

Handleiding voor het gebruik van multivariate analysetechnieken in de ecologie

M.M. van Katwijk¹
C.J.F. ter Braak²

2008

¹Afdeling Milieukunde
Radboud Universiteit Nijmegen
²Afdeling Biometrics
Wageningen Universiteit en Research Centrum

Citeren: van Katwijk MM, ter Braak CJF (2008) Handleiding voor het gebruik van multivariate analysetechnieken in de ecologie. Ecoscience, Universiteit Nijmegen (Versie 1.1).

N.B. Dit is dezelfde handleiding als uit 2003, met kleine verbeteringen, m.n. in Table 3.

Dit is een online publicatie op www.ecoscience.nl

Dankwoord van de eerste auteur: Onno van Tongeren: heel veel dank voor het geduld waarmee je mij alles uitlegde in een tijd waarin er nog geen computerhandleidingen waren, begin tachtiger jaren, en Theo de Boo, dank voor je vele statistische uitleg in de negentiger jaren.

Handleiding voor het gebruik van multivariate analysetechnieken in de ecologie

M.M. van Katwijk
C.J.F. ter Braak

2008

Afdeling Milieukunde
Radboud Universiteit Nijmegen
Postbus 9010
6500 GL Nijmegen

Biometrics
Wageningen Universiteit en Research Centrum
Postbus 100
6700 AC Wageningen

INHOUDSOPGAVE

| | |
|---|----|
| Inhoudsopgave | 5 |
| 1. Inleiding | 1 |
| 2. Voorbewerking van de gegevens | 5 |
| 2.1. Taxonomische reductie | 5 |
| 2.2. Transformatie van de abundantie van de soorten | 5 |
| 2.3. Transformatie van de milieugegevens | 6 |
| 2.3.1. Continue milieuv variabelen (logtransformatie)..... | 6 |
| 2.3.2. Nominale variabelen | 7 |
| 2.3.3. Standaardisatie | 7 |
| 2.3.4. Missing values bij milieuv variabelen | 8 |
| 2.4. Omzetting naar en manipulatie van een 'condensed format'..... | 8 |
| 3. Classificatie..... | 9 |
| 3.1. Inleiding | 9 |
| 3.2. TWINSPAN..... | 9 |
| 3.2.1. Inleiding | 9 |
| 3.2.2. Interpretatie van TWINSPAN..... | 10 |
| 3.3. FLEXCLUS | 10 |
| 3.4. Optimaliteit van clustering..... | 11 |
| 3.5. Algemene richtlijn: hoe nu te classificeren?..... | 11 |
| 3.6. Analyse van CLUSTERS..... | 12 |
| 3.7. Lezen..... | 13 |
| 4. Ordinatie | 15 |
| 4.1. Inleiding | 15 |
| 4.2. Directe versus indirecte gradiëntanalyse technieken..... | 17 |
| 4.3. Ordinatie op basis van een lineair responsmodel (PCA en RDA)..... | 18 |
| 4.3.1. PCA- en RDA-ordinatie op basis van soortensamenstelling..... | 18 |
| 4.3.2. PCA van milieuv variabelen | 19 |
| 4.4. Ordinatie op basis van een eentoppig responsmodel (CA, CCA, DCA en DCCA) | 19 |
| 4.4.1. CA (Correspondence Analysis) | 19 |
| 4.4.2. CCA (Canonical Correspondence Analysis) | 20 |
| 4.4.3. DCA (Detrended Correspondence Analysis)..... | 20 |
| 4.4.4. DCCA (Detrended Canonical Correspondence Analysis)..... | 20 |
| 4.5. Selectie van de milieuv variabelen | 21 |
| 4.5.1. Inleiding | 21 |
| 4.5.2. Het belang van de milieuv variabelen: t-values en V.I.F..... | 21 |
| 4.5.3. Controle van de selectie: eigenvalues vergelijken..... | 22 |
| 4.6. Wanneer welke ordinatietechniek..... | 22 |
| 4.6.1. Direct of indirect..... | 22 |
| 4.6.2. Lineair of ca. Gaussisch..... | 23 |
| 4.6.3. Detrending of niet | 23 |
| 4.7. Synthetische (samengestelde) milieuparameters | 24 |
| 4.7.1. Naar eigen inzicht | 24 |
| 4.7.2. Productvariabelen | 24 |
| 4.7.3. Canonische ordinatieas | 24 |
| 4.8. Verklarende variabele of responsvariabele?..... | 25 |
| 4.9. Constructie van een ordinatiediagram en interpretatie | 26 |

| | |
|--|----|
| 4.9.1. Weergave van soorten en monsterpunten bij PCA en RDA | 26 |
| 4.9.2. Weergave van soorten en monsterpunten in CA, CCA, DCA en DCCA | 26 |
| 4.9.3. Weergave van numerieke milieuvariabelen..... | 26 |
| 4.9.4. Weergave van nominale milieuvariabelen..... | 27 |
| 4.9.5. Eenheden op de assen | 27 |
| 4.10. Lezen..... | 28 |
| 5. Literatuur | 29 |

1. INLEIDING

Hoe moet men gegevens verwerken die betrekking hebben op meerdere objecten en meerdere variabelen? In onderzoek op vrijwel ieder vakgebied heeft men te maken met dergelijke datasets. Een taalkundige heeft misschien 2000 klanken opgenomen en 5 kenmerken van iedere klank onderzocht. Welke eigenschappen heeft een klank om als 'A' herkend te worden? Een psycholoog heeft 100 proefpersonen gevraagd naar een tiental facetten van hun opvoeding en probeert een verband met hun huidige rook-, eet- en drinkgedrag te leggen. Een ecoloog heeft op 200 locaties alle plantensoorten genoteerd en 10 eigenschappen van bodem en grondwater geanalyseerd. Welke verschuivingen in vegetatie kan men verwachten bij toenemende verzuring van de bodem?

Zodra meerdere objecten op meerdere eigenschappen onderzocht zijn, treedt men op het terrein van de multivariate analyse. Bij grote aantallen objecten en/of eigenschappen zijn de analyses niet meer zonder computer uit te voeren. Dit is wellicht de reden dat gegevens die met veel moeite en middelen verzameld zijn, soms niet gebruikt worden.

Deze handleiding is niet bedoeld voor computerfanaten, maar voor mensen die bereid zijn om serieus tijd te besteden aan de verwerking van hun gegevens. Er is uitgegaan van gegevens uit het vakgebied ecologie. Deze gegevens zijn over het algemeen gecompliceerd. De hier gepresenteerde technieken zijn geschikt voor zowel gecompliceerde als eenvoudige datasets. Er zijn m.n. 2 eigenschappen van (sommige) ecologische gegevens die de analyse gecompliceerder maken:

1. Veel verbanden zijn niet lineair. Veelvoorkomend is het eentoppige (unimodale, gaussische) verband. Als een soort bijvoorbeeld voorkomt op plaatsen waar het niet te droog maar ook niet te nat is. Niet te voedselarm, niet te voedselrijk, maar net ertussenin. Dit verband zou je kunnen aanduiden als het 'té-is-nooit-goed'-verband, of het 'net-ertussenin'-verband.
2. Bij ieder object zijn niet evenveel variabelen aanwezig. Bijvoorbeeld in sloot A komen 3 soorten planten voor terwijl in sloot B 18 verschillende plantensoorten worden aangetroffen. In deze handleiding worden technieken voor dit soort gegevens besproken, maar ook voor de 'gewone' datasets, waarbij de gegevens in een rechthoekige, gevulde matrix kunnen worden geplaatst.

Het is van belang dat de computertechnieken als hulpmiddel beschouwd worden. Statistische berekeningen, al dan niet door de computer uitgevoerd, mogen nooit klakkeloos worden uitgevoerd om dan vervolgens met een schijn van objectiviteit 'keiharde' conclusies te trekken.

In deze handleiding worden verschillende stappen in de verwerking van ecologische gegevens beschreven. Het uitgangspunt is een dataset met meerdere objecten (bijvoorbeeld monsterpunten), waarbij van ieder object een aantal eigenschappen (variabelen) bekend is. Deze eigenschappen kunnen gevoelsmatig ('expert judgement') worden ingedeeld in *responsvariabelen* en *verklarende variabelen*. Responsvariabelen zijn variabelen waarvan je verwacht dat ze reageren op onderliggende eigenschappen, dat ze afhankelijk zijn, bijvoorbeeld soorten (die afhangen van milieuvariabelen), eigenschappen, maar soms ook milieuvariabelen (die afhangen van bijvoorbeeld seizoen, klimaat of geografische condities).

Verklarende variabelen zijn variabelen waarvan je verwacht dat ze de responsvariabelen beïnvloeden, en dat ze onafhankelijk zijn, bijvoorbeeld milieuvariabelen, of behandelingen in een experiment. In praktijk is dit onderscheid niet altijd duidelijk te maken (paragraaf 4.8). In alle hier besproken computerprogramma's worden de responsvariabelen aangeduid met 'species', en de verklarende variabelen met 'environmental variables'.

Binnen de ecologie worden multivariate analysetechnieken zoals in deze handleiding beschreven doorgaans gebruikt bij de analyse van veldexperimenten waarbij de responsvariabelen inderdaad de aanwezigheid en abundantie van soorten zijn. Minder bekend is de zeer bruikbare toepassing van multivariate analysetechnieken bij tijdseriemetingen en experimenten waarbij als responsvariabelen een groot aantal plant-, dier- of menseigenschappen, zoals lengte, biomassa, bladoppervlak, aantastingen, vitaliteit, doodsoorzaak, etc. worden genomen (voorbeelden in Kunst et al. 1990, ter Braak & Looman 1994, van Katwijk et al. 1997), alsmede het gebruik van milieuvariabelen als responsvariabelen (voorbeeld in ter Braak & Juggins 1993).

Een dataset waarvoor deze handleiding bedoeld is, bestaat uit objecten (bijvoorbeeld monsterpunten, vegetatieopnamen of experimentele eenheden) waarvan bekend is:

1. welke soorten¹ voorkomen (dat kan betrekking hebben op alle organismen op die plek, of een selectie, bijvoorbeeld alleen vegetatie),
2. eventueel in welke mate iedere soort voorkomt (abundantie) en
3. eventueel welke waarden een aantal milieuparameters hebben (fysisch, chemisch, hydrologisch, etc.)².

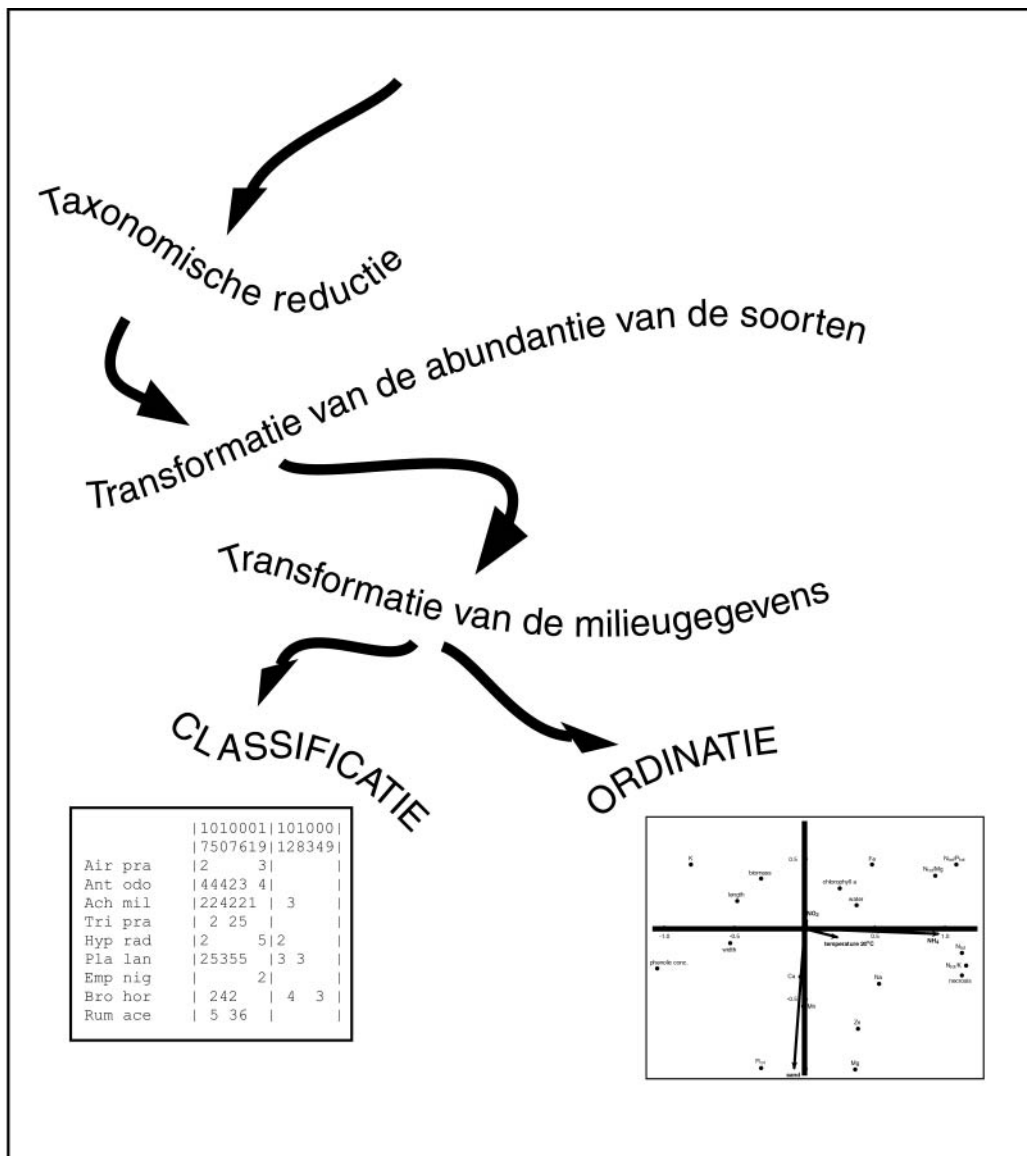
Het resultaat na multivariate analyse is:

1. Een indeling van de gegevens in een aantal groepen (clusters) die gekenmerkt worden door het voorkomen van bepaalde soorten, (bijvoorbeeld in een overzichtelijke tabel).
2. Inzichten in de milieuvariabelen die gecorreleerd zijn met het voorkomen van deze soorten en/of clusters.
3. een overzichtelijke weergave van soorten, clusters en/of milieuvariabelen, bijvoorbeeld in een plaatje (ordinatiediagram).

Gehanteerde multivariate analysetechnieken zijn classificatie en ordinatie. Classificatie kan worden uitgevoerd m.b.v. de programma's TWINSPAN en FLEXCLUS. In hoofdstuk 3 worden deze technieken nader besproken. Ordinatie m.b.v. het programma CANOCO komt aan de orde in het vierde hoofdstuk. Er worden 6 verschillende ordinatietechnieken besproken die alle uitgevoerd kunnen worden met CANOCO. Deze technieken kunnen worden onderverdeeld in 2 typen: indirecte en directe gradiëntanalyse. In het tweede hoofdstuk wordt ingegaan op de voorbereiding van de gegevens. Figuur 1 geeft een overzicht van de achtereenvolgende bewerkingen.

¹ Soorten zijn dan de responsvariabelen. Zoals gezegd kunnen ook eigenschappen als responsvariabelen worden beschouwd, alsmede milieuvariabelen die reageren op bijvoorbeeld behandelingen.

² Verklarende variabelen.



Figuur 1. Overzicht van de bewerkingen die in deze handleiding worden behandeld.

2. VOORBEWERKING VAN DE GEGEVENS

2.1. *Taxonomische reductie*

De meeste multivariate technieken voor gegevensverwerking zijn gebaseerd op een lineair of een eentoppig (gaussisch) responsie model van de taxa met betrekking tot de milieugradiënten. Een dergelijke respons komt het meest voor op soortniveau. Als organismen alleen tot op geslachts- of hoger niveau zijn gedetermineerd, moet overwogen worden of men deze taxa weg zal laten uit de analyse. Keuzes met betrekking tot de deelname van taxa aan de analyse moeten op basis van grondige ecologische kennis en studie gemaakt worden (Verdonschot & Schot 1986).

2.2. *Transformatie van de abundantie van de soorten*

Abundantie van soorten (de mate van aanwezigheid) kan in verschillende eenheden worden uitgedrukt. Bij vegetatieonderzoek wordt vaak de Braun-Blanquet schaal, i.e. de hiervan afgeleide schaal van Barkman/van der Maarel gebruikt, zie tabel 1. Een andere veelgebruikte schaal bij vegetatieonderzoek is de Tansley-methode, zie tabel 2. Een derde mogelijkheid is het inschatten of meten van de bedekking (percentages). Bedekkingswaarden kunnen het beste worden getransformeerd naar de natuurlijke logaritme. Hierdoor krijgen de hoogste bedekkingen minder nadruk en volgt de spreiding meer een normale verdeling (Williamson 1972). Tellingen waarbij ook de waarde 0 voorkomt, logtransformeert men daarom via de formule $\ln(\text{telling} + 1)$.

De aanwezigheid van macrofauna en microflora en -fauna wordt meestal uitgedrukt in absolute of relatieve aantallen. Ook hier is een logtransformatie, om dezelfde reden, wenselijk (zie ook Verdonschot & Schot 1986). Vaak worden hier tellingen getransformeerd naar Preston-klassen; dat is $^2\log(\text{telling} + 1)$, naar beneden afgerond naar een geheel getal. De Preston-klasse is hier al op logschaal.

Tabel 1. Oorspronkelijke schaal van Braun-Blanquet (gebaseerd op bedekkingspercentages), schaal van Barkman et al. (1964) (gebaseerd op bedekkingspercentages en frequentie) en de omzetting van deze schaal in de getallen 1 t/m 9 (van der Maarel 1979). Uit Jongman et al. (1995).

| Braun-Blanquet symbool | Bedekking (%) | Barkman symbool | Bedekking (%) of abundantie | Van der Maarel |
|---------------------------|---------------|--------------------|--------------------------------|----------------|
| | | r | Rare | 1 |
| | | + | Few | 2 |
| 1 | <5% | 1 | Many | 3 |
| | | 2m | Abundant | 4 |
| 2 | 5-25% | 2a | 5-12.5% | 5 |
| | | 2b | 12.5-25% | 6 |
| 3 | 25-50% | 3 | | 7 |
| 4 | 50-75% | 4 | | 8 |
| 5 | >75% | 5 | | 9 |

Tabel 2. Schaal volgens de Tansley-methode. loc = lokaal, plaatselijk. Uit Claassen (1987).

| Aanduiding | Omschrijving | Getalsmatige omzetting |
|------------|---------------------------|------------------------|
| r | Een tot enkele exemplaren | 1 |
| r-o | | 2 |
| o | Hier en daar voorkomend | 3 |
| loc f, o-f | | 4 |
| f | Regelmatig voorkomend | 5 |
| loc a, f-a | | 6 |
| a | Veel voorkomend | 7 |
| co-d | | 8 |
| d | Overvloedig, dominant | 9 |

2.3. Transformatie van de milieugegevens

2.3.1. Continue milieuvariabelen (logtransformatie)

Tabel 3. Berekeningen met de log-transformatie

| | |
|--|---|
| Logtransformatie: | $lx = \log(x+1^*)$ |
| Middelen en standaardafwijking berekenen (op de gewone manier) | $mlx = \text{mean}(lx)$ $slx = \text{std}(lx)$ |
| Terugtransformatie : | $tmx = \exp(mlx) - 1^*$ $tsx = \exp(slx)$ |

x= waarde van variabele

lx=loggetransformeerde waarde

mlx=gemiddelde van een aantal loggetransformeerde waarden

slx=standaardafwijking van een aantal loggetransformeerde waarden

tmx=teruggetransformeerd gemiddelde=geometrisch gemiddelde=benadering van de mediaan

tsx=teruggetransformeerde standaardafwijking

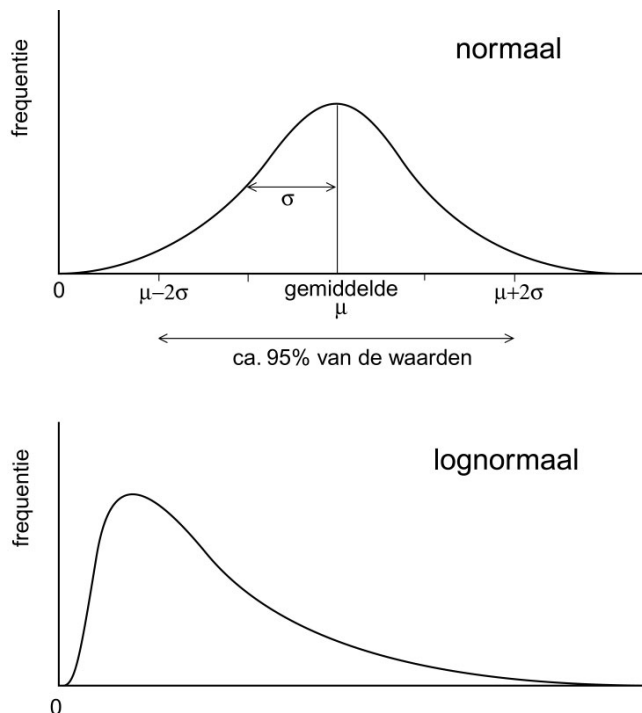
*Kleinste niet-nul-getal, hier 1.

De meeste milieuvariabelen vertonen een verdeling die goed benaderd wordt door een lognormale verdeling (Slob 1987, figuur 2). Om te weten of men inderdaad met een lognormale verdeling te maken heeft, moet men bij grote datasets naar de histogrammen kijken. Voor kleine datasets kan men aannemen dat men met een lognormale verdeling te maken heeft als de standaardafwijking groter is dan het gemiddelde, of als de hoogste waarde meer dan 20 x hoger is dan de laagste waarde (Jongman et al. 1995). Als er sprake is van een lognormale verdeling kunnen de data worden getransformeerd door de natuurlijke logaritme te nemen. De getransformeerde waarden volgen dan een normale verdeling.

Evenals bij de abundanties verdient het aanbeveling om de waarde van de variabelen met een klein getal te verhogen indien 0-waarden voorkomen. Een vuistregel is voor dit kleine getal het kleinste niet-nul getal te kiezen. Voor tellingen (0, 1, 2, 3, ..) verhoogt de waarde dus met 1. Voor concentraties (0, 0.01, 0.02, ...) is het kleinste niet-nul getal in dit voorbeeld 0.01. Bij een variabele als de redoxpotentiaal, waarbij ook negatieve waarden voorkomen, zouden

de waarden kunnen worden verhoogd met een groter getal. Dit uiteraard alleen indien een logtransformatie nodig is!

Als men meerdere waarden voor een (lognormaal verdeelde) milieuvariabele heeft en deze wil middelen, moet eerst de logaritme worden genomen en dan pas worden gemiddeld. Bij interpretatie kan dit gemiddelde worden teruggetransformeerd door de exponent te bepalen. De dan verkregen waarde is het geometrische gemiddelde van de variabele. Dit geometrische gemiddelde is een benadering van de mediaan. Een teruggetransformeerde standaardafwijking berekent men zoals in tabel 3 is aangegeven (Limpert et al. 2001). (Meer hierover kan men lezen in hoofdstuk 2 van Jongman et al. 1995, en in Sokal en Rohlf 1981.)



Figuur 2. Normale en lognormale verdeling (naar Jongman et al. 1995).

2.3.2. Nominale variabelen

Een nominale variabele is een variabele die als waarden geen getallen, maar een bepaalde categorie of klasse aanneemt. Bijvoorbeeld 'geslacht': man - vrouw; of 'landgebruik': hooiland - weiland - akker. Deze moeten voor gegevensverwerking met CANOCO worden opgesplitst in meerdere variabelen, landgebruik wordt bijvoorbeeld 3 variabelen, namelijk hooiland, weiland en akker. Deze kunnen dan ieder de waarden 0 of 1 aannemen (zie handleiding CANOCO).

2.3.3. Standaardisatie

Standaardisatie van de milieuvariabelen houdt in dat ze getransformeerd worden naar een standaardverdeling met bijvoorbeeld gemiddelde = 0 en variantie = 1. Dit is bij de gegevensverwerking van groot belang: anders worden variabelen met bijvoorbeeld waarden tussen de 0 en 0.1 volstrekt onbelangrijk in vergelijking tot variabelen met waarden tussen bijvoorbeeld 0 en 1000. Als milieuvariabelen als verklarende variabelen worden gebruikt hoeft standaardisatie echter niet te worden uitgevoerd omdat dit in de meeste verwerkingsprogramma's

automatisch gebeurt. In CANOCO (technieken PCA, RDA, CA, CCA, DCA en DCCA) is dit het geval. (In FLEXCLUS en TWINSPAN worden, althans bij standaard runs, geen omgevingsvariabelen ingevoerd.). Als milieuvariabelen echter responsvariabelen zijn (dus als 'species' worden ingevoerd) moeten ze wel vaak tevoren worden gestandaardiseerd omdat hun schaal verschilt.

2.3.4. Missing values bij milieuvariabelen

In het programma CANOCO krijgen missing values de waarde 0 in de berekeningen. Dit is uitermate onwenselijk (fosfaatgehalte = 0, pH = 0, etc.; dit zijn misleidende waarden). Daarom is het van belang dat een missing value wordt vervangen door een 'best possible guess'. Men gebruikt hier meestal het gemiddelde voor. Hier wordt bedoeld: het gemiddelde van de waarden die de betreffende variabele in de andere monsterplaatsen aanneemt (ter Braak 1987a).

2.4. Omzetting naar en manipulatie van een 'condensed format'

De basisgegevens moeten worden omgeschreven naar een bestand met 'Cornell condensed format'. In deze vorm kunnen de gegevens worden ingelezen door o.a. TWINSPAN, FLEXCLUS, CANOCO. Hoe het condensed format eruit ziet wordt uitvoerig beschreven in de oude en nieuwe CANOCO-handleiding (resp. ter Braak 1987a en ter Braak & Smilauer 1998 in sectie 4.3). In de windows-versie van CANOCO kunnen excel-bestanden worden gebruikt als invoer voor het programma (W)CanoImp dat voor de omzetting zorgt. M.b.v het programma WINTRAN (<http://www.staff.ncl.ac.uk/stephen.juggins/software/wintran.htm>) is het ook mogelijk om excel- en andere databestanden om te zetten naar cornell condensed format.

3. CLASSIFICATIE

3.1. Inleiding

Met behulp van classificatie worden monsterpunten in groepen ingedeeld. Bij classificatie wordt gestreefd naar een zo groot mogelijke overeenkomst tussen de monsterpunten binnen een groep, en tegelijkertijd ook een zo groot mogelijk verschil tussen de groepen onderling. Gelijkenis en verschil worden vastgesteld aan de hand van de soortensamenstelling. Om het classificeren te vergemakkelijken, en om deze bovendien een objectieve grondslag te geven (hoewel dit discutabel is), zijn computerprogramma's ontwikkeld. In deze handleiding worden twee computerprogramma's besproken die classificaties kunnen uitvoeren: TWINSPAN en FLEXCLUS.

Afhankelijk van doel van het onderzoek en representativiteit van de monsterpunten voor het onderzochte gebied kunnen de clusters worden gebruikt om typen, subtypen of groepen van typen aan te wijzen (of syntaxonomische eenheden bij syntaxonomisch onderzoek; hierbij moeten de typen consistent zijn en in de context van andere syntaxonomische eenheden te worden beschouwd, zie Schaminée et al. 1995). Bij andere doelstellingen worden de clusters bijvoorbeeld alleen gebruikt om de gegevens samen te vatten, eventueel ook voor verdere analyses. Mogelijk worden de clusters ingedeeld bij reeds beschreven typen, syntaxonomische eenheden e.d..

3.2. TWINSPAN

3.2.1. Inleiding

TWINSPAN (Two Way Indicator SPecies ANalysis; Hill 1979) is een divisieve clusteringsmethode waarmee opnamen geïdentificeerd worden en soorten gerangschikt. De methode is gebaseerd op het feit dat een groep opnamen kan worden gekenmerkt door differentiërende soorten. Het programma TWINSPAN stelt eerst differentiërende soorten vast aan de hand van een grove splitsing van de opnamen nadat deze zijn gerangschikt naar de eerste as van CA (Correspondence Analysis, zie hoofdstuk 4). De soorten krijgen een preferentiescore op basis van de mate van voorkeur die ze vertonen voor een van beide 'helften'. Aan de hand hiervan wordt een nieuwe ordening gemaakt en komt de definitieve splitsing tot stand. Dit proces wordt telkens herhaald voor ieder van de splitsingsgroepen.

TWINSPAN is een van de meest gebruikte classificeringsmethoden in de synecologie. Een van de aardige kanten van het programma is dat hoewel het begrip differentiërende soort betrekking heeft op de presentie van een soort, de schrijver van het programma erin geslaagd is om ook de abundantie van de soorten in de berekeningen te betrekken. Iedere soort wordt vervangen door een aantal 'pseudo-species'. Hoe abundantier een soort voorkomt hoe meer 'pseudo-species' er worden gedefinieerd voor die soort. De pseudo-species zijn dan de feitelijke differentiërende soorten (Hill 1979, Jongman et al. 1995).

In voorkomende gevallen kan het programma TWINSPAN met defaultwaarden worden gedraaid, met uitzondering van de cut levels voor de pseudospecies. Een cut level is de abundantiewaarde waarboven een (nieuw) pseudospecimen onderscheiden wordt. Alleen indien de abundantie is uitgedrukt in een (niet log-getransformeerd) percentage kan ook hier

met defaultwaarden volstaan worden. In de TWINSPAN manual (Hill 1979) staan vragen en mogelijke antwoorden op duidelijke wijze beschreven.

3.2.2. Interpretatie van TWINSPAN

Het doel van de TWINSPAN-rangschikking is om de uitspringende, opvallende kenmerken van de gegevensset naar voren te halen. Dit wordt gedaan door gelijkende soorten bijeen te plaatsen en gelijkende opnamen bijeen te plaatsen. Meestal resulteert TWINSPAN in een ruitvormige tabel (een diagonaal van links boven naar rechtsonder). Soorten die weinig overeenkomen met de overige soorten worden op een afwijkende positie geplaatst (bijvoorbeeld onderaan). Gelijkenis tussen soorten heeft hier betrekking op de mate waarin ze in dezelfde opnamen voorkomen. De soortsvolgorde is afhankelijk van de mate waarin de soort beperkt is tot een van de onderscheiden opnamengroepen (Hill 1979).

De splitsingen van de opnamen kunnen worden afgelezen in de nullen en enen die onderaan in de tabel zijn geplaatst.

3.3. FLEXCLUS

FLEXCLUS (FLEXible CLUstering, van Tongeren 1986) is een programma waarmee interactief kan worden geclusterd. Uitgaande van een initiële clustering kan gefuseerd, gesplitst of gerelokeerd worden. Een groot voordeel van het programma is dat men naar eigen inzicht, al of niet gecombineerd met wiskundige maten en technieken, met monsterpunten kan 'schuiven', waarbij de tussentijdse resultaten telkens bekeken kunnen worden. Deze resultaten worden o.a. uitgedrukt in de gelijkenis van de opnamen binnen een cluster, het verschil van ieder cluster met de overige clusters, en hun quotiënt, een maat voor de optimaliteit van ieder cluster (Popma et al. 1983; Van Tongeren, 1986).

Relokatie houdt in dat (de soortensamenstelling van) ieder monsterpunt afzonderlijk vergeleken wordt met ieder cluster. De opnamen die een grotere gelijkenis vertonen met een ander dan hun eigen cluster worden vervolgens naar dat cluster verplaatst. Als maat voor de gelijkenis wordt de similariteitsratio (Wishart 1978) gehanteerd. Na een aantal relokaties (10 zijn meestal genoeg) verandert de clusterindeling nagenoeg niet meer. De clusters zijn dan al of niet 'stabiel'. De clustering is niet stabiel als een monsterpunt, of een groepje van monsterpunten telkens tussen twee clusters heen en weer wordt geschoven. Ten onrechte wordt wel eens door FLEXCLUS-gebruikers gestreefd naar stabiele clusters. Men moet zich realiseren dat ecologische variabelen vaak (deels) continue zijn, waardoor het heel goed kan voorkomen dat sommige monsterpunten precies evenveel gelijkenis vertonen met de ene groep monsterpunten als met een andere. Een clustering die geen stabiele clusters produceert is daarmee feitelijk juist een goede weergave van de werkelijkheid. In paragraaf 3.4 wordt nader ingegaan op de optimalisering van de clustering.

Een initiële clustering kan door FLEXCLUS zelf gemaakt worden, maar het programma kan hiervoor in de plaats ook een bestaande clustering inlezen. Zo'n bestaande clustering zou bijvoorbeeld het resultaat van TWINSPAN (Hill 1979) kunnen zijn³. Voor grotere datasets

³ Daarvoor moet bij het draaien van TWINSPAN positief worden geantwoord als gevraagd wordt om een machine readable copy van de resultaten te maken. Deze machine readable copy is dan namelijk het bestand dat door FLEXCLUS gelezen kan worden

wordt dit laatste aanbevolen.

De initiële clustering die door FLEXCLUS zelf wordt geproduceerd, is, in tegenstelling tot TWINSPAN (waarin een classificatie tot stand komt door splitsingen) in eerste instantie gebaseerd op samenvoegen (agglomeratie). Bij agglomeratieve clustering worden monsterpunten samengevoegd tot groepen van monsterpunten. Samenvoegen (linkage) kan gebeuren met verschillende technieken, bij FLEXCLUS komt de initiële clustering tot stand door single linkage. Deze linkage gaat door tot een (in te voeren) drempel bereikt is.

3.4. Optimaliteit van clustering

Op subjectieve gronden kan gekozen worden voor een bepaalde clustering. Mogelijke criteria zijn:

- Voldoende floristische verschillen om typen te kunnen onderscheiden en deze te kenmerken door (mate van) aanwezigheid van een of meer soorten;
- Een duidelijk statistisch significant verschil tussen de milieuparameters per cluster (zie Jongman et al. 1995, paragraaf 3.3.1);
- Er moeten niet te veel en niet te weinig clusters zijn.

(Deze criteria zijn gemodificeerd en gedeeltelijk afkomstig uit Jongman et al. 1995, p.198).

Op numerieke gronden zijn er twee belangrijke criteria gangbaar (Jongman et al. 1995; Popma et al. 1983):

- De homogeniteit van de clusters (gemiddelde (dis)similariteit van de monsterpunten die tot een cluster behoren),
- De scheiding tussen de clusters (bijvoorbeeld de (dis)similariteit van elk cluster tot het meest erop gelijkende cluster).

Het programma FLEXCLUS geeft deze maten als volgt (per cluster):

homogeneity: de gemiddelde gelijkheid van de 'leden' van het cluster tot de centroid van het cluster,

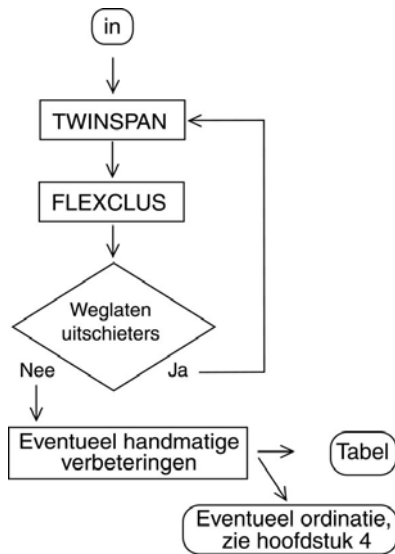
resemblance: de gelijkheid met het meest erop gelijkende cluster.

Hieruit wordt de isolatie van ieder cluster berekend:

isolation: homogeneity/resemblance (waarbij gevallen waarin resemblance=0 worden weggelaten).

3.5. Algemene richtlijn: hoe nu te classificeren?

Aangeraden wordt om eerst een TWINSPAN te draaien, vervolgens FLEXCLUS met de TWINSPAN-resultaten als invoer en dienend als initiële clustering (figuur 3). Binnen FLEXCLUS zou 10x gerelokeerd kunnen worden (als er nog flink wordt geschoven met monsterpunten na 10x relokieren, dan nog eens 10x relokieren), waarna de resultaten grondig bekeken moeten worden. (FLEXCLUS geeft een schema met daarin homogeniteiten en resemblances van ieder cluster). Als de clusters of sommige van de clusters te klein of te groot zijn; te slecht of te goed geïsoleerd ('isolation' zie boven) etc., dan kan men nog wat fuseren of splitsen. Op deze wijze wordt een aardig inzicht verkregen in de structuur van de data.



Figuur 3. Classificatie

Gevolgen van inzicht in de structuur van de data kunnen zijn:

1. Hoogstwaarschijnlijk zijn er een aantal outliers (uitschieters) of groepjes van outliers. Men moet deze nader bestuderen en beslissen of ze (en welke) uit de verdere analyse weggelaten zullen worden. Men kan de gevolgen hiervan voor de clustering alvast enigszins bekijken (en daarmee een inzicht verkrijgen over het belang van weglaten) door dit in FLEXCLUS aan te geven (optie "removal of outliers").
2. Misschien zullen sommige soorten een te grote, en andere juist een te kleine rol spelen in de ontstane clustering. Dat kan dan zowel in FLEXCLUS als in TWINSPAN gemakkelijk worden bijgesteld door een ander gewicht aan de betreffende soorten toe te kennen (default is 1).

Als een of meer sterk afwijkende monsterpunten inderdaad worden weggelaten, of soortgewichten moeten worden bijgesteld, dan wordt TWINSPAN opnieuw gedraaid zonder deze punten, en daarna weer FLEXCLUS. Op dezelfde wijze als bovenbeschreven kan dan een aantal malen gerelokeerd en eventueel gefuseerd of gesplitst etc. worden tot het resultaat naar tevredenheid is. Het inzicht van de ervaren ecoloog is van groot belang tijdens het gehele proces. Vaak worden ook na de numerieke clustering nog handmatige veranderingen in het resultaat aangebracht.

3.6. Analyse van CLUSTERS

Vaak wordt in de verdere analyse van de gegevens niet meer met de individuele monsterpunten gewerkt, maar met de clusters of typen die men onderscheiden heeft. Als maat voor de abundantie der soorten wordt wel de frequentie genomen waarmee iedere soort in het cluster voorkomt (zoals bij een synoptische tabel). In van Katwijk & Roelofs (1988) wordt zowel de frequentie als abundantie der afzonderlijke monsterpunten bij de berekening betrokken. De gemiddelde abundantiescore bij aanwezigheid wordt met de frequentie vermenigvuldigd.

3.7. Lezen

Jongman et al. (1995): Hoofdstuk 6. Cluster analysis (van Tongeren)

Gauch (1982): Hoofdstuk 5. Classification

Schaminée et al. (1995) : Hoofdstuk 10. Numerieke Methoden (Hennekens et al.)

4. ORDINATIE

4.1. Inleiding

De soorten en monsterpunten zijn bij de classificatie ingedeeld in een aantal clusters. Deze worden doorgaans in een tabel geplaatst, en zijn daarmee eigenlijk in een dimensie geplaatst. Het spreekt vanzelf dat deze ene dimensie nooit voldoende kan zijn om de variatie in soortensamenstelling van een divers onderzoeksgebied weer te geven. Met behulp van ordinatie kan men de variatie uitdrukken in meerdere dimensies. Als men de soortensamenstelling van bijvoorbeeld 10 opnamen of vegetatietypen ruimtelijk wil weergeven zijn er feitelijk 10 (!) dimensies nodig. In de praktijk blijkt dat een gering aantal dimensies voldoende is om de belangrijkste variatie in de soortensamenstelling te kunnen treffen. De variatie in soortensamenstelling vertoont immers structuur: sommige opnamen of vegetatietypen lijken op elkaar, er zijn trends.

Ordinatie is een techniek waarmee de dimensies (assen) worden berekend die de grootste variatie in soortensamenstelling treffen. Ordinatieassen kunnen beschouwd worden als latente of hypothetische milieuvariabelen, die zodanig zijn geconstrueerd dat de soorten optimaal passen in een statistisch model dat de soortabundanties langs gradiënten beschrijft (ter Braak 1985). Het is een middel om meerdere soorten gelijktijdig te bestuderen en relaties tussen de soorten en het milieu op te sporen. Op de tweede plaats kan ordinatie helpen om te zien of een belangrijke milieuvariabele over het hoofd is gezien: als er een interpreteerbare ordening van monsters te voorschijn komt in de analyse - eventueel nadat er voor bekende milieuvariabelen is gecorrigeerd (zie covariabelen).

Het resultaat van ordinatie is een rangschikking van soorten, monsterpunten en/of clusters in een laagdimensionale ruimte, zodanig dat gelijkende eenheden bijeen liggen en niet-gelijkende eenheden ver uit elkaar (Gauch 1982). De dimensies worden gedefinieerd door ordinatieassen. Na soortordinatie krijgt iedere soort een score op de soortordinatieassen. Na monsterpuntordinatie krijgt ieder monsterpunt een score op de monsterpuntordinatieassen. Bij de hieronder besproken technieken worden scores voor soorten en monsterpunten gelijktijdig uitgerekend. Dit heeft als voordeel dat de grootste variatie in monsterpunten en de grootste variatie in soorten nooit tot andere ordinatieassen leiden, en samen in één diagram kunnen worden weergegeven. De scores hebben dan bekende verbanden (zie ter Braak & Smilauer 1998, par. 6.3.2)

Besproken zullen worden:

PCA: principale componentenanalyse (principal components analysis)

CA: correspondentieanalyse (correspondence analysis)

DCA: detrended correspondentieanalyse (detrended correspondence analysis)

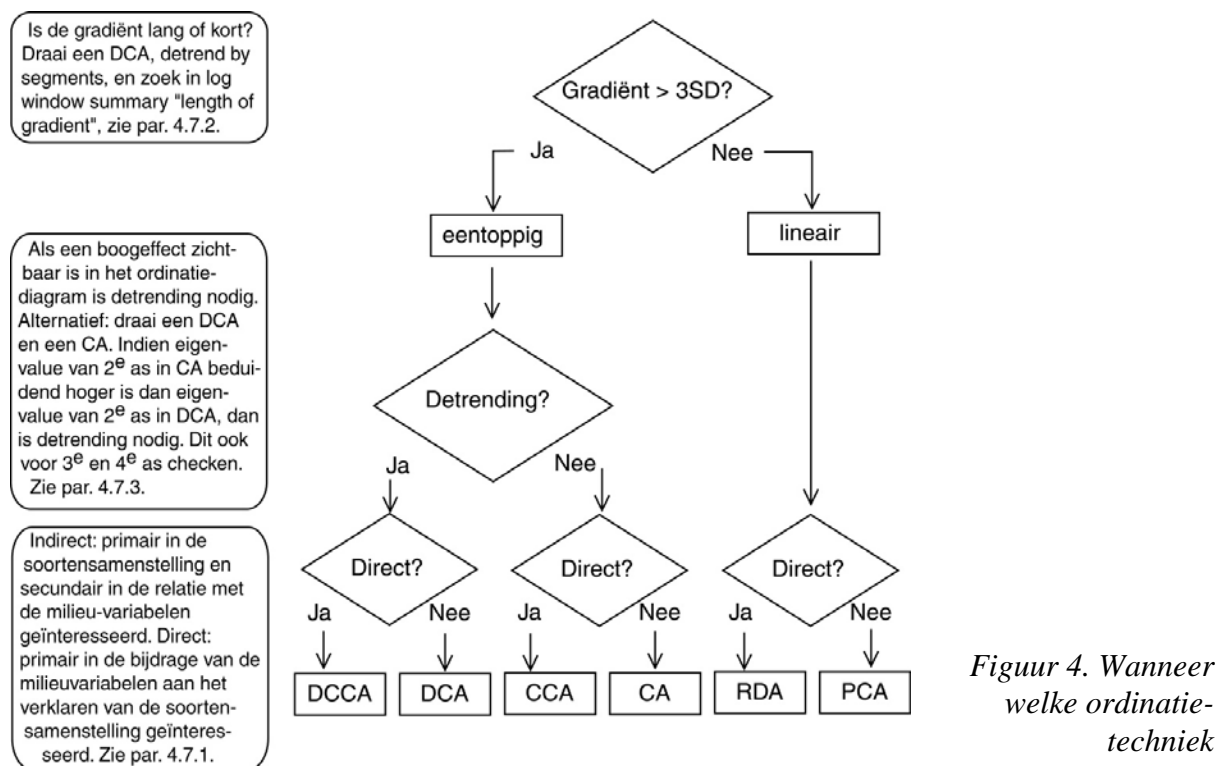
RDA: redundatieanalyse (redundancy analysis)

CCA: canonische correspondentieanalyse (canonical correspondence analysis)

DCCA: detrended correspondentieanalyse (detrended canonical correspondence analysis)

Tabel 3. Overzicht van de overeenkomsten en verschillen van de hier besproken ordinatietechnieken (alle in CANOCO opgenomen).

| | | Indirecte gradiëntanalyse, ofwel ongebonden ordinatie | Directe gradiëntanalyse, ofwel gebonden ordinatie |
|--|-------------------|--|--|
| Lineair responsmodel | | PCA | RDA |
| Gaussisch (eentoppig) responsmodel | Niet detrenden | CA | CCA |
| | Detrenden | DCA | DCCA |



In paragraaf 4.2 t/m 4.4 worden deze technieken nader besproken. Vervolgens wordt ingegaan op de selectie van milieuvariabelen (4.5). Paragraaf 4.6 geeft een leidraad om tot de keuze voor een ordinatietechniek te komen (zie ook figuur 4). Dit hangt af van de onderzoeksvraag, en van de aard van de gegevens. Dan wordt het construeren van samengestelde milieuvariabelen besproken in paragraaf 4.7, de soms optredende verwarring tussen responsvariabelen en verklarende variabelen in 4.8, en tenslotte wordt aangegeven hoe een ordinatiediagram geconstrueerd wordt en hoe dit vervolgens geïnterpreteerd kan worden (paragraaf 4.9).

4.2. Directe versus indirecte gradiëntanalyse technieken

Indirecte (ongebonden) gradiëntanalyse (PCA, CA en DCA) houdt in dat de analyse alleen op basis van de soortensamenstelling plaatsvindt. Milieuvariabelen worden dus niet in de berekening meegenomen, maar, let wel, ze kunnen wel in een tweede stap in het ordinatiediagram worden uitgezet. Bij directe (gebonden) gradiëntanalyse (RDA, CCA en DCCA) worden de milieuvariabelen wél meegenomen in de berekening. Dit wordt hieronder nader uitgelegd.

Kenmerkend voor indirecte gradiëntanalyse (PCA, CA, DCA) is dat de assen in een tweede stap worden geïnterpreteerd naar het milieu. Dit gebeurt op basis van de kennis die men over het gebied heeft, waarbij vaak gebruik wordt gemaakt van grafische technieken (Gauch 1982). Het kan echter evenzeer numeriek gebeuren. Je wilt dan natuurlijk weten in hoeverre iedere milieuvariabele nog wat met die assen te maken heeft. Met andere woorden, wat is de correlatie tussen ieder van de milieuvariabelen en de ordinaatias? Dit komt overeen met het uitvoeren van een multiple regressie op de ordinaatias.

$$Ez_j = b_0 + b_1 * x_1 + b_2 * x_2$$

waarin Ez_j de verwachte waarde van de j -de omgevingsvariabele z_j is; b_1 en b_2 zijn regressiecoëfficiënten en x_1 en x_2 zijn de eerste en tweede ordinaatias.

De regressiecoëfficiënt b_1 is nu een maat voor de correlatie van de j -de milieuvariabele met de eerste as. Analoog is b_2 de coëfficiënt van de tweede as.

Ook zou het interessant zijn om niet alleen te weten wat de correlatie met de ordinaatias voor iedere milieuvariabele afzonderlijk is, maar ook wat hun gecombineerde effect is. Om het gecombineerde effect van de omgevingsvariabelen uit te drukken in correlaties met de ordinaatias moet men een multiple regressie uitvoeren van de ordinaatias op de milieuvariabelen:

voor de eerste ordinaatias:

$$Ex_1 = c_0 + c_1 * z_1 + c_2 * z_2 + .. + c_q * z_q$$

waarin Ex_1 de verwachte waarde van de eerste ordinaatias is; z_j is de j -de (van de q) omgevingsvariabelen en c_j de corresponderende regressiecoëfficiënt is. Deze laatste berekening kan bij indirecte technieken worden uitgevoerd, om te zien welke lineaire combinatie van milieuvariabelen het best overeenkomt met de hypothetische variabele, i.e. de ordinaatias. Dit gebeurt dan na de analyse.

Bij canonische of gebonden ordinaatias (directe technieken) worden de assen 'gedwongen' om zelf deze vorm aan te nemen en een lineaire combinatie van milieuvariabelen te zijn. Multiple regressie wordt dan niet na afloop toegepast, maar tijdens de ordinaatias (in iedere herhalingsstap). De canonische ordinaatias hebben dan ook de vorm:

$$x_i = c_0 + c_1 * z_{1i} + c_2 * z_{2i} + ... + c_q * z_{qi}$$

waarin:

x_i = de waarde van de samengestelde milieuvariabele op monsterpunt i

c_j = het gewicht (kan negatief zijn) van milieuvariabele j

z_{ji} = de waarde van de milieuvariabele j op monsterpunt i

Gewoon een lineaire combinatie van de milieuvariabelen dus. De 'regressiecoëfficiënten' (het

gewicht van de milieuvariabelen, zie bovenstaande formule) van de uiteindelijke canonische ordinaties zijn geen regressiecoëfficiënten in de statistische zin, ze worden daarom aangeduid met "canonische coëfficiënten" (zie p.48 in ter Braak 1987a, of p. 162 in ter Braak & Smilauer 1998).

4.3. Ordinaties op basis van een lineair responsmodel (PCA en RDA)

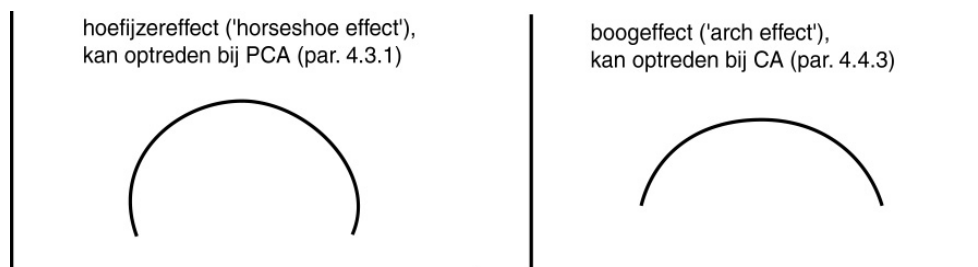
4.3.1. PCA- en RDA-ordinaties op basis van soortensamenstelling

Principale Componenten Analyse is een methode die op een lineair respons model van de soorten is gebaseerd. Dit is een aardige benadering als alleen een beperkt traject van de milieugradiënten wordt bekeken.

De eerste PCA-ordinaties kan men zich voorstellen als een regressielijn in de n-dimensionale ruimte van soorten en/of monsterpunten, waarbij de afstand tussen elk der punten en de lijn zo klein mogelijk is (kleinste kwadraten methode, lineaire regressie). Bij de canonische uitvoering van PCA, namelijk redundancy analysis (RDA), moet de ordinaties tevens een lineaire combinatie zijn van de milieuvariabelen. De selectie van de milieuvariabelen is daarbij van belang.

De tweede en volgende PCA-assen worden op dezelfde wijze berekend als de eerste, waarbij echter een stap is ingebouwd waardoor de as ongecorrleerd wordt gemaakt aan de voorgaande as(sen).

Als PCA wordt toegepast op gegevens van uiteenlopende milieus, treedt het hoefijzereffect op ("horseshoe effect", figuur 5). Dit is begrijpelijk als je bedenkt dat PCA is gebaseerd op een lineair responsmodel. Bijvoorbeeld, in een reeks monsterpunten die een lange gradiënt van voedselarm naar voedselrijk bestrijkt zullen de twee uitersten het verst uit elkaar dienen te liggen in het ordinatiediagram. Echter, een monsterpunt waar soortencombinatie A ontbreekt omdat het te voedselrijk is, wordt over een kam geschoren met een monsterpunt waar soortencombinatie A ontbreekt omdat het te voedselarm is. Beide monsterpunten komen daardoor ten onrechte dicht bij elkaar te liggen (=boogeffect, figuur 5). In technieken die uitgaan van een eentoppig responsmodel gebeurt dit niet (zie paragraaf 4.4). In CANOCO zijn dit de vier CA-technieken CA, DCA, CCA en DCCA.



Figuur 5. Puntenwolken kunnen de vorm van een hoefijzer of de vorm van een boog aannemen. Beide effecten zijn artefacten die voorkomen moeten worden. Om het hoefijzereffect te vermijden dient men een andere ordinatietechniek te kiezen, i.e. CA, DCA, CCA of DCCA, zie 4.3.1. Om het boogeffect te vermijden kiest men voor detrenden, i.e. DCA of DCCA, zie paragraaf 4.4.3.

4.3.2. PCA van milieuvariabelen

PCA kan ook gebruikt worden om de verbanden tussen milieuvariabelen te verhelderen. In plaats van soorten, ordineert men dan milieuvariabelen. Zo kan men zien welke milieuvariabelen een overeenkomstig gedrag vertonen. Men zou vervolgens milieuvariabelen kunnen indelen in groepen en vervolgens de meest representatieve variabele(n) uit elk cluster kunnen selecteren om hiermee verdere analyses uit te voeren. Ook kan men één milieuvariabele samenstellen uit zo'n groep van variabelen. Het beste is om dit te doen aan de hand van de ecologische kennis die men over de samenhang van de variabelen heeft. Het samenstellen van een synthetische milieuvariabele wordt besproken in paragraaf 4.7.

Voorts kan men een PCA van milieuvariabelen uitvoeren om monsterpunten te groeperen op basis van hun milieueigenschappen. Zo kan men een overzicht maken. Net als bij soorten kan men zo proberen te achterhalen wat de onderliggende verklarende variabelen zijn. Voorbeelden van onderliggende verklarende variabelen zijn: het droogvallen van plassen, inundatie door rivierwater, windexpositie van wateren of behandeling in een experiment (maaien, bemesten). Bij een PCA van milieuvariabelen is het van belang om een evenwichtige opbouw van de milieuvariabelen te gebruiken. Als men bijvoorbeeld in een gebied met zowel een zoutgradiënt als een voedingsstoffengradiënt bodemonsters genomen heeft, en hierin Na, Cl, Mg, K, Ca, en NH_4 heeft gemeten, dan zal aan de zoutgradiënt onevenredig veel belang worden toegekend omdat de Na-, Cl-, Mg-, K- en Ca-gehalten alle toenemen in de zoutgradiënt, onderling zeer gecorreleerd zijn en daardoor in feite een 5x zo hoog gewicht krijgen als die ene factor NH_4 die aan de voedingsstoffengradiënt gerelateerd is. In zulke gevallen dient men op basis van ecologisch inzicht een selectie van de milieuvariabelen te maken of kan men de variabelen wegeen.

4.4. Ordinatie op basis van een eentoppig responsmodel (CA, CCA, DCA en DCCA)

Correspondentie Analyse (CA) en daarvan afgeleide technieken vallen onder de categorie Weighted averaging methoden. Dit zijn heuristische methoden. Hieronder worden CA, Canonical CA (CCA), Detrended CA (DCA) en Detrended Canonical CA (DCCA) besproken.

4.4.1. CA (Correspondence Analysis)

CA is synoniem met RA (Reciprokal Averaging) en 'two-way weighted averaging'. De techniek is o.a. beschreven door Hill (1973). Het is een verrassend simpele methode die meerdere malen ontdekt is. Bij CA worden aan de soorten arbitraire scores toegekend. Hieruit worden de monsterpuntscores berekend. De scores van de soorten die in het monsterpunt voorkomen worden dan gemiddeld. Op grond van de monsterpuntscores worden nieuwe soortscores berekend. Dit proces wordt herhaald tot de nieuwe scores nauwelijks meer afwijken van de voorgaande. Ter Braak (1985) heeft aangetoond dat de methode een heuristische benadering is van een ordinatie die op een eentoppig (Gaussisch) responsiemodel is gebaseerd (ter Braak 1985).

Net als bij PCA worden de tweede en volgende assen op dezelfde wijze berekend als de eerste, waarbij een stap is ingebouwd waardoor de as ongecorreleerd wordt gemaakt aan de voorgaande as(sen).

4.4.2. CCA (*Canonical Correspondence Analysis*)

CCA is de canonische versie van CA. Dit houdt in dat in iedere stap in het herhalingsproces zoals bovenbesproken een multiple regressie van de ordinatieassen (d.w.z. de - in de loop van het proces steeds minder arbitraire - scores van soorten en monsterpunten) op de milieuv variabelen wordt uitgevoerd. De zo ontstane ordinatieassen zijn dan een lineaire combinatie van de milieufactoren. De keuze van de milieuv variabelen is daarom van belang. Zie verder paragraaf 4.5.

4.4.3. DCA (*Detrended Correspondence Analysis*)

DCA is een 'verbeterde' vorm van de CA techniek⁴. DCA is oorspronkelijk door Hill & Gauch (1980) ontwikkeld⁵ om een tweetal tekortkomingen van CA te verhelpen met twee heuristisch technieken.

1. Detrending: om het boogeffect ("arch-effect") dat soms bij CA optreedt te verwijderen. Het boogeffect is het verschijnsel dat (kunstmatige) data die in het ordinatiediagram een rechthoek zouden moeten vormen, de vorm van een gebogen strook aannemen. Dit effect is het gevolg van het feit dat de 2e ordinatieas vaak een kwadratische vervorming van de 1e as is. De tweede as is een dubbelgevouwen eerste as (Jongman et al. 1995).
2. Nonlinear rescaling of the axes: om de uiteinden van de ordinatieassen 'op te rekken'. Deze hebben namelijk de neiging om enigszins gecomprimeerd te worden. Dit heeft te maken met het feit dat de soorten die voorkomen in de monsterpunten die aan het einde van de gradiënt liggen in het gegevensmateriaal geen eentoppige (unimodale, gauss-)respons vertonen, maar monotoon of dalend of stijgend zijn.

In CANOCO zijn 4 mogelijkheden om te detrenden. Een van de opties is de oude methode van Hill & Gauch (1980), hierbij wordt de as in segmenten verdeeld, en ieder segment wordt in bepaalde mate opgerekt. Het voordeel van deze methode is dat het oprekken zodanig gebeurt dat de eenheden op de assen gelijk worden aan de SD (standaardafwijking). Op basis van SD's kan men beslissen of men lineair of eentoppig wil ordenen. De andere methodes zijn door ter Braak ontwikkeld, het detrenden met 2^e, 3^e en 4^e orde polynomials (veeltermen), maar blijken geen verbetering op te leveren ten opzichte van de oude methode. Detrenden met polynomials betekent dat de tweede as niet alleen ongecorrleerd moet zijn met eerste as, maar ook met het kwadraat (of 3^e or 4^e macht) van de eerste as. De mogelijkheid 'nonlinear rescaling' is in CANOCO apart opgenomen. Wanneer detrenden en/of rescalen? Dit staat in paragraaf 4.6.3.

4.4.4. DCCA (*Detrended Canonical Correspondence Analysis*)

DCCA voert een ordinatie uit waarbij detrending en nonlinear rescaling (na convergentie) van de assen wordt uitgevoerd, zoals in DCA, en waarbij de ordinatieassen een lineaire combinatie moeten zijn van de milieuv variabelen, zoals in CCA.

⁴ Niet in alle gevallen toepassen, zie paragraaf 4.6.3

⁵ Het wordt toegepast in het nu verouderde programma DECORANA

4.5. Selectie van de milieuvariabelen

4.5.1. Inleiding

Bij directe gradiëntanalysetechnieken (RDA, CCA, DCCA) is het voor de uitkomsten van belang welke milieuvariabelen zijn ingevoerd. Hier is de selectie van de milieuvariabelen dus van groot belang. (Ook bij indirecte technieken moet men niet iedere variabele maar in een plot uitzetten, want dan wordt het wellicht onoverzichtelijk en wordt men verward door overbodige informatie. De selectie van milieuvariabelen heeft bij indirecte technieken echter geen invloed op het ordinatieresultaat zelf.) Soms is een selectie niet nodig omdat men in iedere bekende milieuparameter geïnteresseerd is.

Bij de selectie moet men in eerste instantie weer afgaan op ecologisch inzicht, waarbij m.n. de doelstelling van het onderzoek een rol speelt. Men kan zich echter wel laten helpen door ordinatietechnieken en statistieken die door CANOCO berekend worden.

Een belangrijke overweging om een milieuvariabele niet in de analyse op te nemen is als deze variabele sterk gecorreleerd is aan een andere milieuvariabele. Het verdient dan zelfs aanbeveling om een van beide weg te laten. Mogelijkheden om inzicht te krijgen in de correlaties tussen milieuvariabelen en het belang van iedere milieuvariabele worden besproken in 4.5.2. (t-values en V.I.F.) en 4.4.2. (PCA van milieuvariabelen). Het programma CANOCO heeft een wizard voor 'forward selection of environmental variables', waarmee milieuvariabelen ófwel met hand, ófwel automatisch geselecteerd kunnen worden (ter Braak & Smilauer 1998, paragraaf 5.8).

4.5.2. Het belang van de milieuvariabelen: t-values en V.I.F.

Als men een ordinatie m.b.v. CANOCO uitvoert, wordt in de log-window een 'inflation factor' gegeven. Dit is de V.I.F. (Variance Inflation Factor), een maat voor de correlatie van iedere variabele afzonderlijk met alle andere variabele tegelijk. Als de VIF van een milieuvariabele groter is dan 20 dan is deze variabele vrijwel volledig gecorreleerd met de andere variabelen en heeft daardoor geen eigen bijdrage aan de 'verklaring' van de variatie in soortensamenstelling. (ter Braak 1987a.)

Een andere maat die door CANOCO wordt gegeven zijn de t-waarden (t-val. in de solution file). Als deze minder dan ca. 2.1 bedraagt dan draagt de milieuvariabele niet significant bij aan de regressie (als het aantal monsterpunten minus het aantal milieuvariabelen groter is dan 19). Zo'n milieuvariabele kan worden weggelaten zonder dat er veel aan het ordinatieresultaat verandert. (ter Braak 1987a.)

Er zijn twee risico's verbonden aan het gebruik van V.I.F. en t-waarden:

1. In een vijfde of hogere dimensie (er worden door CANOCO vier assen gegenereerd) is de milieuvariabele misschien wel belangrijk. Dit kan eventueel gecontroleerd worden door een vijfde t/m achtste as te genereren. Men kan dit m.b.v. CANOCO doen door de eerste 4 assen tot covariabele te maken, zie p.20, ter Braak 1987a of ter Braak & Smilauer 1998, p. 205. N.B. In de windows-versie van CANOCO4.x is deze mogelijkheid niet meer aanwezig, wel in de console-versie van CANOCO4.x).
2. Lage t-waarden kunnen veroorzaakt zijn door 'uitdovend' effect van twee gecorreleerde milieuvariabelen. Beide variabelen kunnen dan een zeer lage t-waarde hebben en een hoge V.I.F., maar slechts 1 van beide dient te worden weggelaten (Jongman et al. 1995)

p.55; ter Braak 1986). Daarom is het verstandig per run één variabele met een lage t-waarde (en hoge VIF) weg te laten en dan het resultaat opnieuw te bekijken.

Een voorbeeld: vaak zijn zowel chloride, kalium, natrium, magnesium en calcium gemeten. M.n. chloride en natrium zijn dan vaak zeer sterk gecorreleerd. Het zou dan best zo kunnen zijn dat chloride een V.I.F. van even boven de 20 heeft, terwijl natrium daar net onder zit. Of chloride heeft een wat lagere t-value dan natrium. In zo'n geval kies je er natuurlijk toch voor om natrium weg te laten. Men kan ook een nieuwe milieuvariabele samenstellen, de saliniteit bijvoorbeeld. Meer over het samenstellen van nieuwe milieuvariabelen wordt besproken in paragraaf 4.7. 'Synthetische (samengestelde) milieuparameters.'

4.5.3. Controle van de selectie: eigenvalues vergelijken

De eigenvalue van een as is een getal tussen 0 en 1. Hoe hoger het getal hoe belangrijker de ordinatieas is (meer details zie p.39, ter Braak 1987a). Een eenvoudige manier om te zien of er een belangrijke milieuvariabele ontbreekt is door eerst een ongebonden ordinatie uit te voeren en daarna een gebonden ordinatie met de selectie van de milieuvariabelen. Als nu de eigenvalue van de gebonden ordinatie beduidend minder is dan de eigenvalue van de ongebonden ordinatie dan is er inderdaad een belangrijke variabele over het hoofd gezien. Ga dit na.

4.6. Wanneer welke ordinatietechniek

4.6.1. Direct of indirect

Met behulp van gebonden ordinatie (directe gradiëntanalyse) wordt de variatie in de soortgegevens met 'harde getallen' aan de milieuparameters gecorreleerd, terwijl bij ongebonden ordinatie (indirecte gradiëntanalyse) hypothetische, latente milieuvariabelen worden berekend, waarbij de correlatie met milieuparameters in een tweede stap kan worden berekend. Bij ongebonden ordinatie is men dan ook in eerste instantie gericht op de voornaamste patronen van variatie in de soortensamenstelling. Als de belangrijke milieuvariabelen bekend zijn dan is een gebonden (directe) techniek wellicht effectiever (ter Braak 1987b).

In specifieke gevallen heeft een van beide methoden de voorkeur, maar in de meeste gevallen hebben beide methoden voordelen. Specifieke gevallen:

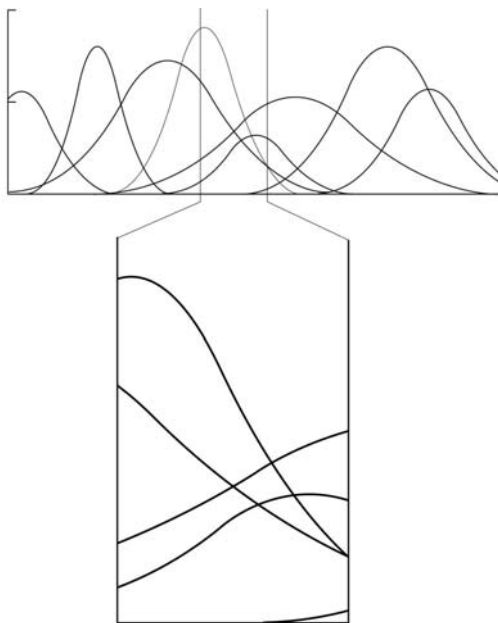
- Directe gradiëntanalyse: als de onderzochte milieuvariabelen slechts zwak zijn gecorreleerd met de eerste paar ongebonden ordinatieassen, maar wel sterk gecorreleerd zijn met de rest van de variatie (dus met de variatie die overblijft na 'aftrek' van de eerste paar assen), dan zouden deze sterke correlaties gemist kunnen worden. Dus, als men het effect van milieuvariabelen wil onderzoeken die (mogelijk) niet direct verband zullen houden met de voornaamste patronen van variatie in de soortensamenstelling, dan moet men directe gradiënt analyse gebruiken.
- Indirecte gradiënt analyse: als men niet de belangrijkste milieuvariabelen gemeten heeft of helemaal geen milieuvariabelen gemeten heeft en/of als men primair geïnteresseerd is in de variatie in soortensamenstelling kan men beter indirecte technieken gebruiken. De assen kunnen dan in een tweede stap gerelateerd worden aan milieuvariabelen die men al of niet gemeten heeft. Met name de indirecte ordinatie techniek wordt dan ook vaak beschouwd als een techniek voor 'hypothesis generation'. (ter Braak 1987b.)

In de meeste gevallen:

Aangeraden wordt om eerst een ongebonden ordinatie uit te voeren om inzicht in de structuur van de data te verkrijgen. Vervolgens kan men met de meest belangrijke en/of interessante milieuv variabelen een gebonden ordinatie uitvoeren, om de relatie tussen deze variabelen en de variatie in soortensamenstelling op directe wijze vast te stellen. De resultaten van de indirecte gradiëntanalyse zijn dan wat globaler, terwijl die der directe gradiënt analyse meer specifiek de relatie met bepaalde milieuv variabelen kan ophelderen. Beide ordinaties kunnen een groter inzicht geven in de gegevens en kunnen naast elkaar gebruikt worden.

4.6.2. Lineair of ca. Gaussisch

We hebben gezien dat de hier besproken ordinatietechnieken zijn gebaseerd op een lineair model (PCA, RDA) of op een benadering van een gaussisch model (CA, CCA, DCA, DCCA). De laatste zijn de 'weighted averaging' methoden. Deze zijn geschikt als een groot aantal soorten hun optimum hebben binnen de dataset. Als de gradiëntlengte minder is dan 3 SD (zie paragraaf 4.9.5. 'eenheden der assen') dan wordt de benadering slechter en bij minder dan 1.5 SD geeft de methode ronduit slechte resultaten, omdat de meeste soorten zich dan (binnen dit bereik, zie figuur 6) monotoon stijgend of dalend gedragen. Dus bij een dataset met grote variatie: weighted averaging methoden (CA-achtigen⁶); bij een smal bereik, weinig variatie: op lineair model gebaseerde methoden (PCA en RDA). Jongman et al. (1995) geven als richtlijn 2 SD (voor de lengte van de gradiënt van de monsterpunten). Om dit vast te kunnen stellen is het handig om een DCA te draaien. Deze berekent namelijk SD's. De lengte van de gradiënt wordt aangeduid in de CANOCO-output onder het hoofdje 'summary'.



Figuur 6. Als de gradiënt lang is, vertonen de meeste soorten een eentoppige (unimodale, gaussische) respons. Als de gradiënt kort is, is de respons monotoon stijgend of dalend. Een lineair model vormt dan de beste benadering (naar Jongman et al. 1995).

4.6.3. Detrending of niet

DCA past een tweetal technieken meer toe dan CA: detrending en nonlinear rescaling (zie paragraaf 4.4.3.). 'Detrending by segments' is de techniek die in het oorspronkelijke DCA,

⁶ N.B. Niet PCA, dit is geen correspondentieanalyse, maar componentenanalyse

namelijk in het programma DECORANA (Hill 1979b), werd uitgevoerd. Deze techniek en/of het nonlinear rescaling hebben als nadeel dat niet alleen het artefact (het boogeffect, zie paragraaf 4.4.3.) wordt weggenomen, maar in sommige gevallen ook ecologisch belangrijke informatie (Minchin 1987, in Jongman et al. 1995). Kenkel & Orloci (in Jongman et al. 1995) vonden dat DCA in sommige gevallen de resultaten van CA vervormt en in elkaar draait. Volgens Jongman et al. (1995) komt dit voor als er weinig soorten zijn per monsterpunt. Men kan controleren of detrending nodig is door de eigenvalues van de tweede en hogere assen te vergelijken. Als de eigenvalue van deze assen bij detrending beduidend lager is dan wanneer geen detrending heeft plaatsgevonden, dan weet men dat de hogere eigenvalue bij niet-detrending is veroorzaakt door een artefact. Detrending is dan nodig.

Bij CCA is detrending vrijwel nooit nodig als er slechts enkele milieuvariabelen in de analyse betrokken zijn. Als een boogeffect optreedt bij CCA is dit geen artefact, maar een indicatie dat er (een) overbodige milieuvariabele(n) zijn opgenomen (ter Braak 1987a).

4.7. Synthetische (samengestelde) milieuparameters

Er zullen hier drie manieren worden besproken om een synthetische milieuparameter samen te stellen. Het beste is het om een synthetische variabele op te bouwen op grond van ecologische kennis. Dit wordt hieronder besproken. Daarna wordt verwezen naar productvariabelen, en als derde wordt het gebruik van een canonische ordinaties als samengestelde milieuparameter besproken. Hierin komen zaken voor die ook elders besproken zijn, maar die voor de duidelijkheid hier herhaald worden.

4.7.1. Naar eigen inzicht

Bijvoorbeeld: de alkaliniteit is een zeer belangrijke parameter, die vaak een goed verband met de variatie in soortensamenstelling vertoont, maar die bij lage pH's geen onderscheid meer maakt. De aciditeit vertoont dat verschijnsel omgekeerd: bij hoge pH's geen onderscheid. Men zou beide parameters als volgt kunnen samenvoegen:

if pH > 6.5 then synthvar=alkaliniteit
else synthvar= -1 * (aciditeit)

Meer voorbeelden kan men vinden in Loucks (1962) en Austin et al. (1984): environmental scalars (gerefereerd in Jongman et al. 1987).

4.7.2. Productvariabelen

Een andere mogelijkheid is het definiëren van productvariabelen (zie ter Braak 1987a, p. 26 of ter Braak & Smilauer 1998 p. 97).

4.7.3. Canonische ordinaties

Het samenstellen van een synthetische variabele uit een selectie van milieuvariabelen die onderling niet sterk gecorreleerd zijn, kan men doen door een canonische ordinatietechniek uit te voeren (bv. RDA, CCA of DCCA) met de geselecteerde milieuvariabelen. De eerste ordinaties is dan de gewenste synthetische variabele. Deze is immers de hypothetische milieuvariabele die is samengesteld uit de optimale combinatie van de milieuvariabelen. Deze heeft de vorm: (zoals in 4.2. ook is beschreven)

$$x_i = c_0 + c_1 * z_{1i} + c_2 * z_{2i} + \dots + c_q * z_{qi}$$

waarin:

x_i = de waarde van de samengestelde milieuvariabele op monsterpunt i

c_j = het gewicht (kan negatief zijn) van milieuvariabele j

z_{ji} = de waarde van de j -de milieuvariabele (van de q) op monsterpunt i

De gewichten c_1 , c_2 etc. zijn de 'canonische coëfficiënten'. Deze worden gegeven in de CANOCO-uitvoer. De milieuv variabelen mogen onderling niet sterk gecorreleerd zijn omdat de canonische coëfficiënten dan instabiel zijn. Een maat voor de onderlinge correlatie is de V.I.F. (Variance Inflation Factor, zie ter Braak 1987a p.40, of ter Braak & Smilauer 1998 p. 119). Deze waarde vindt men in de log-window van CANOCO en mag niet hoger dan 20 zijn.

Men kan controleren of de milieuvariabele eigenlijk wel een zinnige bijdrage levert aan de samengestelde milieuvariabele door te kijken naar de t-values. Dit wordt besproken in paragraaf 4.5.2.

Als een stel milieuv variabelen sterk gecorreleerd zijn, kan men een PCA gebaseerd op een correlatiematrix uitvoeren (i.e. center + standardise by species, ter Braak & Smilauer 1998, p. 91), en de eerste as als synthetische variabele gebruiken, of even zoveel assen als nodig zijn om de variatie in die milieuv variabelen goed te vatten.

4.8. Verklarende variabele of responsvariabele?

Het is niet altijd duidelijk of een variabele een verklarende of een responsvariabele is. Sommige responsvariabelen gedragen zich zelfs tegelijkertijd als verklarende variabele. De vegetatiesamenstelling is bijvoorbeeld een van de belangrijkste bepalende factoren voor de macrofauna- en vissamenstelling in uiterwaardplassen (van den Brink et al. 1994). Als men dit nader wil onderzoeken zou men de soortgegevens als verklarende variabelen kunnen invoeren, als "milieuvariabele". Dit idee is nader uitgewerkt in ter Braak & Schaffers's co-correspondence analysis (2003).

Ook zijn milieuv variabelen, die doorgaans als verklarende variabele worden beschouwd, vaak zelf weer bepaald door onderliggende, grotere en ingrijpende processen. Voorbeelden van dit soort ingrijpende processen zijn (1) het droogvallen van wateren, waardoor bijvoorbeeld het ijzer en fosfaatgehalte in de bodem dramatisch dalen, doordat het aan ijzer gebonden zwavel oxideert, waardoor het vrijkomende ijzer aan fosfaat bindt, etc nog navragen, of (2) inundatie van uiterwaardplassen door rivierwater, waardoor plotseling een enorme input van nutriënten en zouten optreedt (van den Brink et al. 1994). Om de gegevens juist te interpreteren is het van belang om oog te houden voor deze onderliggende processen.

Veelal wordt de keuze bepaald door de onderzoeksvraag (in paleoreconstructies van pH of temperatuur uit soorten, kunnen de milieugegevens als respons worden gedefinieerd en de soorten als verklarend, zoals in WA-PLS, ter Braak & Juggins 1993).

4.9. Constructie van een ordinatiediagram en interpretatie

De ordineringsresultaten worden meestal in 2 dimensies weergegeven. Afhankelijk van deze resultaten kan men plots maken van as1 tegen as2; as1 tegen as3; as1 tegen as4, of as2 tegen as3 etc. In ieder plot kunnen soorten, monsterpunten en milieuvariabelen geplotted worden (als men wil kan dat in 1 diagram). Hiervoor kan men CANODRAW gebruiken vanuit een projectview van CANOCO for windows.

4.9.1. Weergave van soorten en monsterpunten bij PCA en RDA

Soorten worden weergegeven met een pijl, evenals de milieuvariabelen. De monsterpunten worden weergegeven met een punt. De pijl van de soorten wijst in de richting van de maximale variatie in de abundantie van de soort, en de lengte is evenredig met deze maximale mate van variatie. Hoe langer de pijl hoe belangrijker de soort is. (Zie ook p.127 en 128 in Jongman et al. 1995). Coördinaten voor de soorten worden door CANOCO gegeven onder het hoofdje 'species scores'. Coördinaten voor de monsterpuntscores worden gegeven bij 'sample scores'.

4.9.2. Weergave van soorten en monsterpunten in CA, CCA, DCA en DCCA

Bij correspondentieanalyse technieken (CA, CCA, DCA, DCCA) worden soorten weergegeven met een punt. Op dat punt is de kans het grootst dat de soort (met hoge abundantie) aanwezig is. Monsterpunten liggen in het ordinatiediagram op het centroid (gemiddelde) van de punten van de soorten die in dat monsterpunt voorkomen. Zodoende is de kans groot dat monsterpunten die dicht bij een bepaalde soort liggen, ook een hoge abundantie van die soort hebben. Eenvoudig gezegd: soorten en monsterpunten in het diagram geven de variatie in soortensamenstelling van de monsterpunten weer.

Jongman et al. (1995) noemen nog een tweetal aspecten die men in de gaten moeten houden bij de interpretatie van CA en DCA.

- Soorten die aan de randen van het diagram zijn gelegen zijn meestal zeldzame⁷ soorten. Dit kan een gevolg zijn van het feit dat ze een voorkeur hebben voor extreme milieuomstandigheden, maar het kan evengoed een gevolg zijn van het feit dat ze toevallig net voorkomen in monsterpunten met extreme milieuomstandigheden. Deze soorten hebben weinig invloed op de analyse. Het is misschien makkelijker om ze niet te plotten.
- Soorten die in het centrum van het diagram zijn gelegen, kunnen soorten zijn met het optimum in het centrum, maar ze kunnen ook bimodaal zijn of ongecorrleerd aan de ordinaatassen.

4.9.3. Weergave van numerieke milieuvariabelen

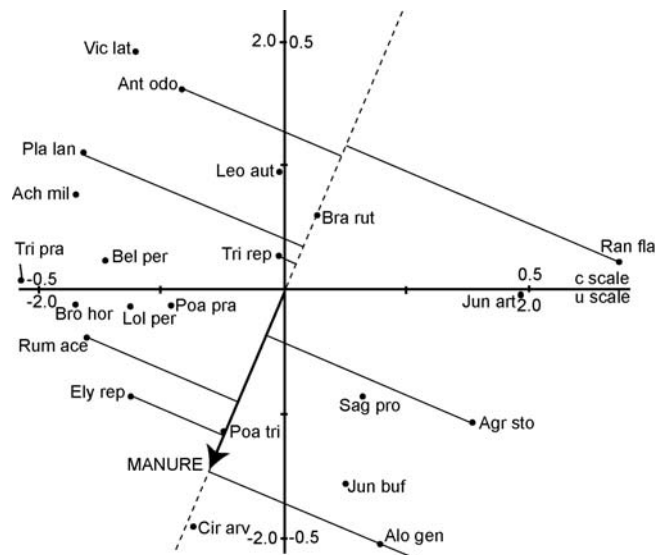
Numerieke milieuvariabelen worden weergegeven met een pijl. De coördinaten van het punt waar de pijl (vanaf de oorsprong) naar wijst kan men in de CANOCO-uitvoer vinden bij 'biplot scores of environmental variables'.

In lineaire methoden (PCA en RDA) geeft een soorten-milieuvariabelen biplot de covariantie weer tussen soorten en milieuvariabelen. In correspondentieanalysemethoden (CA, DCA, CCA, DCCA) geeft het de gewogen gemiddelden met betrekking tot de milieuvariabelen. Milieuvariabelen met een lange pijl zijn de belangrijkste in de analyse; hoe langer de pijl is

⁷ D.i. zeldzaam in de dataset

hoe zekerder men kan zijn over de covariantie of de gewogen gemiddelden (ter Braak 1987a).

Anders gezegd, de pijl wijst in de richting van de maximale verandering van de milieuvvariabele in het diagram, de lengte is evenredig met de mate van verandering in deze richting. Milieuvvariabelen met lange pijlen zijn sterker gecorreleerd met de ordinatieassen, dan die met korte pijlen, en zijn daarom beter gerelateerd aan het patroon van variatie in soortensamenstelling dat in het ordinatiediagram is weergegeven (Jongman et al. 1995).



Figuur 7. Afgeleide volgorde van soorten langs de milieuvvariabele 'manure'. C-scale: schaal van de milieupijlen; U-scale: schaal van de soorten- en opnamepunten. Uit ter Braak (1987b.)

Bij correspondentieanalyse methoden kan de positie van soorten en monsterpunten met betrekking tot de milieuvvariabelen worden afgelezen door een loodlijn op de pijl te trekken, zie figuur 7. (Dit is een voorbeeld van de biplot-rule, ter Braak & Smilauer 1998, p. 37) Het projectiepunt vormt het verwachte gewogen gemiddelde van die variabele voor de betreffende soort of monsterpunt. Een handige regel bij de interpretatie is: het gewogen gemiddelde van de milieuvvariabele voor een soort (of monsterpunt) is naar verwachting hoger dan gemiddeld als de projectie van de soort op de pijl aan dezelfde kant van de oorsprong ligt als de pijl zelf.

4.9.4. Weergave van nominale milieuvvariabelen

Nominale milieuvvariabelen worden in een ordinatiediagram meestal weergegeven met de centroid. Men plot dat als een punt, geen pijl. In CANOCO worden de 'centroids of environmental variables' gegeven. De interpretatie van ligging is net als die van de ligging der monsterpunten.

4.9.5. Eenheden op de assen

Het maken van een ordinatieas komt overeen met het toekennen van een getal (coördinaat) aan de monsterpunten en/of soorten. De as is dus feitelijk niets meer dan de getallen van deze monsterpunten en/of soorten *in volgorde*. De lengte op een as is een maat voor de verandering in de soortensamenstelling der monsterpunten of, omgekeerd, de 'monsterpuntsamenstelling' der soorten. Bij DCA kunnen de eenheden wat preciezer gedefinieerd worden, namelijk in Standaard Deviatie (SD) eenheden. Een gaussische responsiecurve heeft een breedte van ongeveer 4 SD (na rescaling geldt dit voor vrijwel alle soorten in de dataset). Dit wil zeggen dan monsterpunten die 4 SD uit elkaar liggen waarschijnlijk net geen soorten gemeenschap-

pelijk hebben (Gauch 1982; Jongman et al. 1995).

4.10. Lezen

Jongman et al. (1995) hoofdstuk 5: Ordination (ter Braak)

Ter Braak (1987b) hoofdstuk 9: A theory of gradient analysis

Gauch (1982) hoofdstuk 4: Ordination

Jan Lepš & Petr Šmilauer 2003 *Multivariate Analysis of Ecological Data Using CANOCO*
Cambridge University Press, xi + 269 p.

5. LITERATUUR

Claassen, T.H.L., 1987. Typologie en normstelling: een aquatisch-oecologisch onderzoek in Friesland. Dissertatie Katholieke Universiteit Nijmegen, 238 pp.

Gauch, H.G., 1982. Multivariate analysis in community ecology. Cambridge University Press, Cambridge, 298 pp.

Hill, M.O., 1973. Reciprocal averaging: an eigenvector method of ordination. *Journal of Ecology* 61: 237-249.

Hill, M.O., 1979. TWINSpan - a FORTRAN program for arranging multivariate data in an ordered two-way table by classification of the individuals and attributes. Cornell University, Ithaca, N.Y., 31 pp.

Hill, M.O., 1988. How effective is ordination as a means of relating vegetation to ecological factors? In: Hermy, M. & Wilmotte, A. eds. *Multivariate Analysis of Biological Data*. Proc. Young Researchers Meeting, maart 1988, Koninklijke Belgische Botanische Vereniging, Meise: 11-19.

Jongman, R.H.G., ter Braak, C.J.F. & van Tongeren, O.F.R., 1995. *Data analysis in community and landscape ecology*. Cambridge University Press, Cambridge, 299 pp.

Kunst, A.E., Looman, C.W.N. & Mackenbach, J.P., 1990. Socio-economic mortality differences in the Netherlands in 1950-1984: a regional study of cause-specific mortality. *Soc. Sci. Med.*, 31: 141-152

Lepš, J. & Šmilauer, P. 2003 *Multivariate Analysis of Ecological Data Using CANOCO* Cambridge University Press, xi + 269 p.

Limpert E, Stahel WA and Abbt M, 2001 Lognormal distributions across the sciences: keys and clues. *Bioscience* 51: 341-352

Popma, J., Mucina, L., van Tongeren, O. & van der Maarel, E. 1983. On the determination of optimal levels in phytosociological classification. *Vegetatio* 52: 65-76.

Schaminée, J.H.J., Stortelder, A.H.F., Westhoff, V., 1995. *De vegetatie van Nederland. Deel 1. Inleiding tot de plantensociologie - grondslagen, methoden en toepassingen*, Opulus Press Uppsala, 296 pp.

Slob, W., 1987. *Strategies in applying statistics in ecological research*. Dissertatie Vrije Universiteit, Amsterdam, 112 pp.

Sokal, R.R. & Rohlf, F.J., 1981. *Biometry*. 2e druk. Freeman, San Fransisco, 877 pp.

ter Braak, C.J.F., 1985. Correspondence analysis of incidence and abundance data: properties in terms of a unimodal response model. *Biometrics* 41, 859-873.

- ter Braak, C.J.F., 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology* 67: 1167-1179.
- ter Braak, C.J.F., 1987a. CANOCO - a FORTRAN program for canonical community ordination by [partial] [detrended] [canonical] correspondence analysis, principal components analysis and redundancy analysis (version 2.1) ITI-TNO, Wageningen, 95 pp.
- ter Braak, C.J.F., 1987b. Unimodal models to relate species to environment. *Dissertatie Landbouwwuniversiteit Wageningen*, 152 pp.
- ter Braak, C.J.F. & S. Juggins, 1993. Weighted averaging partial least squares regression (WA-PLS): an improved method for reconstructing environmental variables from species assemblages. *Hydrobiologia* 269/270: 485-502
- ter Braak, C.J.F. & Looman, C.W.N. 1994. Biplots in reduced-rank regression. *Biom. J.*, 36: 983-1003
- ter Braak, C.J.F. & Smilauer, P. 1998. CANOCO Reference manual and user's guide tot Canoco for Windows: software for canonical community ordination (version 4). Microcomputer Power (Ithaca, NY, USA), 352 pp.
- ter Braak, C. J. F. & Schaffers, A. P. 2003. Co-correspondence analysis: a new method to relate two species compositions. *Ecology*, in press.
- van den Brink, F.W.B., van Katwijk, M.M., van der Velde G., 1994. Impact of hydrology on phyto- and zooplankton community composition in floodplain lakes along the Lower Rhine and Meuse. *J. Plankton Res.* 16: 351-373.
- van Katwijk, M.M. & J.G.M. Roelofs, 1988. Vegetaties van waterplanten in relatie tot het milieu. *Lab. voor Aquatische Oecologie, K.U. Nijmegen*, pp. 133.
- van Katwijk, M.M., Vergeer, L.H.T., Schmitz, G.H.W., Roelofs, J.G.M. 1997. Ammonium toxicity in eelgrass *Zostera marina*. *Mar. Ecol. Prog. Ser.* 157: 159-173
- van Tongeren, O.F.R., 1986. Flexclus, an interactive program for classification and tabulation of ecological data. *Acta Bot.Neerl.* 35: 137-142.
- Verdonschot, P.F.M. & Schot, J.A., 1987. Macrofaunal community types in helocrene springs. *Jaarverslag 1986 Rijksinstituut voor Natuurbeheer, Arnhem*: 85-103.
- Wishart, D., 1978. CLUSTAN User manual. Program library unit. Edinburgh University Press, Edinburgh, 175 pp.
- Williamson, M., 1972. The analysis of biological populations. Arnold, London, 180 pp.