeft op de oorspronkelijke datawaarden. De eerste voorwaarde waar het model aan moet voldoen is dan dus ook dat de Hosmer-Lemeshow testwaarde, $\hat{C}$, (χ^2-verdeeld) kleiner of gelijk moet zijn aan 15.51 ($\alpha = 0.05$). Dit komt overeen met een kans van 5% dat een model onterecht wordt afgewezen. Van de oorspronkelijke 914 modellen voldoen er 690 aan dit criterium, maar omdat er in totaal 14 modelvarianten per plant zijn afgeleid is het mogelijk dat meerdere varianten aan deze eis voldoen. Dit is voor 622 varianten het geval, voor 75 afgeleide varianten voldoen zelfs alle modellen aan de $\hat{C}$ testwaarden. In die gevallen wanneer meerdere varianten aan de gestelde goodness of fit eis voldoen zijn er nu enkele mogelijkheden om een variant te kiezen:

- selecteren van *die* variant met de laagste $\hat{C}$-waarde of,
- selecteren van *die* variant met de hoogste verklaarde deviantie of,
- selecteren van *die* variant met de hoogste schatter,

De eerste variant levert voor veel modellen een onbevredigend resultaat op. De tweede variant staats haaks op de eis dat het aantal vrijheidsgraden van het model zo laag mogelijk gehouden moet worden; het toevoegen van extra variabelen (toename aantal vrijheidsgraden) leidt altijd tot een toename van de verklaarde deviantie. Dus blijft de laatste variant, maximaliseren van de schatter over.
De toegepaste HosLem eis levert niet voor alle soorten een geschikt model op. Een voorbeeld is de *Orchis morio* (harlekijn, cbs nummer 889), deze wordt wanneer de eis toegepast wordt alleen gemodelleerd door de aanwezige hoeveelheid nutriënten (n en n^2) en door de fysisch geografische regio. Dit terwijl deze soort zeer zeldzaam is en verwacht mag worden dat zeldzame soorten zeer kritisch zijn ten opzichte van standplaatscondities. De hoeveelheid verklaarde deviantie en de maximale schatter van deze soort zijn respectievelijk 28.4% en 0.06. Om deze mindere modelresultaten uit de uiteindelijke selectie te halen is er een tweede criterium gedefinieerd. Van alle modelvarianten die een hogere $\hat{C}$-waarde hebben dan 15.51, is die variant geselecteerd die het dichtst bij de gehanteerde grenswaarde ligt. Vervolgens zijn handmatig de minder geschikte modellen (uit de eerste stap) vergeleken met de selectie uit het tweede criterium, waarbij alleen die modellen zijn meegenomen waarvan de HosLem-

waarde kleiner is dan 20.09 ($\alpha = 0.01$). Het model voor de *Orchis morio* ($\hat{C}$-waarde van 16.19) wordt nu aanzienlijk verbeterd, de variabelen die in het model meegenomen worden zijn nu r, r^2, n, n^2, f, f^2, cpaf2 en fgr. De hoeveelheid verklaarde variantie is nu 52.4 en de maximale schatter is 0.61. Dit levert dus een aanzienlijke verbetering van dit model op met maar een minimaal verlies aan goodness of fit.

Samengevat moet de te gebruiken selectiemethode aan de volgende uitgangspunten voldoen:

- De $\hat{C}$-testwaarde (χ^2-verdeeld) moet kleiner zijn dan 15.51 (kans dat 5% van de modellen onterecht worden afgewezen) en wanneer twee of meer modellen hieraan voldoen wordt het model met de maximale schatter gekozen.
- Uit het eerste criterium kunnen modellen overblijven die een slechte schatter hebben ten opzichte van de maximale schatter uit de verzameling onderzochte modelvarianten. Van deze modellen wordt vervolgens gekeken of er een model bestaat waarvan de $\hat{C}$-waarde het dichtst bij de 5% grenswaarde ligt en bovendien binnen de 1% grenswaarde ligt ($\alpha = 0.01$ => testwaarde kleiner dan 20.09). Dit is een niet geautomatiseerde actie, dit in tegenstelling tot het eerste uitgangspunt.
- Tenslotte moeten optimale modellen gevolgen van veranderingen in de abiotiek in beeld kunnen brengen. Per model moet er minimaal één veranderende variabelen in het model zitten. Wanneer een soort ‘optimaal’ gemodelleerd wordt met alleen de variabele *fgr* dan voldoet dit model niet en moet er een model gekozen worden dat iets minder ‘optimaal’ is dan dit model, maar waarbij abiotische veranderingen wel effect hebben op de verwachte kans van voorkomen.

Het laatste punt betreft een controle achteraf en zal indien het voorkomt dat na toepassen van de eerste twee selectiecriteria een model gekozen wordt met alleen de variabele *fgr* of *veg* toegepast gaan worden.

5.3 Resultaten van de selectiemethode

Toepassen van de criteria uit paragraaf 5.2 leidt ertoe dat voor 690 van de 914 soorten een model wordt afgeleid, zie bijlage II.b (Bakkenes *et al.*, 2002) voor een volledig overzicht van alle 690 vergelijkingen. Voor 15 soorten leverde de goodness of fit maat (uitgangspunt 1) geen biologisch betekenisvol model op. Voor deze soorten is het model geselecteerd dat net buiten de goodness of fit eis lag, maar waarvan $\alpha > 0.01$, en een hogere maximale kans op voorkomen heeft. De derde eis is niet gebruikt; er waren geen modellen met alleen stationaire regressiesvariabelen. Van de 690 soorten die aan de eisen voldoen is de gemiddelde $\hat{C}$-testwaarde gelijk aan 10.77, met een minimum waarde van 0.99 en een maximum waarde van 17.93 met een standaard deviantie van 3.37. In tabel 5.1 staan voor de optimale modellen informatie over het percentage verklaarde deviantie. Gemiddeld wordt circa 39% van de aanwezige deviantie verklaard.

Tabel 5.1 Verklaarde deviantie voor de optimale modellen

gem.	std dev.	std fout	mediaan	min	max	sample variantie
39.26	13.43	0.51	38.37	8.67	80.49	180.43

In de uiteindelijke selectie van 690 modellen zijn alle doorgerekende modelvarianten in min of meer dezelfde mate aanwezig (tabel 5.2). De modellen j en z zijn ten opzichte van de andere modellen sterk vertegenwoordigd, maar niet in die mate dat zij de uiteindelijke modelselectie domineren. Alle modelvarianten komen uiteindelijk in de geoptimaliseerde selectie voor. Dit sluit goed aan bij het doel van dit onderzoek, namelijk het waar mogelijk verbeteren van de MOVE 3 variant van De Heer *et al.* (2000); deze variant is gelijk aan model z. Uit de resultaten blijkt dat modelvariant z, het complete model waarin alle variabelen meegenomen worden maar voor 15% van alle plantensoorten het optimale model is.
In tabel 5.2 staat ook per variabele aangegeven het aantal maal dat deze variabelen geselecteerd is in het optimale model. Uit deze tabel blijkt dat de variabelen zout (s), combipaf (cpaf2) en vegetatietype (veg) vaak niet in het uiteindelijke optimale model meegenomen worden. Voor de variabelen cpaf2 en s is dit niet zo verrassend, omdat deze variabelen maar op enkele plekken hoge waarden hebben en in de rest van Nederland nul of bijna nul zijn, zie bijvoorbeeld de scheve verdeling van deze variabelen in figuren 2.5 en 2.6. Wanneer nu een soort die alleen in het binnenland voorkomt gemodelleerd wordt met behulp van de multiple regressietechniek zal deze soort voor de variabelen zout alleen maar nullen hebben en zal het toevoegen van deze variabelen aan het model dus geen verbetering opleveren. Ditzelfde geldt voor de variabele cpaf2, ook deze variabelen heeft bijna overal in Nederland nullen, behalve op enkele zeer specifieke plekken zoals in de veengebieden (bijvoorbeeld ten westen en zuiden van de Vinkeveense plassen, ten zuiden van Woerden, ten noorden van Amsterdam en in de Peel) en in Zuid-Nederland langs de riviertje de Beerze en de Dommel. Deze hoge waarden zijn een gevolg van de vervuiling door de cadmium fabriek in Budel. Verrassend is dat in meer dan eenderde deel van de modellen de variabele vegetatietype niet meegenomen wordt, terwijl op voorhand gedacht werd dat dit samen met de fysisch geografische regio een belangrijke uitbreiding zou zijn ten opzichte van het MOVE 2 model. Het ontbreken van deze variabele kan waarschijnlijk gedeeltelijk verklaard worden door de aanwezigheid van vele soorten die niet kenmerkend zijn voor één specifiek vegetatietype.

Tabel 5.2 Optimale modellen in uiteindelijke selectie. Per modeltype en per variabelen is aangegeven het aantal maal dat zij geselecteerd zijn.

model	aantal	percentage		variabelen	aantal	percentage
model a	12	1.7		r	647	93.8
model b	22	3.2		r^2	642	93.0
model c	17	2.5		n	666	96.5
model d	25	3.6		n^2	657	95.2
model e	23	3.3		f	684	99.1
model f	37	5.4		f^2	644	93.3
model g	55	8.0		s	399	57.8
model h	50	7.2		cpaf2	265	38.4
model I	58	8.4		fgr	597	86.5
model j	118	17.1		veg	441	63.9
model z	103	14.9		interactie r-n	602	87.2
AIC compleet	53	7.7		interactie r-f	596	86.4
BIC leeg	91	13.2		interactie n-f	595	86.2
BIC MOVE 2	26	3.8				
	690					

In de optimale regressieverzameling komen enkele soorten voor die maar door één of twee variabelen gemodelleerd worden, maar in alle modellen zitten één of meerdere temporele variabelen, zodat het mogelijk is gevolgen van veranderingen in de tijd (scenario analyses) te evalueren. Enkele soorten die slechts door enkele variabelen gemodelleerd worden zijn: *Arctium lappa* (grote klit, cbs nummer 83) deze wordt bijvoorbeeld alleen gemodelleerd door de aanwezige hoeveelheid nutriënten (n^2) en de mate van vochtigheid (f). Volgens het model heeft *Arctium lappa* een sterke voorkeur voor droge, nutriëntrijke standplaatsen *Senecio congestus* (moerasandijvie, cbs nummer 1184) wordt alleen gemodelleerd door de hoeveelheid aanwezige nutriënten (n^2) en de mate van vochtigheid (f en f^2); met een sterke voorkeur voor nutriëntrijke, vochtige standplaatsen.

5.4 Twee voorbeelden

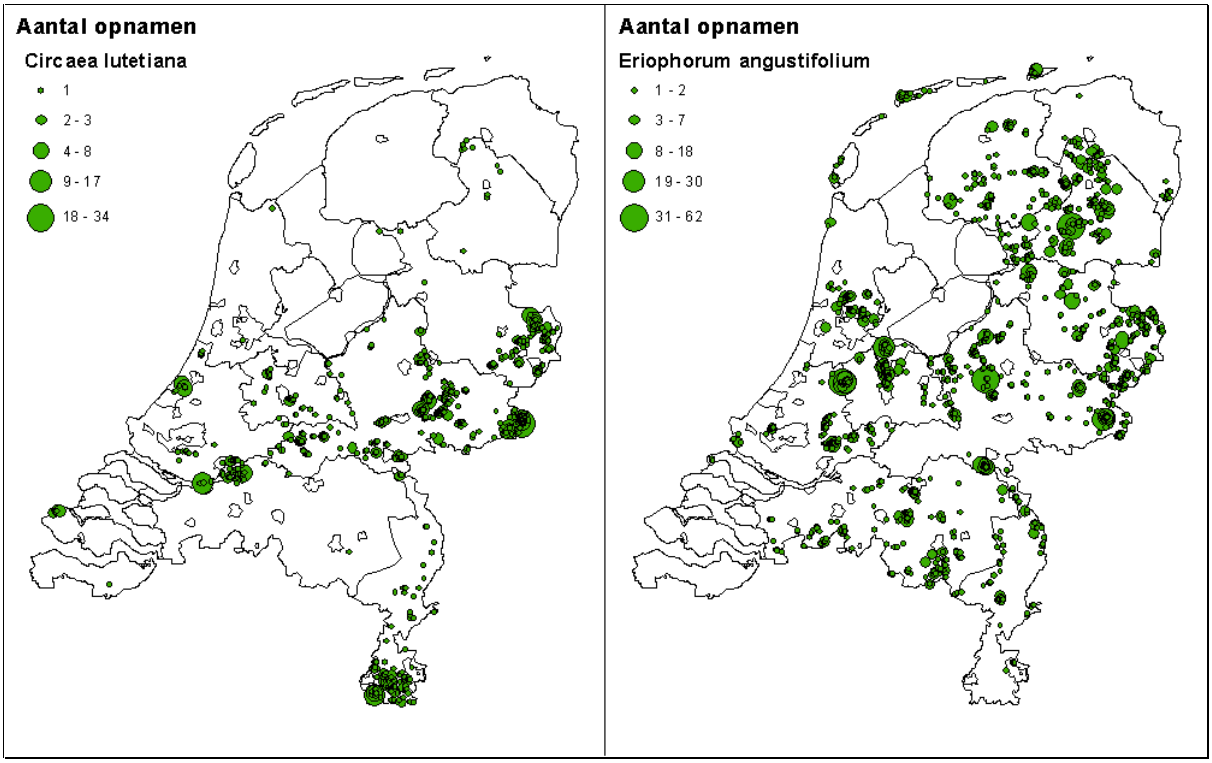
Tot slot nog twee voorbeelden van responsiekrommen van twee willekeurig gekozen soorten, namelijk het groot heksenkruid (*Circaea lutetiana*) en de orchidee het veenpluis (*Eriophorum angustifolium*).
Het groot heksenkruid is volgens Heukels' een plant van vochtige tot natte loofbossen op voedselrijke grond en zij komt tevens voor in grienden. De soort is vrij algemeen in Zuid-Limburg (Heuvelland), plaatselijk komt zij voor in het **S**ubcentreuroop disctrict (oostelijk Gelderland en oostelijk Overijssel) en het **F**luviatiel district (rivierengebied inclusief Zeeland) en langs de binnenduinrand in het **R**enodunaal district (de kalkrijke duinen ten zuiden van Bergen (NH)); in de rest van Nederland is de soort zeldzaam. De soort behoort tot de ecologische groepen H27 (bossen en struwelen op natte, matig voedselrijke bodem) en H47 (bossen en struwelen op vochtige matig voedselrijke bodem)

Het groot heksenkruid is op 1059 opnamepunten (figuur 5.1) waargenomen. Zij heeft een voorkeur (zie bijlage I.b in Bakkenes *et al.,* 2002) voor matig zure tot zwakzure, matig stikstofrijke tot stikstofrijke omstandigheden (r-waarden[9] tussen de 3.6 en 7.7 met een gemiddelde waarde van 6.3; n-waarden[9] tussen 3.8 en 8.0 met een gemiddelde waarde van 6.3); in een vochtige tot natte omgeving (f-waarden[9] tussen 4.2 en 8.7 met een gemiddelde waarde van 6.4). De soort wordt over het complete zoutbereik waargenomen, waarbij de meeste waarnemingen liggen in de niet zoute gebieden, gemiddelde waarde is 0.06 (bijlage I.c in Bakkenes *et al.,* 2002). De meeste waarnemingen liggen in de 'hogere zandgronden', 'rivierengebied' en het 'heuvelland' (bijlage I.d in Bakkenes *et al.*, 2002). Deze waarnemingen komen goed overeen met de beschrijvingen volgens Heukels'.

Volgens de berekende regressievergelijking ligt het optimum (zie bijlage II.c in Bakkenes *et al.*, 2002) van deze plant bij r-, n- en f-waarden[9] van respectievelijk 6.4, 6.1 en 6.7 in de fysisch geografische regio 'Afgesloten zeearmen' en het vegetatietype heide. Maar ook in het 'Heuvelland', 'Zeekleigebied', 'Rivierengebied', 'Hogere zandgronden Noord' en de vegetatietypen heide en loofbos worden hoge kansen op voorkomen voorspeld. Dit komt goed overeen met waarnemingen (figuur 5-1).

Het veenpluis komt volgens Heukels' voor op natte, zure grond in heiden, graslanden, zeggemoerassen, vennen en duinvalleien. Deze soort is vrij algemeen in het Drents en Kempens district, zij is vrij zeldzaam in het Laagveen-, Gelders, Subcentreuroop en Waddendistrict, elders is zij zeer zeldzaam. De soort behoort tot de ecologische groepen G21 (grasland op natte voedselarme zure bodem), G22 (grasland op natte voedselarme zwak zure bodem) en V11 (verlandingsvegetatie in voedselarm zuur water).
Het veenpluis is op 2417 opnamepunten (figuur 5.1) waargenomen. Zij heeft een voorkeur voor redelijk zure, zeer stikstofarme tot matig stikstofrijke omstandigheden (r-waarden[9] tussen de 1.6 en 6.6 met een gemiddelde waarde van 3.7; n-waarden[9] tussen 1.2 en 5.6 met een gemiddelde waarde van 2.7); in een vochtig tot natte omgeving (f-waarden[9] tussen 5.3 en 10.5 met een gemiddelde waarde van 8.4), zie bijlage I.b (Bakkenes *et al.*, 2002). De soort heeft een slechte zouttolerantie en wordt bijn a alleen in zoete gebieden waargenomen (bijlage I.c in Bakkenes *et al.*, 2002). Alle waarnemingen (bijlage I.d in Bakkenes *et al.*, 2002) liggen voornamelijk op heiden en loofbossen in alle fysisch geografische regio's met een voorkeur voor de hogere zandgronden boven de grote rivieren en het zeekleigebied. Deze waarnemingen komen goed overeen met de beschrijvingen volgens Heukels'.

[9] waarden in Ellenberg eenheden. NB Door de uitgevoerde berekeningen zijn dit *niet* de oorspronkelijke Ellenberg eenheden. Allereerst is de schaal verkleind door per opnamen de gemiddelde Ellenberg waarden van de waargenomen soorten te nemen en vervolgens zijn bij de bepaling van de optima de optimum waarden die buiten het Ellenberg bereik vallen teruggeschaald.



Figuur 5.1 Verspreiding van het aantal waarnemingen van groot heksenkruid (links) en veenpluis (rechts)

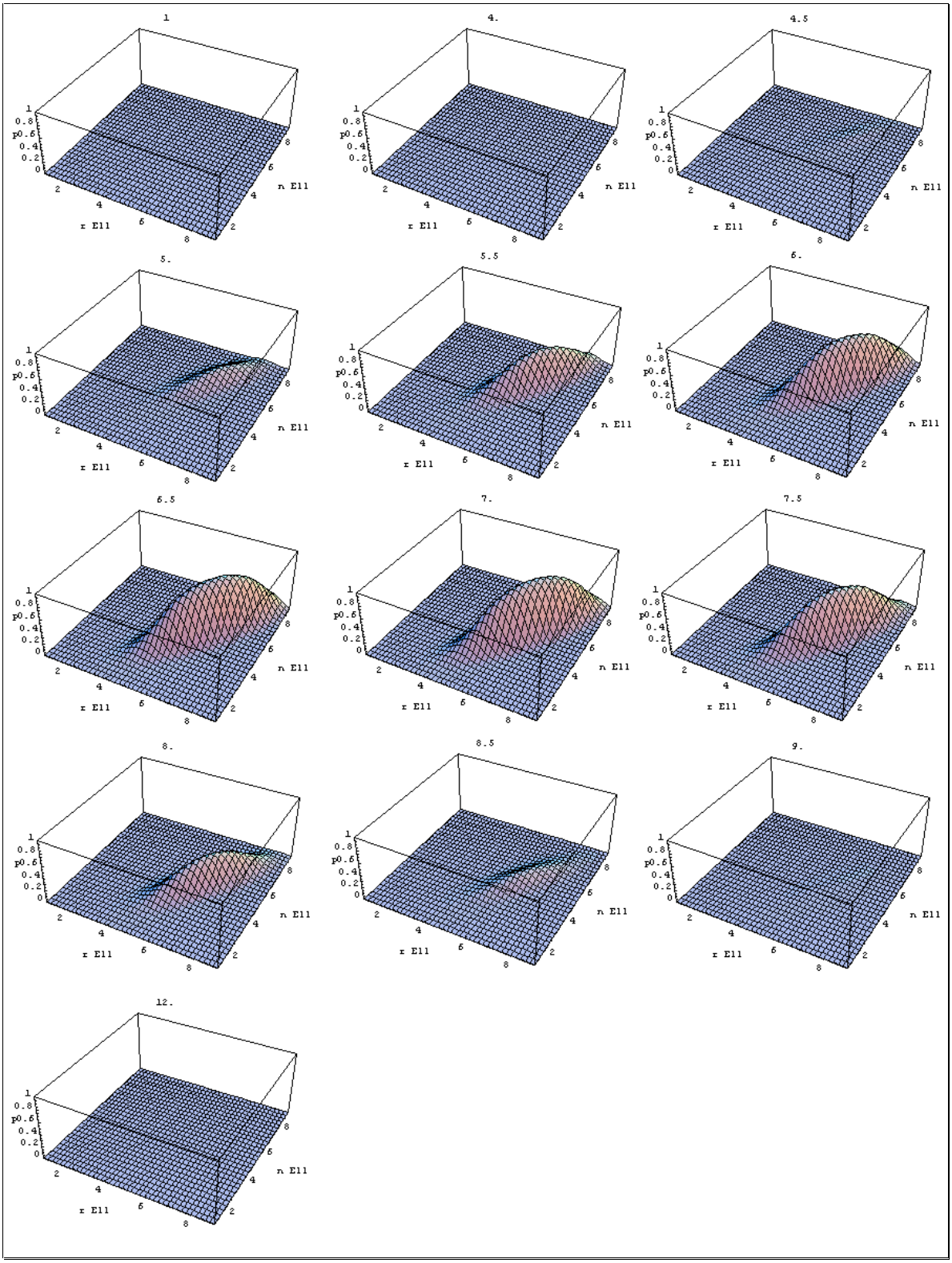
De berekende regressievergelijking voor het veenpluis heeft het optimum (zie bijlage II.c in Bakkenes *et al.*, 2002) liggen bij de r-, n- en f-waarden[10] van respectievelijk 3.0, 1.0 en 9.0. Zoals uit figuur 5-3 blijkt doet deze soort het goed bij hoge vochtigheid (hoge f-waarden) en lage waarden voor de nutriëntenbeschikbaarheid en zuurgraad. Het theoretisch optimum (bijlage II.c in Bakkenes *et al.*, 2002) van deze soort ligt niet in het 'Laagveengebied' (figuur 5.3), maar in het 'Heuvelland'. Maar behalve langs de Maas zullen de benodigde natte omstandigheden in het 'Heuvelland' waarschijnlijk niet gevonden worden. De andere regio's waar deze soort het goed moet doen volgens de regressievergelijking zijn: het 'Laagveengebied', het 'Zeekleigebied', de 'Afgesloten zeearmen' en in mindere mate in het 'Duingebied' en de 'Hogere zandgronden noord'.

De figuren 5-2 en 5-3 geven een drie-dimensionaal beeld van de respons van zowel het groot heksenkruid als het veenpluis onder invloed van (veranderende) (a)biotische omstandigheden. In beide figuren zijn de variabelen zoutgehalte, cpaf2, fysisch geografische regio's en vegetatietype constant en wordt er gevarieerd in de overige drie variabelen, r, n en f. Per figuur is het vochtgetal constant (getal boven figuur) en laten de x- en y-as het bereik voor zuurgraad en nutriëntenbeschikbaarheid zien. Op de z-as staat de kans op voorkomen bij deze combinatie van waarden.

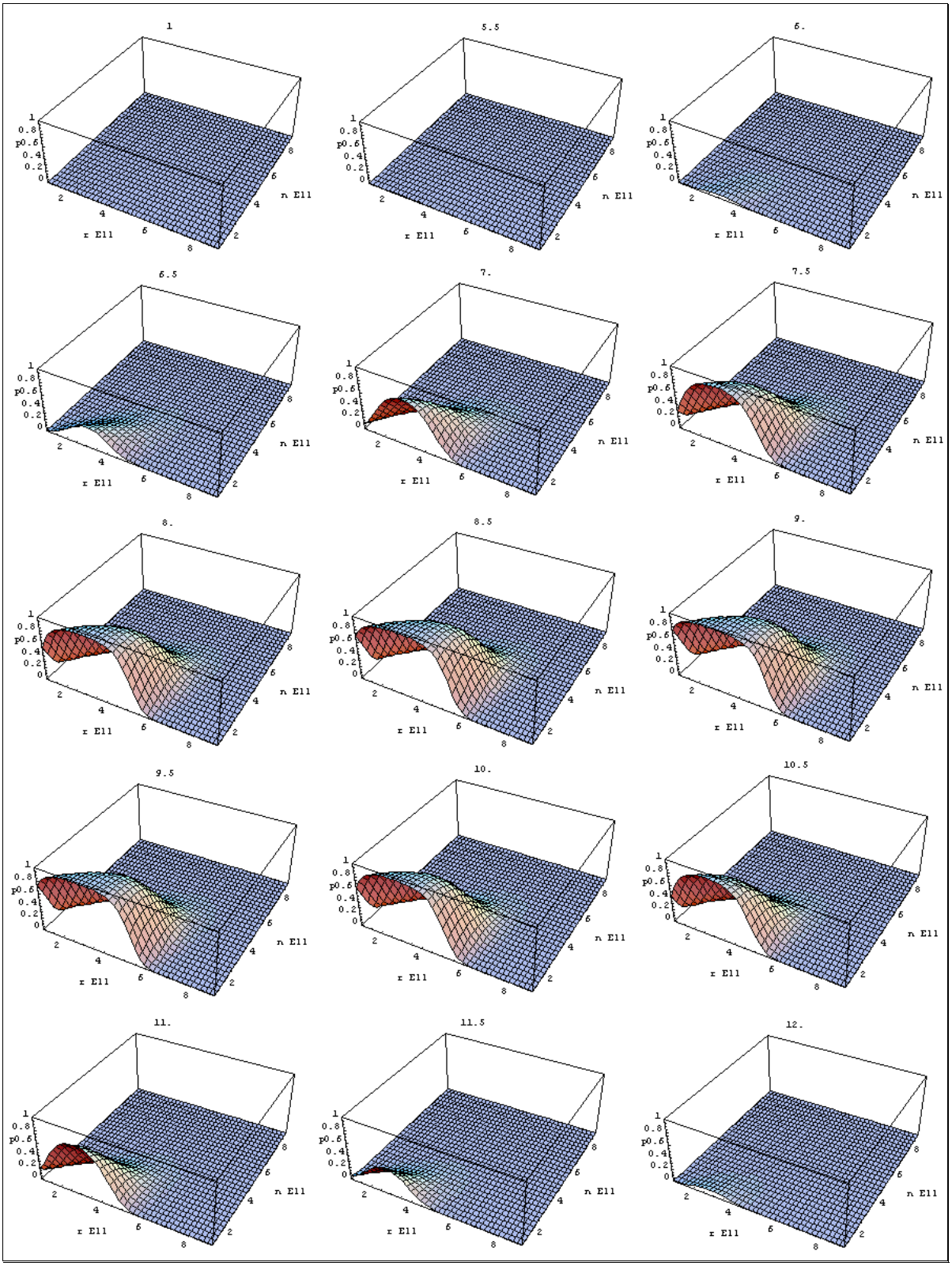
Wat opvalt bij het groot heksenkruid (figuur 5-2) is het bestaan van een lineair verband tussen r en n, wanneer de r-waarde toeneemt moet de n-waarde ook toenemen voor deze soort om zich te kunnen blijven handhaven. Dit zou eventueel een gevolg kunnen zijn van de bestaande correlatie tussen r en n (zie tabel 2.2), maar dan zou dit zich bij de meeste modellen voor moeten doen en dit is niet het geval. Bij vochtwaarden onder de 5 en boven de 8 komt deze soort niet voor.

[10] waarden in Ellenberg eenheden. NB Door de uitgevoerde berekeningen zijn dit *niet* de oorspronkelijke Ellenberg eenheden. Allereerst is de schaal verkleind door per opnamen de gemiddelde Ellenberg waarden van de waargenomen soorten te nemen en vervolgens zijn bij de bepaling van de optima de optimum waarden die buiten het Ellenberg bereik vallen teruggeschaald.

Het veenpluis komt alleen voor bij lage pH-waarden (Ellenberg waarden lager dan vijf) en stikstofarme (Ellenberg waarden lager dan 4) omstandigheden. Wanneer de vochtigheid toeneemt verschijnt deze soort opeens, de kans op het aantreffen van deze soort neemt dan toe



Figuur 5.2 Drie dimensionale weergave van het verloop van de kans van voorkomen voor Circaea lutetiana (groot heksenkruid) onder invloed van een verandering van de Ellenberg vocht, nutriënten- en zuurgetallen. De overige gebruikte instellingen zijn: cpaf2 is 0.0, zout is 0.0, fysisch geografische regio 'Afgesloten zeearmen' en begroeiingstype 'Heide'.



Figuur 5.3 Drie dimensionale weergave van het verloop van de kans van voorkomen voor Eriophorum angustifolium (veenpluis) onder invloed van een verandering van de Ellenberg vocht-, nutriënten-- en zuurgetallen. De overige gebruikte instellingen zijn: cpaf2 is 0.0, zout is 0.0, fysisch geografische regio 'Laagveengebied' en het begroeiingtype 'heide'.

6. Discussie en conclusies

De in dit rapport afgeleide optimale regressievergelijking is niet per definitie het meest optimale model. Bij de stapsgewijze regressie is de invloed van het gekozen startmodel zeer bepalend voor het uiteindelijke berekende resultaat en door het op voorhand compleet definiëren van de modellen, zoals in de gebruikte methode van §3.2 worden de mogelijkheden van tevoren beperkt. Door te selecteren uit varianten die biologisch het meest zeggen, aangevuld met varianten uit de stapsgewijze regressie, wordt de kans dat het beste model wordt gevonden vergroot. Het alternatief is om alle verschillende combinaties door te rekenen, maar dit leidt dan uiteindelijk tot het doorrekenen van 8192 modellen per soort. Dit is een onhaalbare en weinig efficiënte optie.
Veel belangrijker dan het vinden van het allerbeste model is het vinden van een model waarvan de schattingen zo optimaal mogelijk overeenkomen met de waarnemingen. Dit sluit aan bij het doel en gebruik van deze modellen, namelijk het zo goed mogelijk kunnen voorspellen van het potentieel voorkomen van plantensoorten. Dus het belangrijkste criteria is of de afgeleide regressievergelijking goed 'past' op de waarnemingen, oftewel het hebben van een goede 'goodness of fit' waarde. De aangepaste Hosmer-Lemeshow test (Hosmer *et al*,. 1988) door Bio (2000) is zo'n goodness of fit test. In de uiteindelijke selectie zijn alle modellen meegenomen, zowel de handmatig afgeleide als de stapsgewijze modellen, en is per soort dat model geselecteerd dat de beste fit had, dus de laagste χ^2-waarde. Dit heeft ertoe geleid dat er van de oorspronkelijke 914 soorten er 690 soorten overgebleven zijn. Dit zijn allen zeer goede modellen en deze zijn zeer geschikt voor gebruik in de vragen waarbij dit model ingezet zal gaan worden binnen het milieu- en natuurplanbureau van het RIVM.

Regressiemodellen zijn primair afhankelijk van de gegevens die gebruikt worden om de modellen mee af te leiden. De kwaliteit van de modellen is daardoor sterk afhankelijk van de gebruikte gegevens. Er kunnen kanttekeningen geplaatst worden bij de gebruikte verklarende variabelen, zoals de combiPAF-waarden voor zware metalen en de Ellenberg-waarden. Effecten van zware metalen op bodemorganismen kunnen in het algemeen het best beschreven worden op basis van de concentratie in oplossing of met behulp van de vrije metaalactiviteit (Gregor, 1999). Aangezien er te weinig gegevens beschikbaar zijn om de concentratie in oplossing of om de vrije metaalactiviteit betrouwbaar te voorspellen, is gekozen voor de reactieve fractie. Dit leidt wellicht tot afwijkingen van de daadwerkelijke toxiciteit voor planten wat dan ook van invloed zal zijn op de afgeleide regressiemodellen. De manier waarop de toxische druk van zware metalen in het model meegenomen kan worden vraagt daarom nog extra onderzoek.
Ook zijn er veel lage combiPAF-waarden aanwezig in de dataset (zie figuur 2.5), wanneer een soort bij deze lage waarden aanwezig is, kan er een positief verband afgeleid worden. Voorspellingen wijzen uit dat in de toekomst de concentraties aanwezige zware metalen toe zullen blijven nemen, zodat de omstandigheden van deze soort paradoxaal genoeg alleen maar beter zullen gaan worden. Dit effect treedt altijd op wanneer er geëxtrapoleerd wordt buiten het bereik van de waarden die gebruikt zijn voor het afleiden van een regressie-vergelijking en is daarom ook niet een probleem dat gekoppeld moet worden aan deze specifieke variabele. Maar bij deze variabele is het te verwachte effect wel groot. Het treedt ook op bij de variabele zout (s), maar omdat deze sterk gekoppeld is aan de fysisch geografische regio zal hiervoor waarschijnlijk gecorrigeerd worden in de regressie-vergelijking. De variabelen zuurgetal (r), stikstofgetal (n) en vochtgetal (f) komen over het gehele bereik in de gebruikte dataset voor en het gevaar van extrapoleren buiten de modelwaarden doet zich bij deze variabelen dan ook niet voor.

Een probleem bij de Ellenberg-waarden is dat deze waarden geen fysieke metingen zijn, maar inschattingen op basis van expert judgement (Ellenberg, 1992). Er is een aantal studies bekend waarin de Ellenberg waarden worden gerelateerd aan werkelijk gemeten grootheden, maar de variatie in de uitkomsten van die studies is tamelijk groot (Alkemade, 1996; Ertsen, 1998; Wamelink, 2002). Door het ontbreken van geschikte alternatieve voor de Ellenberg indicatiewaarden blijven deze waarden het best mogelijke alternatief.

Een punt van aandacht is het *niet* meenemen van interactietermen voor de geklassificeerde variabelen fgr en veg. Door het niet meenemen is er geen variatie op de ligging van de optima tussen de verschillende fysisch geografisch regio's en/of vegetatietypen mogelijk. Het optimum van een soort voor bijvoorbeeld de zuurgraad verschuift niet afhankelijk van de fysisch geografische regio. In een vervolg onderzoek, wanneer bijvoorbeeld nog extra variabelen of toxiciteitwaarden die op een andere wijze afgeleide zijn worden meegenomen kan het misschien wenselijk zijn om extra interactietermen mee te nemen. Maar toevoegen van extra interactietermen betekent ook het toevoegen van extra vrijheidsgraden aan het model.

Ondanks deze kanttekeningen zijn wij van mening dat de afgeleide modellen significant verbeterd zijn ten opzichte van de vorige versie van MOVE (MOVE 2) en MOVE 3 (De Heer *et al.*, 2000). In het MOVE 3 model van De Heer *et al.* (2002) hebben alle soorten hetzelfde model, model z in deze analyse. Een belangrijk resultaat van deze studie is dat per soort een zo'n geschikt mogelijk model geselecteerd is en dat er dus geen onnodige (ruis)termen meegenomen worden. Een nadelig effect van deze analyse is dat het aantal gemodelleerde soorten is verminderd, maar daarvoor in de plaats zijn alleen die modellen overgebleven met een goede fit op de oorspronkelijke waarnemingen.

Literatuur

Akaike, H., 1973, Information theory and an extension of the maximum likelihood principle. In: B.N. Petrov & F. Cáski, eds., Second international symposium on information theory. Budapest: Akademia Kiadó, pp. 267-281 (reprinted in: S. Kotz & N.L. Johnson, eds., 1992, Breakthroughs in Statistics. Volume 1. New York: Springer-Verlag, pp. 610-624).

Akaike, H., 1974, A new look at statistical model identification. IEEE Transactions on Automatic Control 19, pp. 716-723.

Akaike, H., 1978, A Bayesian analysis of the minimum AIC procedure. Annals of the Institute of Statistical Mathematics 30 A, pp. 9-14.

Alkemade, J.R.M., J. Wiertz & J.B. Latour, 1996. Kalibratie van Ellenbergs milieu-indicatiegetallen aan werkelijk gemeten bodemfactoren, RIVM rapport 711901 016. RIVM, Bilthoven.

Bakkenes, M., D. de Zwart en J.R.M. Alkemade, 2002, Bijlagen bij: MOVE nationaal MOdel voor de VEgetatie versie 3.2, Achtergronden en analyse van modelvarianten, RIVM rapport 408657 010. RIVM, Bilthoven.

Bendel, R.B. & A.A. Afifi, 1977, Comparison of stopping rules in forward regression. Journal of the American Statistical Association 72, pp. 46-53.

Bio, A.M.F. 2000. Does vegetation suit our models? Data and model assumptions and the assessment of species distribution in space, Elinkwijk, Utrecht, 195pp.

Briemle, G. & H. Ellenberg. 1994. Zu Mahvertäglichkeit von Grünpflanzen; Möglichkeiten der practischen Anwendung von Zeigerwerten. Natur und Landschaft, 69Jg. 1994 Heft 4: 139-147.

Buckland, S.T., K.P. Burnham & N.H. Augustin, 1997, Model selection: an integral part of inference. Biometrics 53, pp. 603-618.

Burman, P. & D. Nolan, 1995, A general Akaike-type criterion for model selection in robust regression. Biometrika 82, pp. 877-886.

Carrol, JD, 1972. Individual differences and multidimensional scaling. In: R.N. Shepard, A.K. Romney & S.B. Nerlove (eds.): Multidimensional scaling. Theory and application in the behavioral sciences. Vol. 1, Seminar Press, New York. p. 105-155.

CBS. 1993. Botanisch basisregister. CBS, Voorburg/Heerlen.

Constanze, M.C. & A.A. Afifi, 1979, Comparison of stopping rules in forward stepwise discriminant analysis. Journal of the American Statistical Association 74, pp. 777-785.

Curtis, J.T. & R.P. McIntosh, 1951. An upland forest continuum in the prairieforest border region of Wisconsian. Ecology 32, pp 476-496.

De Heer, M., R. Alkemade, M. Bakkenes, M. v. Esbroek, A. v. Hinsberg, D. de Zwart, 2000, Move: nationaal Model voor de Vegetatie, versie 3. De kans op voorkomen van ca. 900 plantensoorten als functie van 7 omgevingsvariabelen. RIVM rapport 408657 002. RIVM, Bilthoven.

Efron, B., 1982. The Jackknife, the bootstrap and other resampling plans. SIAM, Philadelphia, 92 pp.

Ellenberg, H., H.E. Weber, R. Dull, V. Wirth, W. Werner & D. Paulissen (eds), 1992. Zeigerwerte von Pflanzen in Mitteleuropa. Erich Goltze, Gotting, 258 pp.

Ertsen, A.C.D., J.R.M. Alkemade & M.J. Wassen, 1998, Calibrating Ellenlberg indicator values for moisture, acidity, nutrient availablity and salinity in the Netherlands. Plant Ecology 135, pp. 113-124.

Fresco, L.F.M., 1982. An analysis of species response curves and of competition from field data sets: some results from heath vegetation. Vegetatio 48, pp 175-185.

Gause, G.F., 1930. Studies on the ecology of the Orthoptera. Ecology 11, pp 307-325.

Gregor, H.D. (ed.). 1999. Proceedings of the Workshop on Effects-based approaches for Heavy Metals. UC ECE Convention on lang-range transboundary air pollution, Task force on Mapping. Schwerin, Germany.

Heikkinen, R.K., 1996, Predicting patterns of vascular plant species richness with composite variables: A meso-scale study in Finnish Lapland. Vegetatio 126, pp. 151-165.

Hill, M.O., 1977. Use of simple discrimant functions to classify quantitative phytosociological data. In: E. Didaym, L. Lebart, J.P. Pagès & R. Tomassone (Eds.): First International Symposium on Data Analysis and Informatics. Vol 1: 181-199. Institut de Recherche d'Informatique et d'Automatique, Le Chesnay, pp 181-199.

Hosmer, D.W., S. Lemeshow & J. Klar, 1988, Goodness of fit testing for the logistic regression model when the estimated probabilities are small. Biometrical Journal 30, pp. 911-924.

Hosmer, D.W. JR. & S. Lemeshow, 1989, Applied logistic regression. New York, John Wiley & Sons.

Janssen, M. 1991. Turnover of cadmium through soil arthropods. PhD thesis, Vrije Universiteit, Amsterdam, The Netherlands

Jongman, R.H.G. & Th.J. van de Nes (Eds.), 1982. Beken op de Veluwe. Een onderzoek naar de mogelijkheden voor herstel en behoud. Begeleidingscommissie Proefgebied Nationaal Landschap Veluwe, Arnhem. 112 pp.

Jongman, R.H.G., C.J.F. ter Braak & O.F.R. van Tongeren (Eds.), 1987. Data analysis in community and landscape ecology. Pudoc, Wageningen. 299 pp.

Kaldewaij, A. & J. van Tiel, 1986, Voortgezette wiskunde, Deel 3 Vectorrekening en matrixrekening, Bohn, Scheltema & Holkema, Utrecht/Antwerpen, 149 pp.

Kay, R. & S. Little, 1986, Assessing the fit of the logistic model: a case study of children with the haemolytic uraemic syndrome. Applied Statistics 35, pp. 16-30.

Kitagishi, K. and I. Yamane. 1981. Heavy metal pollution in soils of Japan. Japan Scientific Sociecites Press, Tokyo, Japan.

Klepper O. & D. van de Meent, 1997. Mapping the potentially affected fraction (PAF) of species as an indicator of generic toxic stress. RIVM rapport 607504001. RIVM, Bilthoven.

Kros, J., 1998. De modellering van de effecten van verzuring, vermesting en verdroging voor bossen en natuurterreinen ten behoeve van de Milieubalans, Milieuverkenningen en Natuurverkenningen. SC-DLO, Wageningen.

Latour, J.B., I.G. Staritsky, J.R.M. Alkemade & J. Wiertz, 1997. De Natuurplanner, Decision Support System natuur en milieu, versie 1.1. RIVM rapport 711901 019. RIVM, Bilthoven.

Mallows, C., 1973, Some comments on C_p Technometrics 15, pp. 661-675.

McCullagh, P. & J.A. Nelder, 1983. Generalized linear models. Chapman and Hall, London, 261 pp.

Montgomery, D.C. & E.A. Peck, 1992, Introduction to linear regression analysis, second Editon, John Wiley & Sons, inc., New York, 527 pp.

Oosterbeek, J.G.B., J.R.M. Alkemade, J. Wiertz, H.F. van Dobben & G.W.W. Wamelink, 1997. Het modelleren van de effecten van natuurbeheer ten behoeve van MOVE. RIVM-rapport 715001006. RIVM, Bilthoven.

RIVM, 2000a. Milieubalans 2000. Samsom B.V., Alpen aan de Rijn.

RIVM, 2000b. Nationale Milieuverkenning 5. Samsom B.V., Alphen aan de Rijn.

RIVM, 2001a. Milieubalans 2001. Kluwer, Alpen aan de Rijn.

RIVM, 2001b. Natuurbalans 2001. Kluwer, Alphen aan de Rijn.

Runhaar, J., F.Witte & R. Jongman, 1994. Ellenberg-indicatiewaarden: verbeteringen met reciprocal averaging? Landschap 11/1: 41-48.

Runhaar, J., R. Alkemade, S.M. Hennekens, J. Wiertz & M. van ’t Zelfde, 2002. Afstemming biotische responsmodules DEMNAT-SMART/MOVE. RIVM report 408657 008. RIVM, Bilthoven.

Schaminée, J.H.J., A.H.F. Stortelder, V. Westhoff, J.J. Barkman, H. Doing & L. van Duuren, 1995. Inleiding tot de plantensociologie. Grondslagen, methoden en toepassingen. Opulus Press, Uppsala.

Schwarz, G., 1978, Estimating the dimensions of a model. The Annals of Statistics 6, pp. 461-464.

Shibata, R., 1989, Statistical aspects of model selection. In: J.C. Willems (ed.) From data to model, pp. 215-240. London: Springer-Verlag.

Swartzman, G. & C. Huang, 1992, Spatial analysis of Bering Sea groundfish survey data using generalized additive models. Canadian Journal of Fishery and Aquatic science 49, pp. 1366-1378.

Tiktak A, J.R.M. Alkemade, J.J.M. van Grinsven & GB Makaske. 1998. Modeling Cadmium Accumulation at a regional scale in the Netherlands. Nutrient Cycling in Agroecosystems 50, pp. 209-222.

Tiktak, A. 1999. Modelling non-point source pollutants in soils. Proefschrift Universiteit van Amsterdam.

Tiktak, A., J.G. Otte, P.F.A.M. Römkens & W. de Vries. 2000. Partition relationships for use in Effects-Based Approaches for Heavy Metals. Proceedings of the Workshop on Effects-based approaches for Heavy Metals. UN ECE Convention on long-range transboundary air pollution, Task force on Mapping. Bratislava, Slovakia.

Van de Rijt, C.W.C.J., L. Hazelhoff & C.W.P.M. Blom, 1996, Vegetation zonation in former tidal area: a vegetation-type response model based on DCA and logistic regression using GIS. Journal of Vegetation Science 7, pp. 505-518.

Van der Hoek, D.C.J., M. Bakkenes & J.R.M. Alkemade, 2000. Natuurwaardering in de Natuurplanner. Toepassing voor de VIJNO. RIVM rapport 408657 004. RIVM, Bilthoven.

Van der Hoek, D.C.J., W.H. Hoffmans, A. van Hinsberg, M. van Esbroek, 2002. Ecologische effectberekening voor de 2e Nationale Natuurverkenning: terrestrische systemen. RIVM rapport 408664 002, RIVM Bilthoven.

Van der Meijden, R, 1990. Heukels' Flora van Nederland, Wolters-Noordhoff, Groningen, 662 pp.

Van Hinsberg, A., H. Dijkstra, P. Hinssen, K. Kramer, F. Leus, R. Reiling, R. Reijnen, M. van de Tol & J. Wiertz, 1999. Stroomlijning Natuurplanbureau modellen; inventarisatie en keuze voor modellen voor Natuur, Landschap en Bos. RIVM rapport 408662 001. RIVM, Bilthoven.

Walker, C.H., 1987. Kinetic models for predicting bioaccumulation of pollutants in ecosystems. Environ Poll 44, pp. 22-240.

Wamelink, G.W.W., V. Joosten, H.F. van Dobben & F. Berendse, 2002, Validity of Ellenberg indicator values judged from physico-chemical field measurements, Journal of vegetation science.

Wamelink, G.W.W., H.F. van Dobben, J.R.M. Alkemade & J. Wiertz, 1997. Milieugevoeligheid van de Nederlandse flora; aanvulling van de door Briemle & Ellenberg (1994) geschatte indicatiegetallen. IBN-DLO, Wageningen.

Webster, R. & A.B. McBratney, 1989. On the Akaike Information Criterion for choosing models for variograms of soil properties. Journal of Soil Science 40, pp. 493-496.

Webster, R. & M.A. Oliver (Eds.), 1990, Statistical methods in soil and land resource survey. Oxford University Press, New York, 316 pp.

White, G.C. & R.E. Bennets, 1996, Analysis of frequency count data using the negative binomial distribution. Ecology 77, pp. 2549-2557.

Whittaker, R.H., 1956. Vegetation of the Great Smoky Mountains. Ecological Monographs 26, pp. 1-80.

Wiertz, J., J. van Dijk & J.B. Latour, 1992. MOVE: vegetatiemodule; de kans op voorkomen van ca. 700 plantensoorten als functie van vocht, pH, nutriënten en zout. RIN-rapport 92/94; RIVM rapport 711901006. Wageningen/Bilthoven.

Yee, T.W. & N.D. Mitchell, 1991, Generalized additive models in plant ecology. Journal of Vegetation Science 2, pp. 587-602.

Bijlage 1 Verzendlijst

1. dr. H.F. van Dobben (Alterra)
2. ir. W. Wamelink (Alterra)
3. ir. J.P. Mol (Alterra)
4. dr. J. Schaminée (Alterra)
5. drs. S.M. Hennekens (Alterra)
6. dr. P. Opdam (Alterra)
7. ir. J. Kros (Alterra)
8. dr.ir. W. de Vries (Alterra)
9. drs. W.B. Harms (Alterra)
10. dr. J. Runhaar (Alterra)
11. dr. A. vd. Zande (Alterra)
12. dr. M.J.S.M. Reijnen (Alterra)
13. B. Koolstra (Alterra)
14. D. van Zaane (DLO Centraal)
15. dr. C.J.F. ter Braak (CPRO-DLO)
16. drs. R. van Ek (RIZA)
17. drs. R. Meijers (EC-LNV)
18. dr. H. Smit (EC-LNV)
19. dr. ir. J.P.M. Witte (LUW, vakgroep Waterhuishouding)
20. prof. dr. F. Berendse (LUW, vakgroep TON)
21. dr. A. Schaffers (LUW, vakgroep TON)
22. ir. H. van Oene (LUW, vakgroep TON)
23. prof. dr. K.V. Sykora (LUW, vakgroep TON)
24. dr. H. Olff (LUW, vakgroep TON)
25. dr. R. Jongman (LUW)
26. dr. A. Barendregt (UU, vakgroep Milieukunde)
27. dr. M. Wassen (UU, vakgroep Milieukunde)
28. dr. R. van Diggelen (RUU)
29. prof. dr. J.M. van Groenendael (KUN)
30. prof. dr. J. van Andel (RUG)
31. drs. F. Bekhuis (prov. Gelderland)
32. drs. M. Rijken (prov. Gelderland)
33. drs. L. Jalink (prov. Zuid-Holland)
34. dr. A.J.M. Janssen (KIWA)
35. dr. A. Meuleman (KIWA)
36. dr. D. Ertsen (IWACO)
37. dr. N.J.M. Gremmen (Data Analyse Ecologie)
38. dr. O.F.R. van Tongeren (Data Analyse Ecologie)
39. dr. M. van der Peijl (ESM)
40. dr. R. van der Meijden (Rijksherbarium)

41. dr. C.L.G. Groen (FLORON)
42. dr. K. Kanters (CML)
43. dr. E. de Hullu (SBB)
44. dr. B.F. van Tooren (NM)
45. Depot Nederlandse Publikaties en Nederlandse Bibliografie
46. Directeur Generaal. H.A.P.M. Pont
47. prof. ir. N.D. van Egmond
48. ir. F. Langeweg
49. dr. L.C. Braat
50. ing. H. Bredenoord
51. dr. J. Notenboom
52. drs. R. van Oostenbrugge
53. dr. D. Verkaar
54. ir. T. Bresser
55. drs. A. van der Giessen
56. drs. R. Wortelboer
57. drs. W. Ligtvoet
58. ir. J. van Dam
59. dr. ir. W.A.J. van Pul
60. dr. A.L.M. Dekkers
61. ir. M. Vonk
62. ir. R. van den Berg
63. ing. G.P. Beugelink
64. drs. B.J.E. ten Brink
65. dr. ir. J.J.M. van Grinsven
66. ir. D.C.J. van der Hoek
67. ir. O. Knol
68. dr. A. Tiktak
69. drs. J. Wiertz
70. drs. S. Sollie
71. drs. I. Soenario
72. SBD/Voorlichting & Public Relations
73. Bureau Rapportenregistratie
74. Bibliotheek RIVM
75. Bibliotheek Alterra

74-79. Auteurs

80-94. Bureau Rapportenbeheer

95-120. Reserve exemplaren