Exploratieve multivariate analyse van gegevens uit het MHNO/MSPO-project

Ita Kreft
DSWO/Veldwerk

Jan de Leeuw
Datatheorie

Rijksuniversiteit Leiden

DSWO/R-84/7

We bedanken Jacqueline Meulman (Datatheorie) en Steef de Bie (DSWO) voor hun uitvoerige commentaar.

INHOUD

1 INLEIDING BIJ HET ONDERZOEK

1.1 Verantwoording

Dit rapport is ontstaan naar aanleiding van de behoefte binnen (en ook mogelijkerwijs buiten) het projekt MHNO/MSPO aan een nadere uiteenzetting en verklaring van de gebruikte analysemethoden bij een verkenning van de data van dit projekt. Er is daarbij gepoogd de gebruikte technieken, die op zichzelf zowel nogal onbekend als nogal ingewikkeld zijn, zo helder mogelijk uiteen te zetten.
Het aandeel van de tweede auteur van dit rapport heeft betrekking op de achterliggende methodologische vragen die men zich bij het gebruik van bovengenoemde analysemethoden moet stellen. Ze zijn aangepast aan de nieuwste bevindingen op het gebied van deze methoden, en ze hebben in dit rapport min of meer hun nieuwste aanpassing gekregen. Deze aanpassingen zijn nog niet geheel en al verwerkt in het standaardwerk Gifi (1981).

Voor we nu tot het weergeven van de analyses overgaan willen we een kort overzicht geven van het doel en de vraagstellingen van het besproken onderzoek zoals dit momenteel plaats vindt bij de werkgroep LICOR (Leids Interdisciplinair Centrum voor Onderwijsresearch) met medewerking van de DSWO (Dienst Sociaal Wetenschappelijk Onderzoek). De eerste auteur, drs. G.G. Kreft, is werkzaam bij de DSWO. De tweede auteur, Prof. Dr. J. de Leeuw, bij de vakgroep Datatheorie van de Fakulteit Sociale Wetenschappen.

1.2 Het onderwerp van het MHNO/MSPO onderzoek

Het onderwerp van het onderzoek is de evaluatie van resultaten van een herstrukturering in de sektoren van het Middelbaar Beroepsonderwijs (het MBO). In het 'oude' MBO waren tot nu toe 35 verschillende specifiek op één beroep toegesneden dagopleidingen ondergebracht in het Middelbaar Huishoud- en Nijverheidsonderwijs (MHNO) en het Middelbaar Sociaal-Pedagogisch Onderwijs (MSPO). Het wetsontwerp invoering Middelbaar Diensten, Gezondheids- en Nijverheidsonderwijs (MDGNO) heeft tot doel deze 35 smalle opleidingen te bundelen tot

9 à 10 brede opleidingen. Vanaf augustus 1984: het MDGO; Middelbaar Dienstverlenings- en Gezondheidszorg Onderwijs.

Dit streven tot bundeling van de 35 'oude' opleidingen in 9 'nieuwe' opleidingen moet 1 augustus 1984 gerealiseerd gaan worden voor alle MHNO/MSPO-opleidingen. Voor de experimenteerfase, die loopt van augustus 1979 tot augustus 1983, zijn 38 scholen als projektschool aangewezen. Het zijn deze scholen die bedoeld worden als in het onderzoek gesproken wordt van de opleidingen 'nieuwe stijl'. Alle opleidingen 'nieuwe stijl' zijn betrokken bij dit onderzoek.

De opleidingen 'nieuwe stijl' streven o.m. de volgende doelstellingen na (zie ook Van Dijck, 1981, p. 9-21).

1. Voorzieningen scheppen in alle 9 genoemde soorten opleidingen, die horizontale doorstroming tussen de opleidingen mogelijk maken. Dit geeft leerlingen die aanvankelijk verkeerd gekozen hebben de gelegenheid zonder verlies van studietijd over te schakelen op een andere beroepsopleiding.
2. Voorbereiding geven tot een voortgezette studie bij het HBO.
3. Een verruimde toelaatbaarheid, met name t.a.v. LBO-arbiturièn ten.
4. Het stimuleren van de deelname van jongens aan traditionele meisjesopleidingen.
5. Het bevorderen van deelname aan het MBO-onderwijs van leerlingen uit de lagere SES-groepen in de samenleving.
6. Een algemene verhoging van het opleidingspeil van de genoemde MBO-opleidingen.

1.3 De onderzoekspopulatie

De onderzoeksgroep bestaat uit alle leerlingen MHNO/MSPO 'nieuwe stijl' (NS), voor zover deze de school niet voortijdig hebben verlaten en voorzover zij gerespondeerd hebben. In totaal doen er 38 scholen mee aan dit onderzoek. Alle zijn scholen die experimenteren met een nieuwe opleiding. In dit onderzoek hebben we te maken met de gehele populatie NS-onderwijs, en dus niet met een steekproef. Wel zijn de aangewezen NS-scholen deel van een steekproef, op vrijwilligersbasis, van alle MHNO/MSPO-scholen in Nederland, die voortaan 'oude stijl' (OS) worden genoemd in dit verslag. Deze 38 scholen hebben, op vijf

scholen na, slechts één soort opleiding nieuwe stijl in hun gebouw. De vijf scholen die hier een uitzondering op maken hebben twee soorten opleidingen nieuwe stijl. In totaal zijn er 9 soorten opleidingen overgebleven van de ongeveer 35 verschillende OS-opleidingen. We noemen ze hier in de volgorde die ze in ons onderzoek gekregen hebben. De gebruikte afkorting voor deze negen opleidingen komt eerst, erachter staat de verklaring van de afkorting.

1. AB (aktiviteiten begeleiding)
2. AG (tandarts-, dokters-, apothekersassistent; assisterenden in de gezondheidszorg).
3. HT (huishoudtechnische sektor).
4. MK (mode en kleding).
5. SA1 (sociaal agogisch: I arbeids- en personeelswerk en sociale dienstverlening).
6. SA2 (sociaal agogisch: II inrichtings- en kultureel werk).
7. SB (sport en bewegen).
8. VP (verpleegkundige).
9. VZ (verzorgende: bejaarden- en gezinsverzorging, kraamverzorging).

Per instroomjaar gaat het om de volgende aantallen leerlingen.
Cohort 1 en 2: samen 6005 (aantal respondenten 5405).
Cohort 3 : totaal 3984 (aantal respondenten 3705).
Cohort 4 : totaal 4119 (aantal respondenten 3672).
Over de vier cohorten samen betekent dit een totaal van ± 14108 leerlingen bij de instroommeting, d.w.z. de eerste meting. Bij de uitstroommeting, de tweede meting, van de cohorts 1 en 2, na aftrek van de uitvallers, zal het totale leerlingenaantal naar schatting 4000 zijn (2 metingen van ± 2000 leerlingen).

1.4 De onderzoeksinstrumenten

Bij de bestudering van dergelijke grote populaties als we in 1:4 aangaven is het gebruik van een schriftelijke vragenlijst als observatie-instrument een goed middel. Een onderzoek dat gebruik maakt van vragenlijsten wordt in de literatuur aangeduid als een survey-onderzoek. Ook dit onderzoek maakt gebruik van vragenlijsten en wel twee

maal een verschillende vragenlijst, die ook beide meerdere malen worden afgenomen. Op deze verschillende afnametijdstippen gaan we hier niet nader in (zie Kreft, 1982). De twee verschillende vragenlijsten worden gebruikt voor twee verschillende meetmomenten. De eerste voor een afname in het derde en laatste jaar van de opleiding.

De instroomvragenlijst "beoogt een inzicht te verkrijgen in een aantal instroomkenmerken van de eerstejaarsleerlingen. Aan de orde komen onder meer: vooropleiding, sociaal-economisch milieu, woonsituatie, motivatie en beroepsverwachting." De uitstroomvragenlijst "zal worden voorgelegd aan alle derdejaarsleerlingen. Ook nu zal er een kombinatie van postenquete met klassikale afname worden gehanteerd. De verzamelde uitstroomgegevens zullen o.a. een inzicht geven in de verwachtingen ten aanzien van beroepsmogelijkheden, vertikale doorstroming, meningen t.a.v. de opleiding, de verhouding (in tijd) tussen theorie en praktijk en de gang van zaken rond examens en diploma." Beide citaten zijn uit Faddegon en De Rooy (1981).

2 INLEIDING BIJ DATA-ANALYSE

"De keuze van een bepaalde techniek voor analyse van een datamatrix hangt af van de onderzoeksvraagstelling en deze wordt mede bepaald door de aard van de gegevens." (Meerling, 1981, p. 161). In het in dit paper besproken onderzoek is er weinig kennis vooraf ten aanzien van datgene wat men meet. De gebruikte vragenlijsten zijn nieuwe instrumenten, die niet vooraf op validiteit en betrouwbaarheid getoetst zijn. Men heeft tevens geen preciese ideeën over de samenhang en de relaties van de variabelen onderling. Men meet een nieuwe, onbekende situatie, het design is 'observationeel', dat wil zeggen: je kijkt, observeert, en tracht een systeem in de gegevens te vinden of aan te brengen. Zie ook Kreft (1982). Hierdoor gebruikt men in een dergelijk onderzoek dikwijls technieken die de samenhang tussen een groot aantal variabelen kunnen beschrijven. Dit soort technieken kan samengevat worden onder de noemer 'multivariate analyse technieken'. De aard van de samenhang die multivariate analyse technieken onderzoeken wordt ook wel 'de struktuur' van de gegevens genoemd. Voor een

uitgebreide behandeling van dit soort problemen verwijzen we naar Meerling (1981), Gifi (1981), Van de Geer (1971).

Eén van deze multivariate technieken, die dikwijls aanbevolen wordt in 'observationele' situaties, is principale componenten analyse, voortaan afgekort tot PCA. Omdat we in het hier besproken onderzoek te maken hebben met 'kategorische' in plaats van 'numerieke' data, hebben we dus vormen van PCA voor kategorische data nodig. Met name zijn dit de programma's HOMALS, PRIMALS, en PRINCALS, die voor dit doel ontwikkeld zijn door de vakgroep Datatheorie van de RUL. Op deze programma's komen we later in dit paper uitvoerig terug.

Multivariate analyse technieken worden dikwijls beschreven als technieken voor data reduktie. Wat moet men onder data reduktie verstaan? We gaan hier uitvoerig op in, omdat er in de praktijk nog al wat misverstanden op dit punt blijken te bestaan. In een onderzoek als MHNO/MSPO is men gewend van kruistabellen uit te gaan. En dit niet in de laatste plaats omdat opdrachtgevers, meestal zelf geen onderzoekers, daarom vragen. Zelfs als ze daar niet om vragen zijn kruistabellen toch vaak de eenvoudigste bewijsvoeringen voor beweerde verbanden. Dat de met behulp van kruistabellen gevonden verbanden dikwijls meer in het hoofd van de onderzoeker en/of opdrachtgever zitten dan in de data is al meerdere malen in de literatuur aangetoond. We verwijzen naar de diskussie in Gifi (1981, p. 40 e.v.).

Eén van de opvallendste eigenschappen van een kruistabel, in tegenstelling tot een multivariate techniek, is dat een kruistabel geen data reduktie toepast. Men vindt in de kruistabel de exacte aantallen en/of percentages over meestal twee variabelen, samen met bijv. een chi-kwadraat toets. Deze toets geeft bij 5405 respondenten (zoals in dit onderzoek) vrijwel altijd een signifikante afwijking van de nulhypothese van onafhankelijkheid. Dat wil zeggen: welke kruistabel je ook berekent, het verband zal vrijwel altijd signifikant zijn. Ondanks de op het eerste gezicht prettige eigenschap dat een kruistabel geen data-reduktie toepast, dat wil hier zeggen geen informatie verloren laat gaan, schuilt er toch een addertje onder het gras. Bij 37 variabelen (zoals in dit onderzoek in eerste instantie voor de analyse zijn uitgekozen) kan men $\frac{1}{2}n(n-1)$, dus 666 kruistabellen van twee variabelen

uitdraaien. Dit aantal wordt nog zeer vele malen groter als men ook drie- of meerdimensionale tabellen wil bekijken, in totaal zijn er dan 2^{37}, dat wil zeggen meer dan honderd miljard mogelijkheden.

Natuurlijk draait niemand zo veel kruistabellen, om de eenvoudige reden dat daar geen konklusies uit te halen vallen. Men kiest op voorhand welke tabellen men wil hebben. Dit kiezen kan op verschillende gronden gebeuren. Om er enkele te noemen: op grond van de verwachtingen van de opdrachtgever, op grond van de eigen verwachtingen, die natuurlijk ook al impliciet een rol spelen bij het vervaardigen van het instrument, op grond van intuitie of common sense. Het grappige is nu dat je op deze manier ook data reduktie toepast, maar dan nu niet op grond van een 'objektieve' data reduktie techniek, maar op grond van 'subjektieve' en/of 'intuitieve' keuzes. Er is echter nog een bezwaar tegen het gebruik van uitsluitend kruistabellen. Met behulp van kruistabellen vindt men de verdeling van een variabele in samenhang met één of meerdere variabelen. De groepering van een variabele ten opzichte van andere variabelen kan echter steeds wisselen, afhankelijk van de variabele waarmee je kruist. Algemene conclusies zijn daardoor meestal moeilijk te trekken, tenzij je je beperkt tot twee of drie variabelen. Laat ik dit toelichten aan de hand van een analyse, die verderop in dit paper wat uitvoeriger besproken wordt.

In het MHNO/MSPO-projekt zijn er leerlingen uit negen soorten MBO-opleidingen. We onderzoeken of deze negen opleidingen verschillen in termen van vijf achtergrondsvariabelen. Uit PCA-analyses blijkt inderdaad dat er drie typische groepen opleidingen zijn. Als men echter vijf maal een kruistabel analyseert, voor elke achtergrondsvariabele apart, dan vindt men een steeds wisselende groepering van de negen MBO-opleidingen. Zo is de groepering voor sexe anders dan de groepering voor SES en/of leeftijd. Als men vervolgens deze kruistabellen apart, dus los van elkaar, met de uitkomst van de PCA-analyse vergelijkt, dan vindt men dat de PCA-groepering niet noodzakelijk overeenkomt met één van de vijf kruistabel-groeperingen. Deze verschillen zijn te verklaren uit de data reduktie eigenschappen van multivariate technieken, in dit geval dus de data reduktie eigenschappen van PCA. Uit de analyse blijkt dat de vijf achtergrondsvariabelen sterk samenhangen. De onderlinge korrelaties zijn hoog, men kan met relatief

weinig verlies van informatie de vijf achtergrondsvariabelen vervangen door één enkele nieuwe variabele, een gewogen som van de oorspronkelijke vijf. PCA groepeert de opleidingen nu in termen van die nieuwe variabele. In tegenstelling tot de kruistabellen kijken we nu dus naar alle vijf variabelen tegelijk.

In principe zou het dus mogelijk zijn alle 37 variabelen tegelijkertijd in een PCA-analyse te verwerken, waarbij deze 37 'dimensies' dan gereduceerd worden tot bijvoorbeeld twee dimensies (dat wil zeggen: twee nieuwe variabelen, gewogen kombinaties van de 37 oorspronkelijke variabelen, met gewichten gekozen op zo'n manier dat zo weinig mogelijk informatie verloren gaat). Natuurlijk gaat daarbij informatie verloren, namelijk de informatie uit de 35 andere dimensies. Het grote voordeel nu is echter dat we de gegevens teruggebracht hebben tot een weliswaar vereenvoudigde, maar daartegenover beter interpreteerbare en begrijpelijker samenhang. We gaan ervan uit dat de 35 dimensies die we 'weggooien' weinig informatie, althans weinig interpreteerbare informatie, bevatten.

Nogmaals: bij multivariate analyse technieken, en bij PCA in het bijzonder, let men niet op het 'unieke', maar op samenhangen en overeenkomsten van variabelen. Wellicht ten overvloede nog een voorbeeld. Op een schaal samengesteld uit politieke items zullen VVD-ers bij VVD-ers terechtkomen, en daar geheel van gescheiden zal de groep CPN-ers liggen. Hoewel er, zowel bij de VVD-ers als bij de CPN-ers, geen twee personen precies hetzelfde geantwoord hebben op alle vragen, lijken de VVD-ers meer op elkaar dan op de CPN-ers, en vise versa. Of om het in vaktermen te zeggen: de verschillen binnen groepen (de 'within'-variantie) zijn kleiner dan de verschillen tussen groepen (de 'between'-variantie). We kunnen daarom zonder veel verlies van informatie de schaal beschrijven in termen van politieke groepen, een duidelijk voorbeeld van data reduktie dus. Het is zaak 'groot' te kijken bij de interpretatie van resultaten van een multivariate analyse techniek, het gaat om globale effekten.

3 MEER OVER PCA

Met behulp van PCA beschrijft men de belangrijkste verbanden in de data. Het grote aantal oorspronkelijke variabelen wordt teruggebracht tot een meestal aanzienlijk kleiner aantal nieuwe variabelen, die dan de belangrijkste aspecten weergeven die door de oorspronkelijke variabelen gemeten zijn. Men bepaalt zodoende principale componenten, ofwel nieuwe dimensies. De eerste van deze dimensies is de richting waarin de spreiding tussen de punten het grootst is, ofwel de richting die de variantie maximaliseert. Met 'punten' worden in ons geval respondenten bedoeld, iedere respondent wordt namelijk als een punt in de ruimte van de principale componenten afgebeeld. De eerste dimensie wordt bij de door ons gebruikte technieken gevonden op basis van de verschillen in antwoordpatronen van individuen. Als men dan achtereenvolgens kijkt waarom de individuen uit elkaar liggen, op basis van welke antwoorden op welke vragen dit geschiedt, dan kan men deze dimensie interpreteren en daar vervolgens conclusies uit trekken. Deze conclusies kunnen van verschillende aard zijn. Zo kan men concluderen dat een andere analyse met andere of met meer of met mindere variabelen nodig of gewenst is, of men kan inhoudelijke conclusies trekken ten aanzien van de structuur van de data.

De tweede principale component, of ook wel de tweede dimensie, is ook weer een gewogen combinatie van de scores op de te analyseren variabelen. We eisen nu evenwel dat deze nieuwe lineaire combinatie ongecorreleerd is met de vorige lineaire combinatie, de eerste dimensie of eerste principale component. En verder eisen we dat de tweede principale component de grootst mogelijke variantie heeft van al die lineaire combinaties (gewogen optellingen) die gecorreleerd zijn met de eerste component. Bij het zoeken naar de derde en achtereenvolgende dimensies gaat dit evenzo. De derde dimensie heeft een maximale spreiding van de respondenten onder alle gewogen sommen die ongecorreleerd zijn met de eerste en de tweede principale component, en zo verder. Men begrijpt dat de eerste dimensie de meeste variantie 'verklaart', de tweede minder, de derde nog minder, en zo verder. Het is om deze reden vaak voldoende om alleen te letten op de eerste twee dimensies. Verklaren deze niet veel, dan zullen de volgende nog

minder verklaren. Men krijgt bij meer dan twee dimensies bovendien moeilijkheden bij de interpretatie en de naamgeving van de derde en hogere dimensies.

PCA kan men ook formuleren in termen van eigenwaarden van de correlatiematrix, en daarbij behorende eigenvectoren. In de verderop in dit paper gerapporteerde analyses zult u deze term voortdurend tegenkomen. Aan de eigenwaarden kan men zien hoeveel de dimensies aan variantie verklaren. De eerste eigenwaarde is altijd de hoogste, omdat deze behoort bij de eerste dimensie die immers de meeste variantie verklaart. Een correlatiematrix van de orde m (tussen m variabelen) heeft m eigenwaarden, die optellen tot m. Deelt men de eigenwaarde door m, dan krijgt men de variantie die door de betreffende dimensie verklaard wordt, of precieser: men krijgt de proportie van de totale variantie die door de betreffende dimensie verklaard wordt.
Ieder dimensie heeft dus een eigenwaarde, die aangeeft hoe belangrijk die dimensie is. We moeten echter het belang van de eigenwaarde (ofwel van de proportie verklaarde variantie) enigszins afzwakken. De discriminatiematen geven meer gedetailleerde informatie. Iedere variabele heeft op iedere dimensie een discriminatiemaat, de eigenwaarde is de som van de discriminatiematen (of: de proportie verklaarde variantie is het gemiddelde van de discriminatiematen). Een discriminatiemaat, het woord zegt het al, geeft aan hoe goed een variabele discrimineert, ofwel hoe goed een variabele groepen mensen uit elkaar trekt (op een bepaalde dimensie). De discriminatiemaat is het kwadraat van de componentlading (van een variabele op een component), de componentlading is de correlatie tussen de variabele en de component. In het geval van categorische variabelen, waar we hier mee te maken hebben, geeft een hoge discriminatiemaat aan dat de vraag goed groepen respondenten (bijvoorbeeld ja-zeggers, nee-zeggers, en weet-niet-zeggers) onderscheidt, of precieser dat de door de vraag gedefinieerde groepen op de betreffende component goed onderscheiden worden. Een lage discriminatiemaat op alle dimensies geeft aan dat de vraag geen duidelijke relatie vertoont met de andere vragen uit de analyse.
Dit laatste wil natuurlijk niet zeggen dat vragen met een lage discriminatiemaat in een nieuwe analyse, samen met andere variabelen, ook

niet goed zullen discrimineren. Om dat uit te zoeken moet men opnieuw analyses uitvoeren, al dan niet met behulp van een theorie of hypothese, en al dan niet geleid door resultaten van eerdere analyses. Als men in een analyse enkele vragen heeft met een lage discriminatiemaat, dan hebben die vragen geen invloed op de uitslag van de analyse. Met andere woorden: doe je analyse over zonder de 'slechte' variabelen, dan zal men geen wezenlijk andere, dus ook geen betere, oplossing vinden. Waar lage discriminatiematen wel invloed hebben is op de proportie verklaarde variantie. Deze wordt immers kleiner naarmate het aantal variabelen waardoor men de eigenwaarde deelt groter is. Verwijdert men 'slechte' variabelen, dan verandert de eigenwaarde niet veel, maar het aantal waardoor gedeeld wordt wel. Daarmee is dan tegelijk de proportie verklaarde variantie groter geworden. De discriminatiematen zullen, evenals de eigenwaarden, bij de hierna volgende analyses steeds vermeld worden.

4 GEBRUIKTE PROGRAMMA'S, MET VOORBEELDEN

Zoals al in de inleiding werd gezegd, hebben we gebruik gemaakt van PCA-technieken die speciaal geschikt zijn voor gebruik bij niet-numerieke data. De programma's die we gebruiken bij onze exploratie van MHNO/MSPO zijn achtereenvolgens HOMALS, PRIMALS, en PRINCALS.

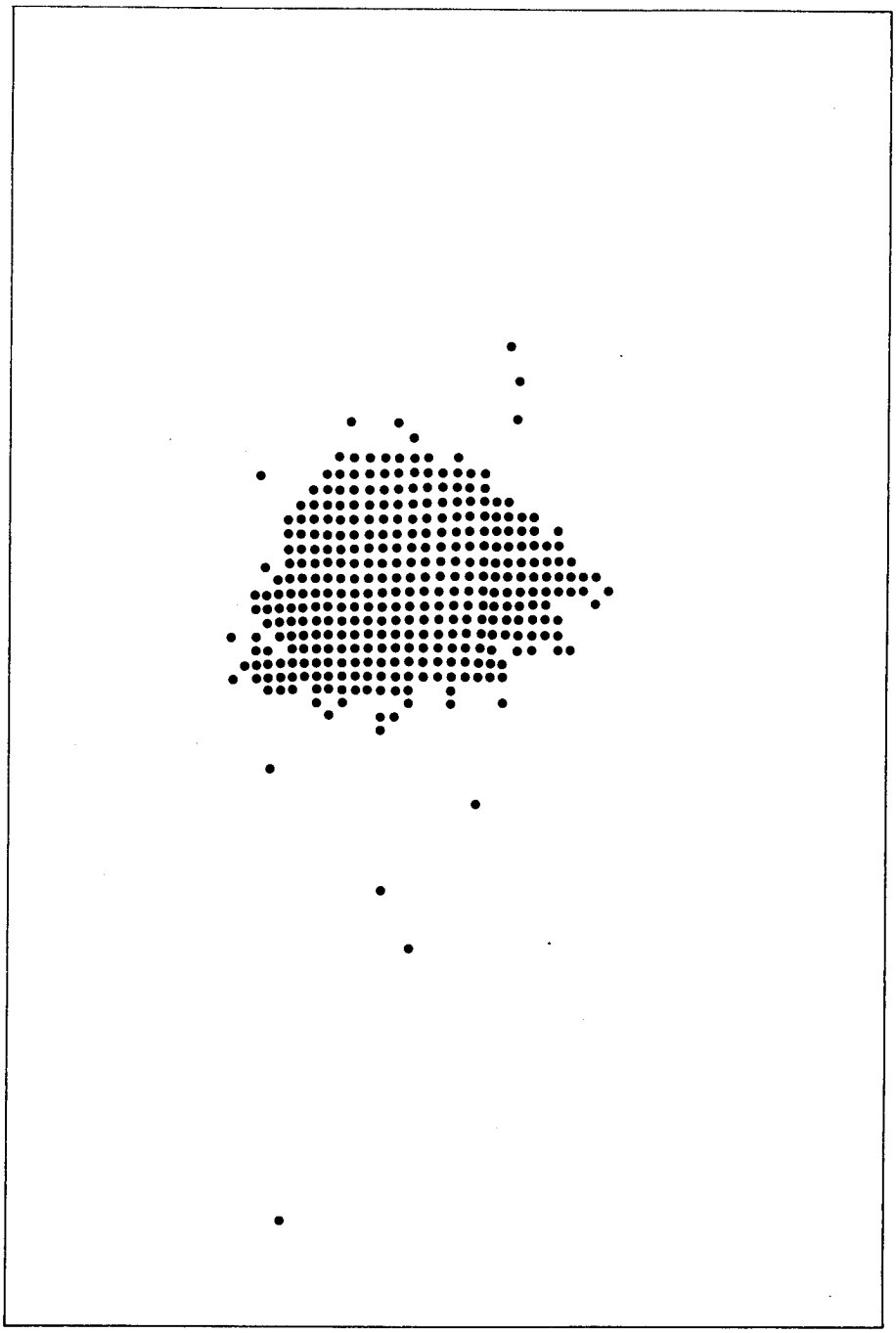
4.1 HOMALS, met drie voorbeelden

De meest bekende en algemeen verbreide techniek van de bovenstaande drie is HOMALS. Als we precieser willen zijn moeten we eigenlijk zeggen dat HOMALS het meest bekende computer-programma is. Het computer-programma HOMALS is de laatste tien jaar door de vakgroep Datatheorie van de RUL ontwikkeld en geperfectioneerd, de data reductie techniek waarop het programma gebaseerd is werd reeds rond 1940 door Guttman geintroduceerd, en wordt onder geheel andere namen als HOMALS over de gehele wereld intensief gebruikt. HOMALS is een afkorting van homogeneity analysis by alternating

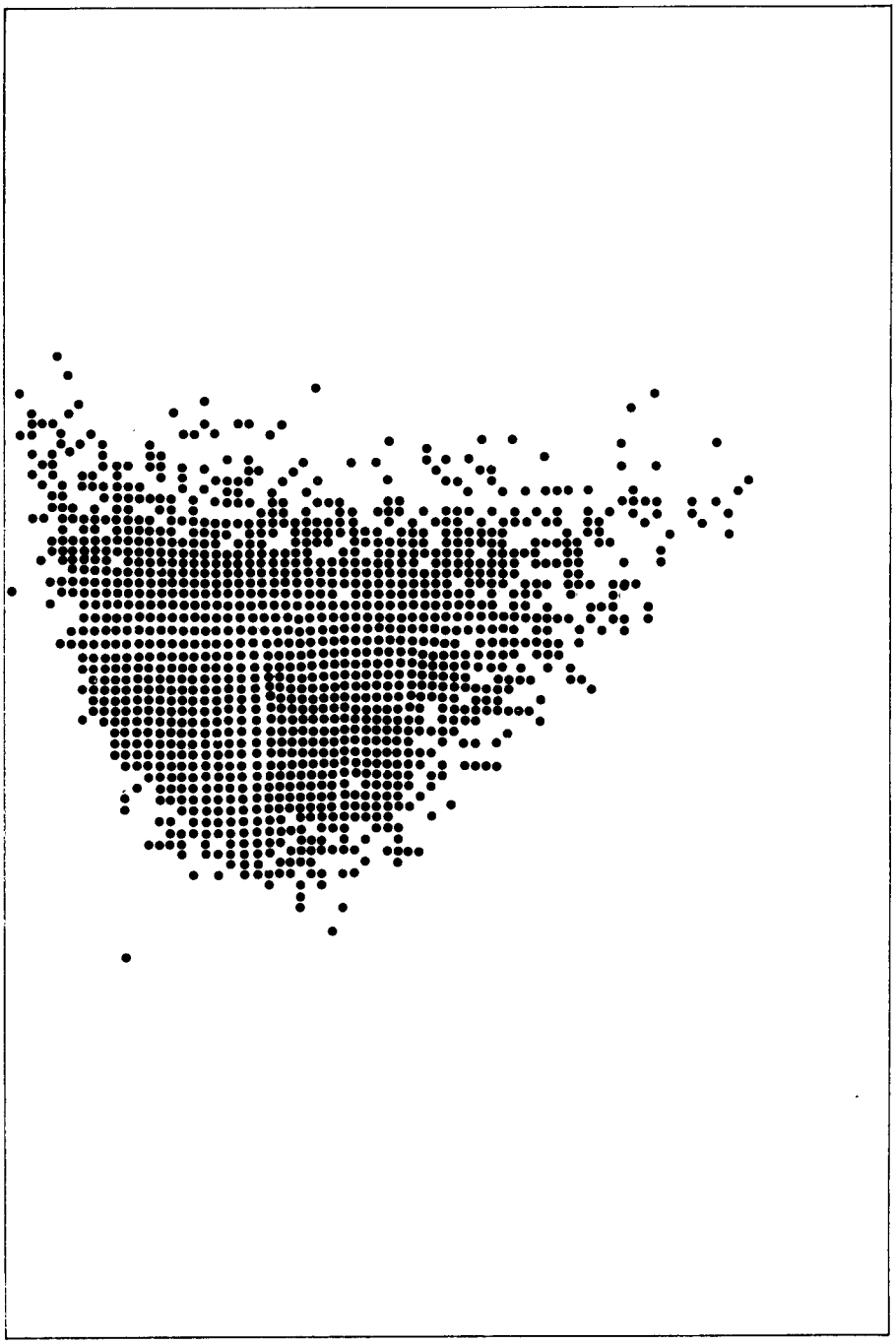
least squares. Het is een globale techniek, die in een grote mate van data reductie toepast. Met deze techniek exploreer je de data. Uit de resultaten van een HOMALS analyse kan je vaak afleiden welke analyses en welke analyse methoden je vervolgens kan toepassen. Vaak ook kan men volstaan met een HOMALS analyse. Op basis van sommige resultaten uit HOMALS, toegepast op MHNO/MSPO, hebben we een aantal groepen van vragen geselecteerd en die vervolgens nader bekeken met PRIMALS en PRINCALS. Om welke reden we dat gedaan hebben wordt bij de betreffende analyses verduidelijkt.

Wat doet HOMALS ? Bij deze techniek wordt voor iedere vragenlijst invuller, hier dus iedere leerling, een plaats gezocht in een (meestal twee-dimensionale) ruimte. Iedere persoon wordt voorgesteld als een punt in een plat vlak. Hierdoor ontstaat een 'puntenwolk'. De personen worden op zo'n manier in het vlak geplaatst, dat zij die antwoordpatronen hebben die het meest op elkaar lijken dicht bij elkaar in een groep worden geplaatst, terwijl personen met antwoordpatronen die weinig met elkaar gemeen hebben ver van elkaar afliggen. Voorbeelden van puntenwolken zijn de figuren 1 en 2 op pagina's 18 en 19. Iedere leerling is een 'punt', en de plaats van de leerling is per analyse vast. De plaats houdt rekening met allerlei variabelen in de analyse, en is dus ook voor alle variabelen hetzelfde. Je kunt, als je wilt, een leerling steeds herkennen aan de plek die hij inneemt. Dit is belangrijk omdat puntenwolken op twee manieren in de output van HOMALS terugkeren. In de eerste plaats als in figuur 1, als plot van 'object scores, unlabeled'. De labels van de punten in de plot geven dan slechts aan hoeveel leerlingen er op dat plaatsje in de plot liggen. En in de tweede plaats als in figuur 2, als plot van 'object scores, labeled by .... (een variabele naam). In deze tweede soort plots zien we dezelfde puntenwolk, die is immers per analyse constant, maar de punten hebben andere labels. We kunnen ieder van de variabelen gebruiken om de labels toe te kennen, een punt in de plot krijgt dan als label het nummer van de categorie waarin de overeenkomstige leerling op die variabele gescoord heeft. In een gelabelde plot kun je ook zien of een vraag in een analyse het goed doet. Als iedereen die ja zegt op een vraag dicht bij elkaar ligt, en ook zo voor de andere antwoorden, dan doet de vraag (of variabele) het goed. In de gelabelde plot liggen dan gelijke labels dicht bij elkaar, of zijn de punten met

FIGUUR 1: HOMALS-Puntenwolk 'Balletje'



FIGUUR 2: HOMALS-Puntenwolk 'Dikke V'



gelijke labels (leerlingen met dezelfde score op die variabele) althans goed van anderen te onderscheiden. De vraag krijgt in dat geval ook een hoge discriminatiemaat, of beter gezegd hoge discriminatiematen op beide dimensies. Slecht is een vraag waarbij alle mogelijke antwoorden in alle groepen voorkomen. De vraag krijgt dan ook lage discriminatiematen.
Je kan dus op twee manieren zien of een vraag het 'goed doet' in een analyse, n.l. door naar de gelabelde plots van de vraag te kijken én door naar de discriminatiematen te kijken. Dit laatste is bij analyses van meer dan 100 personen aan te raden, omdat de plaatjes (zoals figuur 2) bij een (veel) groter aantal personen bijna geheel uit plusjes bestaan, die aangeven dat er op dat plekje in de plot meerdere personen met mogelijkerwijs een verschillend antwoord op de betreffende vraag voorkomen. Hoeveel personen dat zijn, en welke score ze hebben, staat in HOMALS output onder het plaatje. Bij een bestand als van MHNO/MSPO, met 5405 respondenten, is het aantal 'dubbelpunten' zo groot dat de uitleg onder het plaatje enkele pagina's beslaat, waar men verder niets aan heeft. Bovendien kan men aan figuur 2 ook nauwelijks zien of variabele 4 het nu inderdaad goed doet, door de grote hoeveelheid plusjes. In onze analyse hebben we ons daarom beperkt tot het bekijken van de discriminatiematen. Plaatjes hebben we alleen gevraagd van de categorie kwantificaties en van de 'object scores, unlabeled', zoals in figuur 1. Hoe we naar deze twee soorten plaatjes moeten kijken, bespreken we later. Eerst nog iets over discriminatiematen in HOMALS.

In paragraaf 2 van dit paper hebben we gezien dat discriminatiematen in PCA kwadraten zijn van componentladingen, en dat componentladingen correlaties tussen variabelen en principale componenten zijn. De principale componenten tenslotte zijn ongecorreleerde gewogen sommen van de oorspronkelijke variabelen, gekozen op zo'n manier dat ze een zo groot mogelijk percentage van de variantie verklaren. Het percentage van de variantie dat ze verklaren, kunnen we vinden door de overeenkomstige eigenwaarde van de correlatiematrix door het aantal variabelen te delen. HOMALS nu past niet helemaal in dit algemene beeld van PCA. In de eerste plaats moeten we ons realiseren dat onze variabelen categorisch zijn, dat wil zeggen ze nemen geen numerieke waarden aan, en correlaties tussen variabelen en tussen

variabelen en principale componenten zijn niet zonder meer te berekenen. We kunnen correlaties tussen variabelen uitrekenen als we willekeurige namen van de kategorieën van een variable vervangen door getallen; Daartoe moeten we van een variabele 'ja' vervangen door een getal, 'nee' vervangen door een getal, en 'geen mening' of 'weet niet' vervangen door een getal. In de meeste gevallen waarin men PCA op categorische variabelen toepast, vervangt men bijvoorbeeld 'ja' door +1, 'nee' door -1, en 'weet niet' door 0, en vervolgens doet men een gewone PCA. Dit noemt men kwantificeren van categorieën, het is duidelijk dat de zojuist geschetste procedure in grote mate willekeurig is. Als we andere kwantificaties kiezen, dan veranderen de correlaties tussen de variabelen, en daardoor veranderen eigenwaarden en discriminatiematen eveneens. De oplossing van dit dilemma, gebruikt in HOMALS, PRIMALS, en PRINCALS, is optimale kwantificatie, dat wil zeggen we kiezen de kwantificaties op zo'n manier dat ze zo goed mogelijk zijn. Omdat 'zo goed mogelijk' op verschillende manieren gedefinieerd kan worden, zijn er ook verschillende technieken voor categorische (dat wil zeggen kwantificerende) PCA.

In HOMALS (merk op dat we steeds deze naam gebruiken voor de techniek èn voor het programma) kiezen we kwantificaties van de categorieën op zo'n manier dat de som van de covarianties van de variabelen zo groot mogelijk wordt. Omdat covarianties schaalafhankelijk zijn (als men een variabele met twee vermenigvuldigt, dan vermenigvuldigt men daardoor ook zijn covarianties met andere variabelen met twee), moeten we zorgen dat de kwantificaties niet willekeurig groot gekozen kunnen worden. We eisen daarom dat de som van de varianties van de variabelen gelijk is aan een constante, bijvoorbeeld gelijk aan m, het aantal variabelen. Deze definitie van optimale kwantificatie leidt dus tot kwantificatie van de categorieën. Men kan nu op twee manieren verder gaan. In de eerste plaats kan men de kwantificaties gebruiken om een correlatiematrix te construeren, waarop dan vervolgens een gewone PCA gedaan wordt. Dit doen we in PRIMALS (zie verderop). De tweede manier om verder te gaan is een volgend stel kwantificaties te vinden dat weer optimaal is in de zin van maximale som van de covarianties, maar dat aan de voorwaarde voldoet dat het ongecorreleerd is met het eerste stel kwantificaties. Dit doet HOMALS. Dat wil dus zeggen: de tweede HOMALS dimensie bestaat uit

een tweede stel kwantificaties, in vaktermen: HOMALS in meer dimensies geeft multipele kwantificatie van de variabelen. Gebruik van multipele kwantificaties in HOMALS maakt het lastig om HOMALS als een vorm van PCA op te vatten (behalve natuurlijk als we maar één stel kwantificaties, dus één HOMALS dimensie bekijken). Het is daarom beter HOMALS in meer dimensies (gewoonlijk dus twee) te bekijken als een schaaltechniek, dat wil zeggen dat we van het begrip correlatie overstappen op het begrip afstand (Meerling, 1981, geeft een algemene inleiding in meerdimensionale schaaltechnieken; Gifi, 1981, Heiser, 1981, bespreken HOMALS als meerdimensionale schaaltechniek).
We hebben al gezien hoe we het begrip afstand in HOMALS gebruiken. Leerlingen met een overeenkomstig antwoordprofiel liggen dicht bij elkaar, leerlingen die in dezelfde categorie van een variabele scoren, horen in de gelabelde plot van die variabele een duidelijk onderscheiden groep punten te zijn. Er zijn voor iedere variabele binnen-groeps-afstanden (leerlingen in dezelfde categorie van die variabele) en tussen-groeps-afstanden (leerlingen in verschillende categorieën). De schaaltechniek geimplementeerd in HOMALS probeert nu de binnen-groeps-afstanden zo klein mogelijk te maken ten opzichte van de tussen-groeps-afstanden Of: probeert de tussen-groeps-afstanden zo groot mogelijk te maken ten opzichte van alle mogelijke afstanden. HOMALS-eigenwaarden zijn ratio's van tussen-groeps-afstanden ten opzichte van alle afstanden op de desbetreffende dimensie, HOMALS-discriminatiematen zijn vergelijkbare ratio's, maar dan voor iedere variabele apart berekend. HOMALS-categorie kwantificaties zijn zwaartepunten (middelpunten) van alle leerlingen die in de categorie gescoord hebben. Het is verreweg het handigst HOMALS-output te interpreteren aan de hand van dit eenvoudige meetkundige begrippenapparaat.

Elke variabele heeft dus een zelfde aantal categorie kwantificaties als hij categorieën heeft. Hoe verder de categoriegemiddelden van eenzelfde vraag uit elkaar liggen, hoe beter de vraag discrimineert (immers: de tussen-groepen-variantie is hetzelfde als de spreiding van de groepsgemiddelden). Hoe beter de vraag discrimineert, hoe beter hij over het algemeen groepen van leerlingen onderscheidt. Hier is echter een waarschuwing op zijn plaats. HOMALS, evenals de

andere technieken die we later bespreken, doet het het best als de variabelen ongeveer evenveel categorieën hebben, terwijl bovendien de individuen niet al te ongelijk over die categorieën verdeeld moeten zijn. Dit wil zeggen dat er bij een erg scheve verdeling het volgende gebeurt: de 'volle' categorie komt dicht bij het centrum van de plot, de 'lege' categorie komt aan de rand van de plot. "If there is only a very small number of objects in a category then HOMALS can make between-category distances larger by placing the corresponding object points (and thus the category points) near the periphery of the plot. If almost everybody scores in the same category, then that category will be qualified close to the centroid of all object scores, which is the origin of the plot." (Gifi, 1981b, p. 4). Oppassen dus bij de interpretatie van 'scheve' variabelen.
Al eerder merkten we op, dat het niet aan te raden is plaatjes te vragen van object scores gelabeled naar de variabelen als je met erg grote aantallen objecten werkt. Toch is de 'puntenwolk' die gevormd wordt door de objecten (individuen, leerlingen) van belang. Er is een belangrijke globale eigenschap van puntenwolken, waar men op moet letten. Het beste kan men daarvoor de plot van de object scores in ongelabelde vorm nemen. Puntenwolken kunnen namelijk allerlei vormen aannemen: rond, vierkant, V-vormig of hoefijzervormig, of een stelletje ongeregeld. Echt mis is het echter als de punten allemaal op een hoopje liggen, als het ware een klein balletje vormen waarbuiten enkele personen zweven (zie figuur 1, bijvoorbeeld). Deze personen worden uitbijters genoemd. Deze uitbijters hebben een zodanig afwijkend gedragspatroon (in de vragenlijst natuurlijk) dat HOMALS deze personen af gaat zetten tegen de rest van de groep. Net zoals voor extreme anarchisten de rest van de Hollanders één pot nat is, zo is voor HOMALS de rest van de respondenten ook één pot nat. HOMALS gooit hen op één hoop (of 'balletje'), en zet hen af tegen de uitbijters. Wat men dan moet doen is duidelijk: de analyse overdoen zonder de uitbijters (een goed voorbeeld hiervan staat in Gifi, 1981b). Tenzij men natuurlijk juist geinteresseerd is in deze mensen (of objecten).

Een ander patroon dat men kan vinden in de puntenwolk, en waar men in tegenstelling tot bovengenoemd balletje, blij mee is, is de 'dikke V', ook wel het 'hoefijzer' genoemd. In figuur 2 ziet men een

dergelijke wolk. Waarom is men met een dikke V zo blij? Een wolk als deze laat zien dat de tweede HOMALS-dimensie, hoewel ongecorreleerd met de eerste dimensie, daar toch een funktie van is, en wel een kwadratische functie. Dat betekent weer dat de tweede dimensie hetzelfde meet als de eerste. Men kan hieruit concluderen dat de variabelen één grote gemeenschappelijke onderliggende dimensie hebben, die dus ook veel variantie verklaart, en die op verschillende manieren in de HOMALS-dimensies terugkomt: in ieder geval lineair en kwadratisch. De vorm van de puntenwolk bestudeert men dus gewoonlijk via de plot 'object scores, unlabeled'. Als er erg veel individuen zijn kan het echter lastig zijn om het hoefijzer te ontdekken. Men kan dan beter naar de plot van de categoriekwantificaties voor alle variabelen gezamenlijk kijken. Dit zijn vaak veel minder punten, en omdat ze in het zwaartepunt van groepen objectpunten liggen geven ze dikwijls een duidelijker beeld (zie bv. figuur 5) op p. 32.

We geven nu drie voorbeelden van een HOMALS analyse op MHNO/-MSPO-materiaal. Een volledige beschrijving van de gebruikte variabelen staat in de appendix van dit rapport. De bij de analyses behorende plots zijn geen standaard HOMALS output, maar zijn op de VERSATEC plotter gemaakt met gebruikmaking van de HOMALS output.

Eerste HOMALS-voorbeeld

In tabel 1 (p. 25) staan de variabelen uit onze eerste HOMALS-analyse, met hun aantal categorieën, en met de discriminatiematen op de eerste twee HOMALS-dimensies.
Door de vele niet-discriminerende variabelen zijn beide eigenwaarden laag (deze zijn immers het gemiddelde van de discriminatiematen). Deze eigenwaarden worden aanzienlijk hoger wanneer men de zeven niet-discriminerende variabelen verwijdert, voor de overblijvende variabelen heeft dat verder geen gevolg. Uit de discriminatiematen in tabel 1 kan afgeleid worden dat slechts de variabelen 1, 8, 9, 10, 12, en 13 'goede' variabelen zijn, in de zin dat ze goed discrimineren. De variabelen 9, 10, en 13 discrimineren ook goed op de tweede dimensie. Als men de categorie kwantificaties van de variabelen plot, moet men er rekening mee houden dat de categorieën van niet-discriminerende

TABEL 1: eerste HOMALS analyse met categorieën en discriminatiematen van de gebruikte variabelen

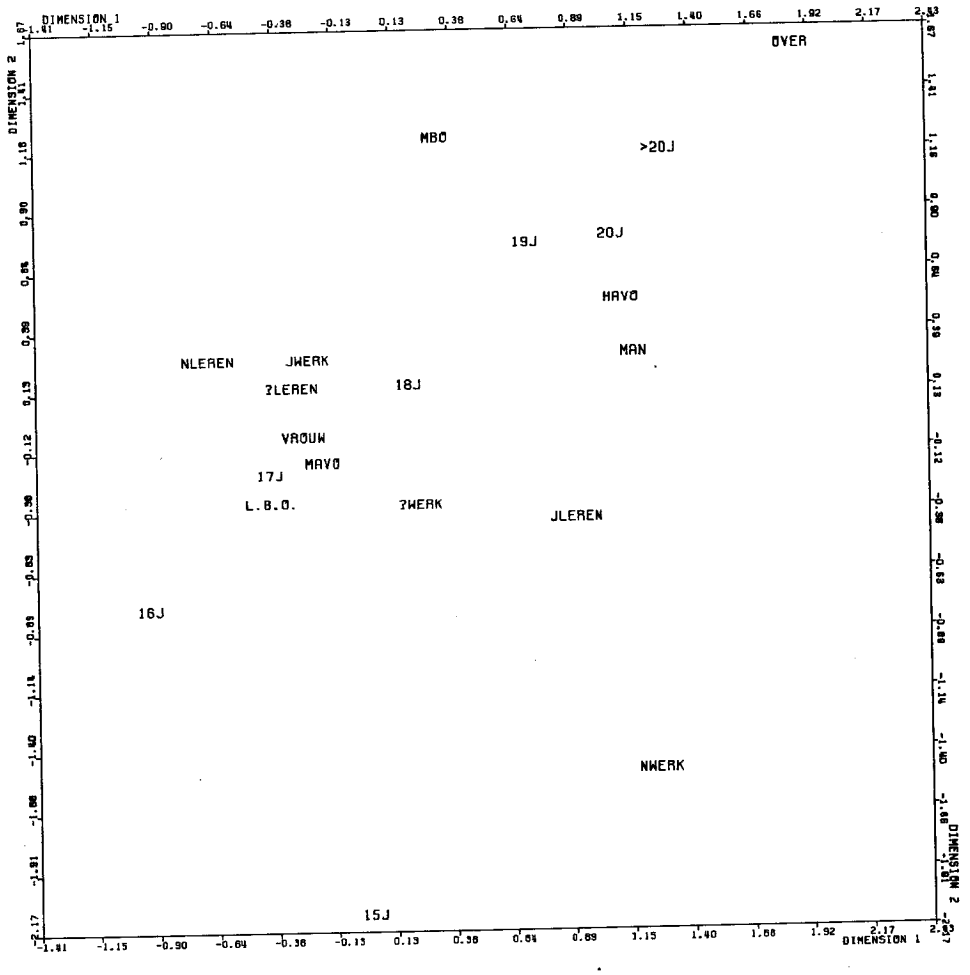
variabele	categorieën	discriminatiematen	
		dimensie 1	dimensie 2
1. motivatie 24:09	5	.260	.108
2. motivatie 24:10	5	.023	.055
3. motivatie 24:21	5	.096	.127
4. al gewerkt? Ja/Nee.	2	.091	.008
5. was de school je eerste keuze? Ja/Nee.	2	.083	.152
6. schoolmelding? Ja/Nee.	2	.000	.030
7. opleidingsmelding? Ja/Nee.	2	.000	.017
8. ga je verder leren? Ja/Weet niet/Nee.	3	.380	.075
9. ga je je beroep uitoefenen? Idem.	3	.157	.223
10. leeftijd van 15 t/m 20, ouder dan 20.	7	.448	.336
11. start van de opleiding. Cohort 1/2.	2	.085	.147
12. sexe	2	.295	.016
13. laatste diploma (vooropleiding).	5	.237	.238
eigenwaarden		.166	.118

variabelen om het centrum van de plot (het nulpunt dus) komen te liggen. We hebben daarom ook de categorieën van de zeven niet-discriminerende variabelen niet geplot in figuur 3 (p. 26).

Uit figuur 3 zijn twee belangrijke conclusies te trekken.

a. Op de eerste dimensie liggen ja-leren, nee-werken, mannen, HAVO-diploma, en 20 jaar en ouder hoog. Laag liggen 15, 16, en 17-jarigen, LBO en MAVO diploma, nee-leren en weet-niet-leren, ja-beroep en weet-niet-beroep. De tendens is dus dat jonge mensen (die meestal vrouw zijn) met een LBO of MAVO diploma niet verder willen leren, maar wel hun beroep zullen gaan uitoefenen. En omgekeerd: oudere mannen met HAVO diploma gaan verder leren.
b. Dat een nadere analyse over de achtergrondsvariabelen gerechtvaardigd lijkt. Immers: leeftijd, vooropleiding, en sexe gedragen zich ongeveer gelijk, hun categorie kwantificaties lopen ongeveer parallel met elkaar als men ze in de juiste volgorde verbindt. Dit doet vermoeden dat de achtergrondsvariabelen een grotendeels ééndimensionale structuur hebben, een geschikt programma om dit

FIGUUR 3



na te gaan is PRIMALS. De PRIMALS analyse over de achtergrondsvariabelen wordt verderop in dit paper besproken.

We merken naar aanleiding van plot 3 nog een aantal dingen op. Er zijn maar 17 leerlingen van 15 jaar. Deze 'lege' categorie komt daardoor aan de rand van de plot. De labels voor ja-leren, nee-leren en weet-niet-leren zijn in de plot jverder, nverder, en ?verder. Het label 'over' bovenaan de plot zijn de overige vooropleidingen (niet LBO, MBO, MAVO, HAVO).

Tweede HOMALS-voorbeeld

Deze analyse is identiek aan de vorige, op één punt na. Variabele 10, leeftijd, is vervangen door MBO-opleidingen. Als men de in tabel 2 opgenomen discriminatiematen vergelijkt met die van de eerste analyse, dan kan men concluderen dat dezelfde variabelen wel en ook dezelfde variabelen niet discrimineren. Ook is de verhouding tussen de 'goede' variabelen (de verhouding tussen discriminatiematen dus) constant gebleven. De getallen zijn echter iets gewijzigd.

Tabel 2: tweede HOMALS-analyse
categorieën en discriminatiematen van de gebruikte variabelen

variabele	categorieën	discriminatiematen	
		dimensie 1	dimensie 2
1. motivatie 24:09	5	.354	.060
2. motivatie 24:10	5	.012	.097
3. motivatie 24:21	5	.040	.056
4. al gewerkt? Ja/Nee.	2	.080	.018
5. was de school je eerste keus? Ja/Nee.	2	.028	.116
6. schoolmelding? Ja/Nee.	2	.000	.067
7. opleidingsmelding? Ja/Nee.	2	.002	.006
8. ga je verder leren? Ja/Weet niet/Nee.	3	.456	.046
9. ga je je beroep uitoefenen? Idem.	3	.145	.259
10. negen MBO-opleidingen.	9	.507	.602
11. start van de opleiding. Cohort 1/2.	2	.003	.040
12. sexe.	2	.435	.070
13. laatste diploma (vooropleiding).	5	.170	.307
eigenwaarden		.172	.134

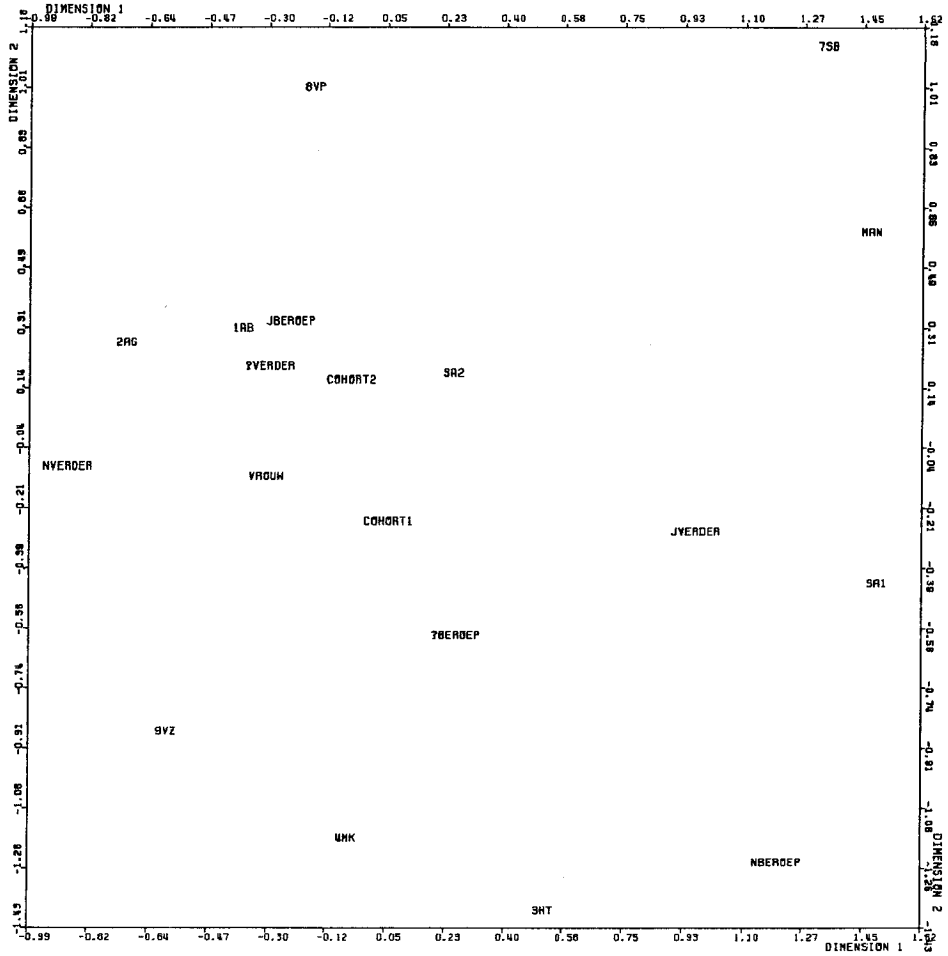
Evenals in de vorige analyse zien we dat motivatie 24:10 en motivatie 24:21 niet discrimineren op de eerste dimensie. Maar 24:21 discrimineert nu ook niet meer op de tweede dimensie, terwijl dat in de vorige analyse nog wel het geval was. Dit kan veroorzaakt zijn door variabele 10, in tabel 1 leeftijd en in tabel 2 MBO-opleidingen. Evenals in de eerste analyse discrimineert variabele 5 beter op de tweede dimensie dan op de eerste, hetzelfde geldt (in beide analyses) voor

motivatie 24:10, beroep uitoefenen, start opleiding, en vooropleiding. Deze groep variabelen meet blijkbaar iets gemeenschappelijks wat niet in de eerste dimensie past. Op deze eerste dimensie ziet men hetzelfde patroon terugkomen als bij de vorige analyse. Sexe heeft echter een aanzienlijk hogere discriminatiemaat. Dit zou erop kunnen wijzen dat sexe sterker samenhangt met variabele 10 uit deze analyse (MBO-opleiding) dan met variabele 10 uit de vorige analyse (leeftijd). We zullen dit nogmaals nagaan met behulp van PRIMALS, wanneer we de achtergrondsvariabelen leeftijd, sexe, vooropleiding, MBO-opleidingen en daarbij ook nog SES (gemeten aan de opleiding van de vader) tezamen analyseren. Merk tenslotte op dat variabelen 4, 6, en 7 in beide analyses niet of nauwelijks discrimineren. We moeten aannemen dat deze variabelen niet met de anderen samenhangen, en nader bekeken moeten worden.

De categorie kwantificaties uit de tweede HOMALS-analyse staan in figuur 4 (p. 29). Het verschil tussen de figuren 3 en 4, voor zover het dezelfde variabelen betreft, is gering.
Als we figuur 3 en figuur 4 met elkaar vergelijken blijkt dat dezelfde verhoudingen tussen de categorieën van de verschillende variabelen weer terug komen, met name voor de twee variabelen waar we het meest in geinteresseerd zijn: beroep uitoefenen en verder leren. In figuur 4 staat ook de niet-discriminerende variabele start (cohort 1 en 2). Men ziet dat de beide categorieën rond het middelpunt liggen. Andere niet-discriminerende variabelen, alsmede variabele 13 (voor-opleiding) zijn voor de duidelijkheid niet geplot. Vooropleiding vindt men wel geplot in figuur 3. Uit figuur 4 zou men kunnen concluderen dat er verschillen bestaan tussen opleidingen met betrekking tot verder leren. SA1 en SB met man en ja-verder en nee-beroep liggen hoog op de eerste dimensie, terwijl AG en VZ met vrouw en nee-verder laag op de eerste dimensie liggen. Beroep discrimineert ook op de tweede dimensie: ja-beroep ligt hoog en nee-beroep ligt laag. Hier kan men uit afleiden dat VP en SB positief staan tegenover hun beroep, hoewel SB ook dicht ligt bij verder leren. HT en MK liggen laag op de tweede dimensie, wat zou kunnen duiden op een dikwijls nee zeggen tegen beroepsuitoefening bij deze beide opleidingen. We benadrukken nogmaals dat het bij HOMALS, zoals bij alle datareductie technieken, gaat om gemiddelde tendenties. Niet alle HAVO-mannen willen verder leren,

niet alle VZ of AG vrouwen willen zo snel mogelijk een beroep gaan uitoefenen. (Voor de afkortingen: zie appendix 1 of paragraaf 1:4).

FIGUUR 4: Categorie kwantificaties tweede HOMALS-voorbeeld



Derde HOMALS-voorbeeld

Als illustratie van wat we eerder beweerd hebben, volgt hier kort een beschrijving van een HOMALS-analyse die van de andere verschilt, omdat we alleen de variabelen opgenomen hebben die het in de eerdere analyses 'goed' deden. We hebben immers betoogd, dat de relatieve belangrijkheid van de variabelen (dat wil zeggen de relatieve discriminatiemaat) niet verandert, als men de 'slechte' variabelen uit de analyse verwijdert. Het enige wat dan wel verandert zijn de eigenwaarden oftewel de gemiddelde discriminatiemaat over alle variabelen. In wezen is de nu volgende analyse overbodig, omdat er nauwelijks nieuwe informatie toegevoegd wordt aan hetgeen we al weten. Zoals men zal constateren blijft de verhouding tussen de variabelen ongeveer gelijk, maar omdat ze nu allemaal goed discrimineren zijn de eigenwaarden hoger. In tabel 3 staan de discriminatiematen uit de drie analyses nogmaals naast elkaar.

TABEL 3: discriminatiematen van drie HOMALS-analyses

variabele	analyse 1		analyse 2		analyse 3	
	dim.1	dim.2	dim.1	dim.2	dim.1	dim.2
1. MBO-opleidingen	----	----	.507	.602	.592	.433
2. leeftijd	.448	.336	----	----	.672	.048
3. laatste diploma	.237	.238	.170	.307	.556	.087
4. eerste keus	.083	.152	.028	.116	.003	.106
5. verder leren	.380	.075	.456	.046	.167	.376
6. beroep uitoefenen	.157	.223	.145	.259	.007	.562
eigenwaarden	.166	.118	.172	.134	.333	.263

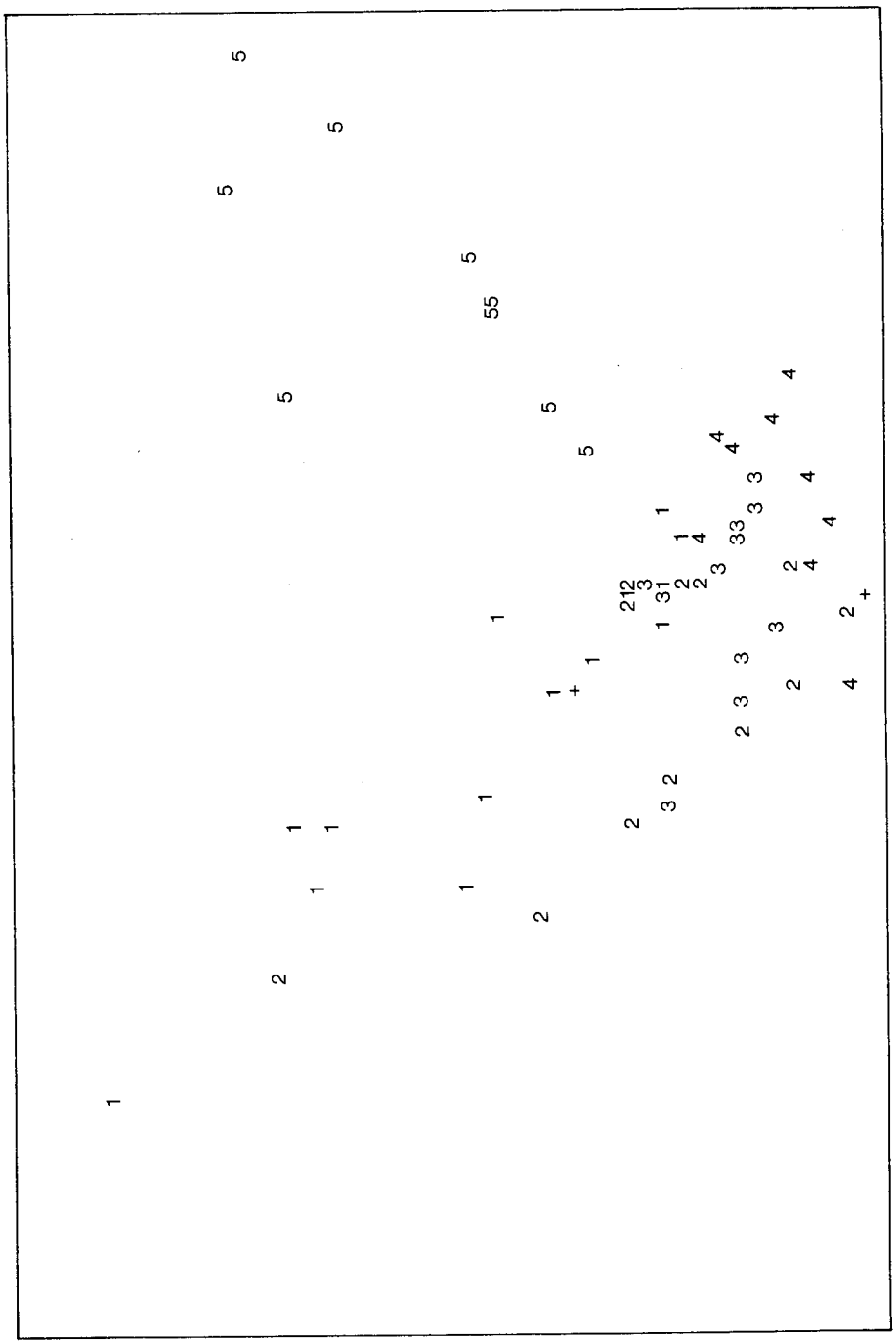
In bovenstaande tabel zijn een aantal verschuivingen te constateren, die het beeld wat duidelijker maken. Zo kan men concluderen dat in analyse 3 variabele 4 (eerste keus) wat dichter bij verder leren en beroep uitoefenen komt te liggen. Deze drie variabelen discrimineren allen op de tweede dimensie. Wat ook in analyse 3 duidelijker naar voren komt, is de scheiding die er is tussen de achtergrondsvariabelen

1 t/m 3 en de overige variabelen 4 t/m 6. In principe is het niet altijd aan te bevelen achtergrondsvariabelen actief in de analyse mee te laten doen. Je krijgt dan al gauw wat er hier gebeurt: de eerste dimensie wordt gedomineerd door de eerste set (de achtergrondvariabelen), de tweede dimensie door de tweede set (de onderzoeksvariabelen). Dit gebeurt echter lang niet altijd, voorwaarde is natuurlijk dat de achtergrondsvariabelen sterk met elkaar samenhangen, sterker tenminste dan de overige variabelen. De reden dat we in deze eerste analyses de achtergrondsvariabelen toch mee hebben laten doen, is dat het hier gaat om een exploratie van de data set. We weten nog niet of en in hoeverre de achtergrondsvariabelen samenhangen. Uit de hierna volgende PRIMALS-analyse zal blijken hoe sterk die samenhang is, dit geeft aanwijzingen om in verdere analyses niet meer dan één achtergrondsvariabele mee te nemen, omdat hooggecorreleerde variabelen de eertse dimensie gaan domineren. Het geeft ook aanwijzingen om de achtergrondsvariabelen passief te behandelen, dat wil zeggen: men doet een HOMALS over de overige variabelen en berekent hierbij de observatiescores, voor iedere leerling één. Op basis van deze observatiescores berekent men categorie kwantificaties van de achtergrondsvariabelen volgens het bekende recept: iedere categorie kwantificaties is het middelpunt van de objectscores van die leerlingen die in die categorie vallen. Hoewel dus de achtergrondsvariabelen niet meedoen in de analyse, worden ze wel in de plaatjes gepresenteerd. Als men dit soort variabelen passief mee wil nemen, moet men de observatiescores uit HOMALS wel wegschrijven naar een tijdelijke data set en vervolgens zelf de desbetreffende gemiddelden berekenen, de tegenwoordige versie van HOMALS kan dit nog niet intern.

4.2 PRIMALS, met twee voorbeelden

We hebben gezien dat HOMALS in een aantal gevallen een hoefijzervormige puntenwolk vindt. Figuur 2 (p. 19) is daar een voorbeeld van, een duidelijker voorbeeld (omdat we daar categorie kwantificaties plotten) is figuur 5 (p. 32). Wat doet men als men een dergelijke V-vormige puntenwolk vindt? Er is dan alle reden om de aandacht te beperken tot de eerste HOMALS dimensie, met bijbehorende categorie kwantificaties, en met bijbehorende correlatiematrix. Tot nu toe hebben

FIGUUR 5: Hoefijzer van categorie-kwalificaties



we HOMALS steeds gebruikt als techniek om de twee-dimensionale afbeeldingen te maken, objectpunten en kategoriekwantificaties worden in het platte vlak geplaatst. In veel situaties echter ligt de keus voor ééndimensionale kwantificatie meer voor de hand. We gebruiken dan HOMALS bijvoorbeeld als een methode voor schaalconstructie. We kunnen a priori besluiten slechts één dimensie te bekijken, bijvoorbeeld omdat we uit een aantal hoog gecorreleerde variabelen een ééndimensionale attitude of achievement-schaal willen construeren, we kunnen ook op basis van een twee-dimensionale HOMALS beslissen om terug te gaan naar één dimensie, en die dimensie wat meer in detail te bekijken. Dit laatste gebeurt wanneer we een V- of hoefijzervormige puntenwolk vinden, en het gebeurt met het programma PRIMALS. In 3.1 hebben we al uitgelegd dat PRIMALS hetzelfde doet als HOMALS in één dimensie, het vindt optimale kwantificaties van de categorieën. Optimaal is gedefinieerd als: maximale som van de co-varianties voor vaste som van de varianties. Of, wat eleganter, PRIMALS vindt kwantificaties van de variabelen op zo'n manier dat de eerste eigenwaarde van de correlatiematrix van de variabelen zo groot mogelijk wordt. PRIMALS doet bovendien nog een aantal extra dingen, en geeft daarom ook andere output dan HOMALS.

Er worden bijvoorbeeld twee correlatiematrixen uitgeschreven: een correlatiematrix vóór herschaling van de variabelen (gebruikt de categorienummers, dat wil zeggen vat de nominale variabelen numeriek op) en een correlatiematrix ná herschaling (na optimale kwantificatie). Als men variabelen heeft met een mooi regelmatig oplopende schaal, dan zal dat blijken uit de herschaling. Immers als de variabelen al redelijk ordinaal zijn, zal de herschaling niet veel te verbeteren hebben. Heeft men echter nominale categorieën dan kunnen de twee correlatiematrices nogal verschillen. Op beide correlatiematrices doet PRIMALS vervolgens een gewone PCA, met eigenwaarden en componentladingen. De eigenwaarden geven aan hoeveel variantie de eerste component 'verklaart', hoeveel de tweede, enzovoorts. Discriminatiematen worden ook geprint, het zijn de kwadraten van de componentladingen. Let op: componentladingen hebben een teken, kunnen positief en negatief zijn, discriminatiematen zijn altijd positief. Of een componentlading positief of negatief is, hangt af van de kwantificatie van de variabele. De door een component 'verklaarde' proportie variantie berekent men weer door de som van de discriminatiematen te

delen door m, of door de eigenwaarde te delen door m. Uit de definitie van PRIMALS volgt dat de correlatiematrix na herschaling een grotere eerste eigenwaarde heeft dan de correlatiematrix voor herschaling, met andere woorden PRIMALS probeert zo te herschalen dat de variabelen zo één-dimensionaal mogelijk zijn.

Het is belangrijk er nogmaals op te wijzen dat het aantal dimensies in HOMALS en in PRIMALS een andere betekenis heeft. In HOMALS is het aantal dimensies gelijk aan het aantal ongecorreleerde kwantificaties, HOMALS is multipel, iedere HOMALS-dimensie kan gebruikt worden om een correlatiematrix te construeren, die dan vervolgens met PCA in m (aantal variabelen) dimensies ontbonden kan worden. PRIMALS gebruikt maar één kwantificatie, dezelfde als ééndimensionale HOMALS, en ontbindt de geconstrueerde correlatiematrix met PCA in een aantal dimensies. Dimensionaliteit in PRIMALS is de dimensionaliteit van de geïnduceerde (optimale) correlatiematrix. Het is daarom onzin de eerste twee HOMALS eigenwaarden te vergelijken met de eerste twee PRIMALS eigenwaarden. De eerste twee PRIMALS eigenwaarden behoren bij één en dezelfde correlatiematrix, de eerste twee HOMALS eigenwaarden behoren bij verschillende correlatiematrices.

Eerste PRIMALS voorbeeld

Figuur 5 (p. 32) komt uit een HOMALS-analyse van de 21 motivatie variabelen. De categorie kwantificaties laten duidelijk een V-vorm zien. Zoals we eerder gezien hebben is dit een aanwijzing voor een onderliggende één-dimensionale structuur. We bekijken daarom de eerste HOMALS-dimensie wat meer in detail met behulp van PRIMALS. De uitkomsten zijn wat teleurstellend. In de eerste plaats zijn alle 21 variabelen na kwantificatie positief gecorreleerd. Dit is opmerkelijk, omdat er in de set motivatie variabelen duidelijk aan elkaar tegengestelde items zijn. Dit laatste geeft de verwachting, dat sommige items negatief gecorreleerd of tenminste ongecorreleerd zullen zijn. Als dan bovendien nog blijkt dat de hoog met elkaar correlerende items op elkaar lijken wat de scheefheid, dus wat de marginalen betreft, dan heeft men alle grond aan te nemen dat de eerste dimensie iets anders meet dan 'motivatie'. De component-ladingen wezen uit dat items met linksscheve verdelingen dezelfde ladingen hebben, de rechts-scheve ver-

delingen kregen ook op elkaar lijkende ladingen, enzovoort. De conclusie tot nu toe luidt daarom: de eerste dimensie in HOMALS en dus ook PRIMALS meet antwoordtendentie, ofwel aard en mate van scheefheid. Dit illustreert dat HOMALS en PRIMALS met een zekere voorzichtigheid gebruikt moeten worden wanneer variabelen zeer scheef zijn, en aan de andere kant dat de 21 motivatievariabelen zich vooral onderscheiden door hun aard en mate van scheefheid, een nogal teleurstellende conclusie.

Tweede PRIMALS voorbeeld

Interessante inhoudelijke informatie krijgen we uit een PRIMALS op de achtergrondsvariabelen. In tabel 4 staan de component ladingen.

TABEL 4: PRIMALS component ladingen

variabele	schaling	ladingen	
1. MBO-opleidingen	942813657	.772	.240
2. leeftijd	15 t/m ≦ 20	.774	-.214
3. sexe	van vrouw naar man	.653	.642
4. SES	SES1 t/m SES9	.701	-.086
5. Vooropleiding	LBO naar HAVO	.678	-.559

Merk op dat we bij PRIMALS aan moeten geven in welke volgorde de categorieën van een variabele geschaald zijn, omdat anders het teken van de componentlading niet te interpreteren is. PRIMALS geeft als eigenwaarden van de correlatiematrix respectievelijk: 2.573, 0.836, 0.629, 0.500, 0.463. Hieruit blijkt dat de eerste component ongeveer evenveel verklaart als de andere vier dimensies tezamen, de anderen verklaren allemaal ongeveer evenveel. Dit patroon vindt men dikwijls bij gegevens waarbij maar één dimensie een rol speelt, de overige componenten zijn 'ruis' en niet te interpreteren.

De eerste component verklaart dus ongeveer 50% van de variantie, men kan zeggen dat de achtergrondsvariabelen voor 50% uit elkaar te verklaren zijn, of preciezer dat ze voor 50% te verklaren zijn uit

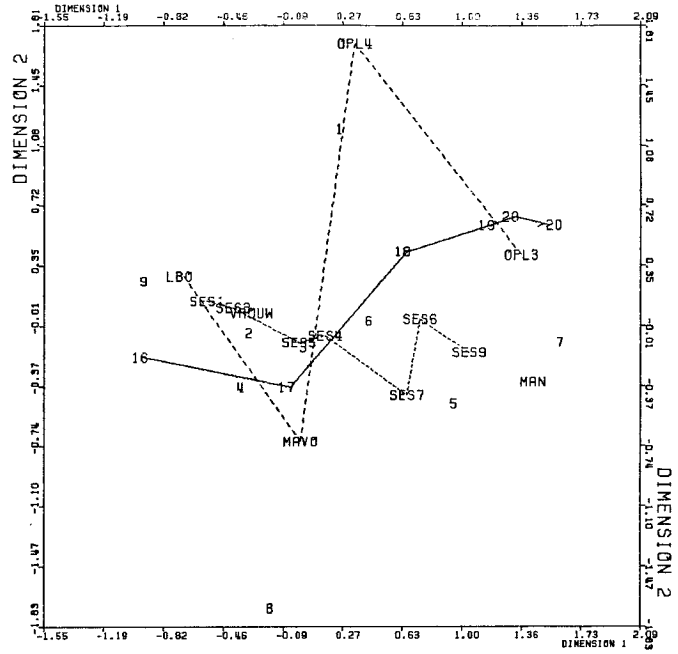
de eerste component, die een gewogen som is van de gestandaardiseerde vijf achtergrondsvariabelen. De optimale gewichten zijn daarbij de componentladingen, uit tabel 4 volgt dat de gewichten ongeveer gelijk zijn, en dat dus de som van de achtergrondsvariabelen een goede benadering van de eerste component is. Dat de correlaties onderling hoog zijn blijkt ook uit tabel 5.

TABEL 5: correlaties na PRIMALS van de 5 achtergrondsvariabelen

	MBO	LEEFT	SEXE	SES	OPLEI
1. opleiding soort	1.00	0.46	0.48	0.40	0.39
2. leeftijd		1.00	0.36	0.43	0.48
3. sexe			1.00	0.33	0.21
4. sociaal-econ. klasse				1.00	0.36
5. vooropleiding					1.00

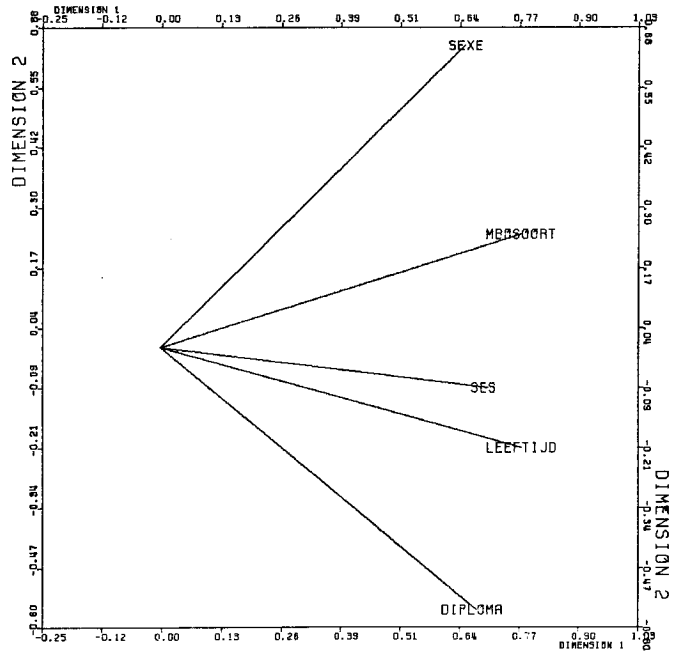
Figuur 6 en 7 (p. 37) behoren ook bij de analyse van de vijf achtergrondsvariabelen. De eerste plot is een HOMALS plot, de tweede een PRIMALS plot. Weliswaar moet dit voorbeeld PRIMALS illustreren, maar het is ook nuttig nog eens met HOMALS te vergelijken. In de HOMALS plot, figuur 6, staan de categorie kwantificaties in twee dimensies, iedere dimensie correspondeert met een aparte kwantificatie van de categorieën. In de PRIMALS plot, figuur 7, worden de categoriekwantificaties van de eerste HOMALS dimensie gebruikt om de eerste twee principale componenten van de bijbehorende correlatiematrix uit te rekenen en te plotten. Wat ziet men aan beide plots?

HOMALS VOOR ACHTERGRONDSVARIABELEN



FIGUUR 6

PRIMALS VOOR ACHTERGRONDSVARIABELEN



FIGUUR 7

1. In de HOMALS plot, figuur 6, worden de opleidingen gelabeled met getallen. De betekenis van de getallen vindt men in de appendix.
2. In de HOMALS plot ziet men de categorieën van de variabelen leeftijd, SES, en vooropleiding verbonden met lijnen. Deze drie variabelen gedragen zich met betrekking tot de eerste dimensie ongeveer gelijk, ze klimmen op van laag naar hoog (opleiding 3 is HAVO, opleiding 4 is MBO). Als we de lijn van MAVO naar HAVO zouden trekken, dan wordt nog duidelijker hoeveel deze drie variabelen met elkaar gemeen hebben. Duidelijk wordt dan ook dat de vooropleidingen in een vierkant liggen, met als twee contrasten hoog (MBO, HAVO) versus laag (LBO, MAVO) en beroepsonderwijs (MBO, LBO) versus algemeen vormend onderwijs (MAVO, HAVO). De twee contrasten worden de twee paren overstaande zijden van het vierkant.

De categorieën van sexe zijn niet met elkaar verbonden. Als men de lijn in gedachten trekt ziet men ook sexe opklimmen van laag (vrouw) naar hoog (man). Bij de MBO-opleidingen ziet men als lage opleidingen 9, 2, en 4, en aan de rechterkant de hoge opleidingen 6, 5, 7. Deze volgorde is vanzelfsprekend dezelfde als in tabel 4. Voor de volledigheid nogmaals: als we over 'hoog' en 'laag' praten, bedoelen we daar geen waardering mee, maar uitsluitend de plaats op de eerste principale component. Hoog op deze component komen oudere mannen met HAVO uit de SES9 groep die op MBO-opleiding 5, 6, of 7 zitten. Laag komen jonge vrouwen met LBO of MAVO uit SES1 of SES2 die op opleiding 2, 4, of 9 zitten. Dat deze dimensie sterk samenhangt met maatschappelijke waardering is niet iets wat de techniek vindt, dit komt pas ter sprake in het stadium van de interpretatie.
3. De conclusies uit deze HOMALS (behalve de eerder genoemde één-dimensionale structuur) luiden dan ook als volgt. Hoe hoger de leeftijd, hoe hoger de vooropleiding, hoe hoger de SES (zie ook de correlaties in tabel 5 op p. 36, bij een steekproefgrootte zoals deze zijn die zéér significant).
4. In de PRIMALS plot, figuur 7 op p. 37, zien we dat de lengte van de pijlen voor alle vijf de variabelen even groot is. Ze scoren alle vijf ook ongeveer even hoog op de eerste dimensie (zie ook tabel 4). Op de tweede component zijn er zowel positieve als negatieve ladingen. Op de tweede dimensie in figuur 7 ziet men de pijlen dan ook ver uit elkaar liggen. Zo ver zelfs dat figuur 7 suggereert dat sexe en

diploma ongecorreleerd zijn, omdat ze in de figuur loodrecht op elkaar staan. De feitelijke correlatie (zie tabel 5) is echter .21, weliswaar veruit de laagste uit de tabel, maar nog steeds veel groter dan nul. Aan de tweede dimensie moet men niet veel waarde toekennen. Deze verklaart immers slechts een kwart van de overgebleven helft na verwijdering van de eerste dimensie. En de derde en volgende dimensie verklaren ongeveer evenveel, gezien de eigenwaarden. Als we ook deze dimensies zouden uitrekenen (PRIMALS kan dat niet, maar we kunnen PCA op de correlatiematrix uit tabel 5 doen) zouden de pijlen er wel eens heel anders uit kunnen zien.

4.3 PRINCALS, met één voorbeeld

We kunnen PRINCALS het beste introduceren door het te vergelijken met PRIMALS. In PRIMALS kiezen we de kwantificaties van de categorieën op zo'n manier dat de eerste eigenwaarde van de correlatiematrix zo groot mogelijk wordt, met andere woorden dat de variabelen zo ééndimensionaal mogelijk worden. PRINCALS is in twee opzichten meer algemeen. In de eerste plaats is er de mogelijkheid om de kwantificaties zo te kiezen dat de som van de eerste twee (of van de eerste drie, enzovoort) eigenwaarden zo groot mogelijk wordt. In PRINCALS kun je dus kiezen hoeveel-dimensionaal je de correlatiematrix wilt maken, PRIMALS is het speciale geval waarin je zo goed mogelijk mikt op één-dimensionaliteit. Het tweede verschil tussen PRIMALS en PRINCALS is minstens even interessant. Alle variabelen in PRIMALS worden behandeld als nominaal, dat wil zeggen er worden van te voren geen categorie kwantificaties uitgesloten, optimale categorie kwantificaties worden uitsluitend berekend door rekening te houden met het criterium, de grootste eigenwaarde van de correlatiematrix. Als variabelen (zoals bijvoorbeeld in MHNO/MSPO de motivatievariabelen, of leeftijd) duidelijk ordinaal zijn (dus categorieën hebben die duidelijk geordend zijn), dan houdt PRIMALS daar geen rekening mee. Het kan zijn dat we a priori ordening terugvinden, het kan ook zijn dat het programma een geheel andere ordening oplevert dan we verwacht hadden. Dat laatste is niet noodzakelijk een nadeel, bij variabelen als SES en vooropleiding kan het zelfs een voordeel zijn, omdat daar de a priori ordening niet 100% zeker is, en meer het karakter heeft van

een hypothese. Aan de andere kant kan het mogelijk zijn dat we in sommige gevallen (door de aard van de variabele of door eerder onderzoek) zo overtuigd zijn van de ordinale eigenschappen, dat we ze als eis stellen. Dit kán met PRINCALS, men kan in dit programma eisen dat de categorie kwantificaties monotoon zijn met een van tevoren opgegeven volgorde. Men kan tevens eisen met PRINCALS dat de gevonden categorie kwantificaties lineair zijn met van te voren opgegeven getallen. Van iedere variabele in PRINCALS kan men dus van tevoren kiezen of hij nominaal, ordinaal, of numeriek opgevat moet worden. Hieruit volgt dat PRIMALS een speciaal geval is van PRINCALS (kies alle variabelen nominaal, en maximaliseer de eerste eigenwaarde), en tevens is gewone PCA een speciaal geval van PRINCALS (kies alle variabelen numeriek).

Om nog even te recapituleren: PRIMALS zowel als PRINCALS geven enkelvoudige kwantificaties, dat wil zeggen per analyse één stel categoriekwantificaties en één bijbehorende correlatiematrix. HOMALS geeft meervoudige kwantificaties , dat wil zeggen iedere HOMALS dimensie geeft categoriekwantificaties op basis waarvan men een correlatiematrix uit zou kunnen rekenen. HOMALS rekent overigens die correlatiematrixen zelf niet uit, omdat HOMALS output beter geïnterpreteerd kan worden in termen van afstanden tussen categoriepunten en observatiepunten in een twee-dimensionale ruimte. PRINCALS generaliseert PRIMALS, omdat het de mogelijkheid geeft om per variabele het meetniveau verschillend te kiezen, en omdat het de mogelijkheid heeft om het optimaliteitscriterium (het aantal eigenwaarden van de correlatiematrix dat gemaximaliseerd wordt) te kiezen. HOMALS generaliseert PRIMALS omdat het de mogelijkheid tot meervoudige kwantificatie heeft, uit een HOMALS oplossing kan men altijd de PRIMALS oplossing afiezen, omdat PRIMALS eenvoudigweg overeenkomt met de eerste HOMALS dimensie. De programma's verschillen overigens in organisatie van de output, en ook in een groot aantal andere details. Zo heeft PRIMALS een aantal verschillende opties voor de aanpak van ontbrekende gegevens, terwijl HOMALS en PRINCALS slechts één mogelijke aanpak hebben.

Het bovenstaande geeft een enigszins vereenvoudigd beeld van de werkelijke situatie, omdat we een complicatie buiten beschouwing

gelaten hebben. Het programma PRINCALS is zo geschreven dat het ook HOMALS generaliseert. Anders gezegd: men kan voor iedere variabele bovendien nog kiezen of hij meervoudig of enkelvoudig behandeld wordt, zodat de mogelijke meetniveau's worden: meervoudig nominaal, enkelvoudig nominaal, enkelvoudig ordinaal, enkelvoudig numeriek. HOMALS is het speciale geval waarin alle variabelen meervoudig nominaal behandeld worden. Wanneer er in een PRINCALS toepassing zowel meervoudige als enkelvoudige variabele voorkomen, dan is het weer beter af te stappen van de PCA-interpretatie, en over te gaan op de meetkundige interpretatie à la HOMALS. De meetkunde van PRINCALS staat in detail in Gifi (1981a).

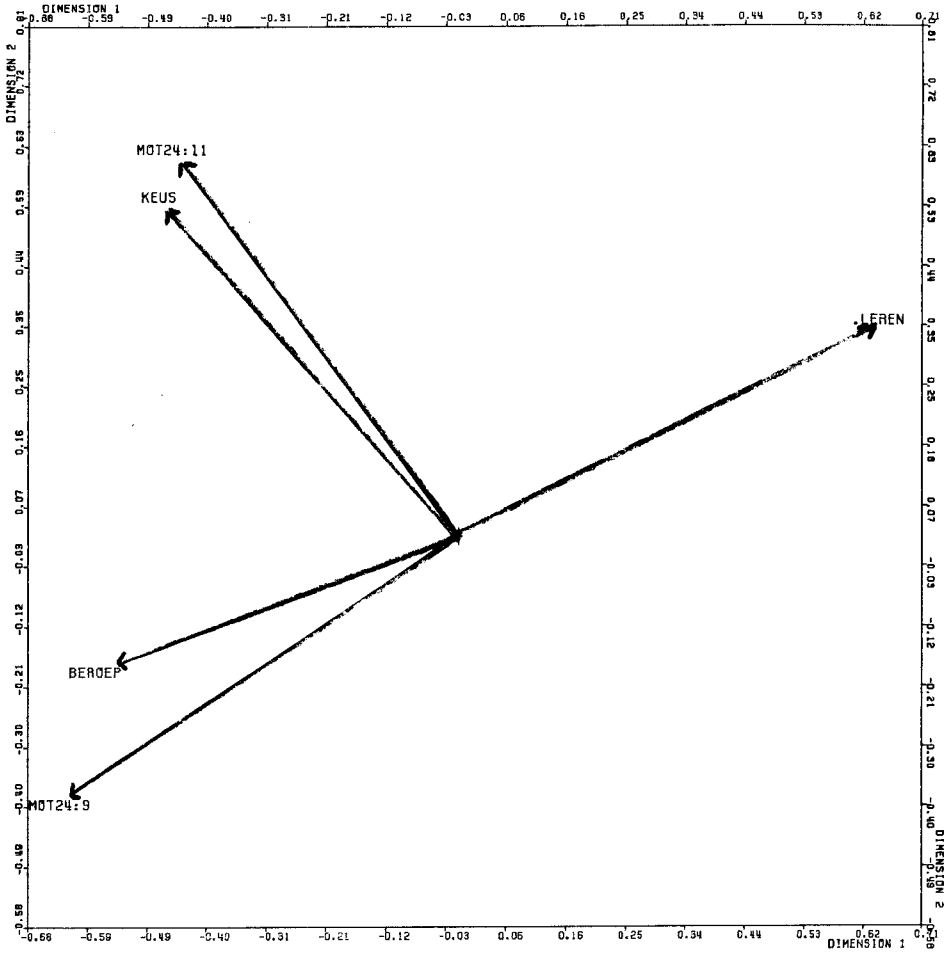
PRINCALS-voorbeeld

PRINCALS met alle variabelen enkelvoudig (single) ordinaal hebben we uitgedraaid over de variabelen in tabel 6 op p. 42. In die tabel vindt men de ladingen op de eerste twee componenten (we hebben de som van de eerste twee eigenwaarden gemaximaliseerd, dat wil zeggen de correlatiematrix zo twee-dimensionaal mogelijk gemaakt). De variabelen met een hoge lading op de eerste dimensie hebben we bovendien geplot in figuur 8 op p. 43. In tabel 6 hebben variabelen 1, 3, 5, en 9 hoge negatieve ladingen, terwijl variabele 8 een hoge positieve lading heeft. De ladingen op de eerste component kunnen we vergelijken met de overeenkomstige HOMALS discriminatiematen in tabel 1 of tabel 2 op p. 25 en 27 (bedenk daarbij dat de discriminatiematen kwadraten van componentladingen zijn). In beide gevallen zijn variabelen 1, 8, en 9 het best.

TABEL 6: PRINCALS analyse over 9 variabelen met component ladingen

variabele	componentladingen	
1. motivatie 24:09.	-.624	-.393
2. motivatie 24:10.	-.274	.056
3. motivatie 24:21.	-.456	.621
4. al gewerkt ? Ja/Nee.	.217	.053
5. was de school je eerste keus ? Ja/Nee.	-.477	.548
6. schoolmelding ? Ja/Nee.	.061	-.543
7. opleidingsmelding ? Ja/Nee.	.023	-.242
8. ga je verder leren ? Ja/Weet niet/Nee.	.659	.355
9. ga je je beroep uitoefenen ? Idem.	-.575	-.188
eigenwaarden	.191	.151

FIGUUR 8: Componentladingen PRINCALS-voorbeeld



De bovenstaande figuur 8 moet men als volgt interpreteren. Leren-ja ligt rechts van het centrum, leren-nee en leren-weet-niet liggen aan de andere kant. Beroep-ja ligt links van het centrum van de plot, beroep-nee en beroep-weet-niet liggen aan de andere kant. De pijl wijst dus steeds naar de ja-categorie. Doordat de ordinale optie gebruikt is, zijn de categorieën in de volgorde 1 - 2 - 3 gedwongen. Bij de overige variabelen is dat soepel gelopen, maar bij 'beroep' en

'leren' niet. De oorspronkelijke categorieën zijn 1 = ja, 2 = nee, 3 = weet niet. In een nominale analyse (PRIMALS, of PRINCALS nominaal) wordt de volgende 1 = ja, 2 = weet niet, 3 = nee. Dit kunnen we zien aan bijvoorbeeld de eerste HOMALS dimensie in de twee eerste HOMALS voorbeelden. Een nominaal programma draait dus, heel logisch, de volgorde van de twee laatste categorieën om. Door de volgorde 1 - 2 - 3 op te leggen, doet men eigenlijk iets onnatuurlijks. Twee alternatieven zijn beter, namelijk eerst de variabelen in de logische volgorde hercoderen en dan PRINCALS ordinaal, of beide variabelen nominaal behandelen in PRINCALS. Beide oplossingen geven in dit geval hetzelfde resultaat, hoewel dit niet noodzakelijk zo is. We rapporteren deze analyses niet, omdat het verschil minimaal is. Zowel 'beroep' als 'leren' krijgen een iets hogere lading op de eerste component. Een gevolg van de foutieve codering die we gebruikt hebben is dat PRINCALS leren-nee en leren-weet-niet een gelijke kwantificatie geeft, en hetzelfde voor beroep-nee en beroep-weet-niet.
Waar het label 'mot24:9' of 'mot24:21' staat moet men het volgende lezen: deze motivatie heeft helemaal geen invloed gehad op mijn keuze voor deze opleiding. Aan de andere kant van het centrum van de plot ligt dus categorie 5 van deze variabele: deze motivatiereden heeft zeer veel invloed gehad op mijn keuze voor deze opleiding. Voor 'keus' geldt: waar het label staat en de pijl eindigt ligt 'eerste-keus-ja', daartegenover ligt 'eerste-keus-nee'.
We zien in figuur 8 dat 'beroep' tegenover 'leren' staat. Motivatie 24:9 ligt bij beroep, maar we weten dat dit betekent dat voor mensen die hun beroep gaan uitoefenen deze motivatiereden géén rol speelt, maar juist wel voor verder leerders. 'Keus' ligt op de eerste dimensie gelijk met beroep, zoals we al eerder gezien hebben, maar verschilt er op de tweede dimensie aanzienlijk van. We zien dat motivatie 24:21 niet van belang is voor de 'eerste keuzers' maar wel voor de 'tweede keuzers'. Uit de formulering van de motivatievragen is dat ook duidelijk genoeg in te zien. Dat eerste-keus/tweede-keus loodrecht staat op de belangrijkste dimensie beroep-uitoefenen/verder-leren hoeft, evenals bij PRIMALS, op zichzelf niet veel te betekenen. Om de waarde van dit resultaat goed in te kunnen schatten zouden we ook de hogere eigenwaarden van de correlatiematrix moeten bekijken.

5 SAMENVATTING

Bij deze exploratie van de eerste dataset uit het MHNO/MSPO-onderzoek vinden we samenhangen die we bij de analyse van de tweede dataset (het cohort 1981/1982) kunnen (en zullen trachten te) verifiëren. De meest opvallende en sterkste samenhang is die tussen de achtergrondsvariabelen leeftijd, sexe, vooropleiding, SES, en de huidige MBO-opleiding. De negen verschillende MBO-opleidingen in dit onderzoek verschillen ook van elkaar wat hun populatie leerlingen betreft, en wel in een zo hoge mate dat we voorzichtig moeten zijn met algemene uitspraken die de gehele MHNO/MSPO populatie betreffen. Er zijn opleidingen met meer dan te verwachten aantallen jongens van een hogere leeftijd uit een hoger SES-milieu en met een hogere vooropleiding met daartegenover opleidingen met meer meisjes dan te verwachten met een lagere opleiding, lager SES, en lagere leeftijd. Beantwoording van de vraagstellingen in het onderzoek zal, gegeven deze grote verschillen, beter per opleiding of per groep opleidingen gegeven kunnen worden dan voor het gehele MHNO/MSPO. Een andere consequentie van de sterke samenhang tussen achtergrondsvariabelen onderling, is dat het moeilijk wordt uitspraken te doen over samenhang van afzonderlijke achtergrondsvariabelen met motivatievariabelen en toekomstplannen. Dat jongens, leerlingen uit hogere SES-milieus, en oudere leerlingen relatief vaak willen doorstuderen zijn verre van onafhankelijke conclusies, het gaat hierbij voor een groot deel om dezelfde groep. Het spreekt dus vanzelf dat dit en soortgelijke effecten alleen te ontdekken zijn met een multivariate aanpak, en dat multivariate analysetechnieken dit soort effecten op een directe en inzichtelijke manier afbeelden.

REFERENTIES:

M. Van Dijck: Doorstroming, differentiatie en stage in het MHNO/MSPO nieuwe stijl. Interim-rapportage over de periode 1-10-'80 tot en met 31-8-'81. LICOR, augustus 1981.

M. van Dijck: Vragen betreffende de gebruikte analyse-methoden. Handgeschreven notitie, november 1981.

M. Faddegon & R. de Rooy: Projektnieuws 81/04/02. LICOR/DSWO, 1981.

J.P. van de Geer: Introduction to multivariate analysis for the social sciences. Freeman, San Francisco, 1971.

A. Gifi: Nonlinear multivariate analysis. Datatheorie FSW/RUL, 1981.

A. Gifi: HOMALS users guide. Datatheorie FSW/RUL, 1981.

W.J. Heiser: Unfolding analysis of proximity data. Datatheorie FSW/RUL, 1981.

G.G. Kreft: Methodologische kanttekeningen bij het MHNO/MSPO onderzoek. DSWO/Veldwerk, 1982.

LICOR: Doorstroming, differentiatie, stage in het MHNO/MSPO nieuwe stijl. Subsidieaan-vraag. LICOR, september 1980.

Meerling: Methoden en technieken van psychologisch onderzoek. Deel I en II. Boom, Meppel, 1980 en 1981.

APPENDIX 1: Cohorts, gebruikte vragen, marginalen

1:

De analyse is uitgevoerd op de eerste twee cohorts tezamen. Cohort 1, instroomjaar 1979/1980, afnamejaar 1981, N = 2146, en cohort 2, instroomjaar 1980/1981, afnamejaar 1981, N = 3237.

2:

De gekozen analysevariabelen zijn de volgende. Er zijn zeven achtergrondsvariabelen (1 t/m 7) en negen motivatie- en andere onderzoeksvariabelen (8 t/m 16). Variabele 10 bestaat uit 21 aparte vragen, waarbij gevraagd wordt of en in hoeverre 21 bepaalde motivaties bij de schoolkeuze een rol gespeeld hebben. Eigenlijk zijn er dus 29 niet-achtergrondsvariabelen. Variabele 11 lijkt op 10, alleen wordt er bij 11 gevraagd welke van de genoemde 21 motivaties de belangrijkste was. Vraag 11 heeft dus 21 categorieën. In dit rapport wordt vraag 11 niet geanalyseerd, analyses waarbij vraag 11 een rol speelt zijn uitgevoerd met het programma ANACOR. Zij worden gerapporteerd in G.G. Kreft en J. de Leeuw: Differentiële motivatie bij keuze van MBO-opleiding: een toepassing van correspondentieanalyse. Datatheorie, 1982.

3:

De 36 variabelen spelen in beide cohorts een gelijke rol, met uitzondering van vraag 1. De MBO-opleiding VP startte pas in 1980, en komt dus alleen in het tweede cohort voor. De vragen zijn zo gekozen dat ze, naar wij vermoeden, aan de beantwoording van de diverse vraagstellingen van het projekt meewerken (zie voor een nadere toelichting Kreft, 1982). In eerste instantie bestaan de vragen uit twee verschillende sets, achtergrondsvariabelen en onderzoeksvariabelen, waarbij de laatste soort vooral uit motivatievariabelen bestaat. Het ligt voor de hand deze twee sets van variabelen aan elkaar te relateren met de gegeneraliseerde techniek voor canonische analyse die in Gifi (1981) beschreven saat (programma CANALS). Deze analyses zullen elders gerapporteerd worden.

4:

We geven nu gebruikte variabelen, met marginale percentages, en diverse coderingen.

1. MBO-opleidingen:	1:AB	: 15%
	2:AG	: 14%
	3:HT	: 3%
	4:MK	: 8%
	5:SA/1	: 5%
	6:SA/2	: 14%
	7:SB	: 12%
	8:VP	: 9%
	9:VZ	: 20%
2. Start van de opleiding:	1: cohort 1979	: 40%
	2: cohort 1980	: 60%
3. Leeftijd:	1: 15 jaar	: 1%
	2: 16 jaar	: 34%
	3: 17 jaar	: 30%
	4: 18 jaar	: 20%
	5: 19 jaar	: 8%
	6: 20 jaar	: 4%
	7: ≥ 20 jr.	: 4%
4. Sexe:	1: man	: 17%
	2: vrouw	: 83%
5. SES (opleiding vader):	1: SES1	: 18%
	2: SES2	: 0%
	3: SES3	: 29%
	4: SES4	: 16%
	5: SES5	: 13%
	6: SES6	: 9%
	7: SES7	: 11%
	8: SES8	: 0%
	9: SES9	: 3%
6. Blijven zitten:	1: Ja	: 41%
	2: Nee	: 59%
7. Laatste diploma:	1: LBO	: 33%
	2: MAVO	: 42%
	3: HAVO	: 11%
	4: MBO	: 10%
	5: anders	: 5%
8. Vroeger gewerkt	1: Ja	: 81%
	2: Nee	: 19%

9. Nu een bijbaan:	1: Ja	: 41%
	2: Nee	: 59%
10. Motivatievraag 24 (zie verderop).		
11. Motivatievraag 25 (zie verderop).		
12. Is deze school je eerste keus:	1: Ja	: 67%
	2: Nee	: 33%
13. Bij andere scholen aangemeld:	1: Ja	: 30%
	2: Nee	: 70%
14. Bij andere opleidingen aangemeld:	1: Ja	: 39%
	2: Nee	: 61%
15. Verder leren:	1: Ja	: 31%
	2: Nee	: 17%
	3: Weet niet	: 52%
16. Beroep uitoefenen:	1: Ja	: 69%
	2: Nee	: 7%
	3: Weet niet	: 25%

N.B. De percentages zijn steeds berekend over het aantal respondenten dat de vraag beantwoord heeft, met ontbrekende gegevens is hierbij geen rekening gehouden (zie Kreft, 1982). De belangrijkste onderzoeksvragen zijn 15 en 16. We geven daarom van deze twee vragen de letterlijke vraagstelling.

15. Ben je van plan na deze opleiding nog verder te gaan leren?
16. Ben je van plan het beroep uit te gaan oefenen waarvoor je nu opgeleid wordt?

5: De MBO-opleidingen

De volgende coderingen werden gebruikt:

AB : Aktiviteitenbegeleiding.
AG : Tandarts-, dokters-, apothekersassistent, assisterende in de gezondheidszorg.
HT : Huishoudtechnische sector, civiele dienst.
MK : Mode en kleding, kostuumnaaien, couture.
SA1: Arbeids- en personeelwerk, sociale dienstverlening.
SA2: Inrichtingswerk en cultureel werk, KV/JV.
SB : Sport en bewegen, CIOS.
VP : Verpleegkundige.
VZ : Bejaarden- en gezinsverzorging.

6: De bepaling van het SES

De sociaal-economische status (het SES) kan in deze vragenlijst bepaald worden aan de hand van vraag 13 (opleiding vader, beroep vader, opleiding moeder, beroep moeder). We veronderstellen dat aan deze vier variabelen één onderliggende dimensie, het SES, ten grondslag ligt. Een aangewezen programma om dit te onderzoeken is PRIMALS. De voornaamste conclusies uit de PRIMALS-analyse zijn de volgende.

a: De opleiding van de vader correleert het hoogst met de door deze vier variabelen gedefinieerde principale component. De componentlading is .91. Dit betekent dat men zonder veel verlies van informatie deze variabele kan gebruiken voor het bepalen van het SES, althans in deze populatie.

b: We geven een aantal correlaties tussen gekwantificeerde variabelen die het bovenstaande ondersteunen.

Correlatie tussen: beroep moeder - opleiding moeder = .46
Correlatie tussen: beroep moeder - beroep vader = .53

We zien dus: moeders beroep correleert hoger met vaders beroep dan met haar eigen opleiding. Hetzelfde zien we, maar dan nog sterker, terug in de correlaties tussen de opleidingen van vader en moeder.

Correlatie tussen: opleiding moeder - opleiding vader = .61
Correlatie tussen: beroep vader - opleiding vader = .72

De laatste correlatie is de hoogste. Vaders beroep past beter bij zijn opleiding dan moeders beroep, vaders opleiding past beter bij moeders opleiding dan bij moeders beroep. Opleiding vader is daarom een goede predictor van SES.

We tekenen hierbij overigens het volgende aan. Bij de codering van de beroepen is uitgegaan van de CBS-beroepenklapper. Met aanvullingen voor die beroepen waarover de klapper geen informatie bevat. Beroep 'huisvrouw' is niet als beroep gecodeerd. Het is of als ontbrekend gegeven beschouwd, of afgeleid uit de opleiding en als zodanig gecodeerd. Hieruit volgt natuurlijk dat beroep moeder de minst betrouwbare van de vier SES-vragen zal zijn.

De opleiding vader is in dit onderzoek als volgt gecodeerd (hetzelfde geldt natuurlijk ook voor opleiding moeder).

SES1 Alleen lager onderwijs.
SES2 VGLO of LAVO. Deze categorie bevat slechts 4 personen. In de analyses wordt deze categorie samen genomen met categorie SES1.
SES3 LBO.
SES4 ULO, MULO, driejarige HBS.
SES5 MBO.
SES6 VHMO, HBS, Gymnasium, Lyceum, Atheneum.
SES7 HBO.
SES8 Semi-hoger onderwijs (bijv. KMA, Landbouwhogeschool). Ook hier komen weinig personen in voor (totaal 18), deze categorie wordt daarom samen genomen met het HBO, met SES7 dus.
SES9 Universitaire opleiding.

7: Laatste diploma
De codering van het diploma is geschied naar het laatste diploma dat de leerling behaald heeft. Het is dus mogelijk dat een leerling, die hier bij MAVO staat, eerst een LBO-diploma heeft gehaald. Wat echter vaker voorkomt is dat leerlingen met een MBO-diploma (meestal Inas Intas VK/VZ en vormingsklas) eerst MAVO of LBO gehaald hebben. Er wordt dus geen onderscheid gemaakt tussen leerlingen, die MAVO + MBO en leerlingen die LBO + MBO hebben gedaan. Hetzelfde doet zich voor bij leerlingen die een HAVO-diploma hebben. Ook hier wordt geen onderscheid gemaakt tussen leerlingen die de weg MAVO - HAVO kozen en leerlingen die regelrecht naar de HAVO zijn gegaan. De bedoeling is hier in volgende analyses wél onderscheid in te maken.

8: Motivatievraag 24
Deze vraag bestaat uit 21 vragen, ieder met vijf antwoordcategorieën. De vragen zijn 21 motivatieredenen, de antwoordcategorieën zijn geen invloed - weinig invloed - tamelijk veel invloed - veel invloed - zeer veel invloed. We geven voor alle 21 vragen de verdeling in percentages over de vijf categorieën.

	1	2	3	4	5	vraag 25
01: Op mijn vorige school werd mij deze opleiding aangeraden.	47	20	18	09	06	91 = 2%
02: Met deze opleiding kan je later veel kanten op.	11	16	29	25	19	407 = 8%
03: Ik was nog (partieel) leerplichtig.	78	11	05	03	03	35 = 1%
04: De opleiding die ik eigenlijk wilde volgen was te duur.	95	02	01	01	01	21 = 0%
05: Ik wilde al langer voor dit beroep gaan leren.	19	10	17	20	34	704 = 13%
06: Deze opleiding werd mij aangeraden door een adviesinstantie.	86	05	04	02	02	32 = 1%
07: Ik heb al gewerkt in het werk waarvoor ik hier opgeleid word.	73	07	07	05	08	131 = 3%
08: Er is veel vraag naar mensen met deze opleiding.	32	22	26	13	07	72 = 1%
09: Ik heb deze opleiding nodig om verder te kunnen studeren voor het beroep dat ik eigenlijk wil uitoefenen.	56	12	11	09	12	294 = 6%
10: Door de samenstelling van mijn vakkenpakket had ik niet veel keuzemogelijkheden.	85	07	05	02	02	18 = 0%
11: In deze opleiding hoef ik nog niet definitief te beslissen welk beroep ik later wil uitoefenen.	57	16	04	07	06	88 = 2%
12: Bij deze opleiding zit je niet alleen maar te studeren, je bent ook praktisch bezig.	10	13	27	25	26	488 = 9%
13: Ik wist dat er op deze school een leuke sfeer heerst.	62	16	13	06	04	11 = 0%
14: Het vak waarvoor ik nu opgeleid word, wordt goed betaald.	56	24	13	05	02	13 = 0%
15: Deze school was dicht bij mijn huis.	84	08	04	02	02	9 = 0%
16: Dit beroep kan je blijven uitoefenen als je voor je kinderen moet zorgen.	56	16	13	08	07	82 = 2%
17: Ik was nog te jong voor de opleiding die ik eigenlijk wilde volgen.	87	03	03	02	04	121 = 2%
18: Mijn ouders wilden graag dat ik deze opleiding ging volgen.	82	10	05	02	01	17 = 0%
19: Het beroep waarvoor deze school opleidt, trekt mij erg aan.	04	03	14	23	57	2236 = 43%
20: Ik kon geen geschikte baan vinden na mijn vorige opleiding.	92	03	02	01	02	35 = 1%
21: De opleiding die ik eigenlijk wilde volgen zat vol.	91	02	02	01	03	90 = 2%

9: Motivatievraag 25

Hier wordt gevraagd: Welke van de bij vraag 24 genoemde redenen heeft voor jou de meeste invloed gehad bij het maken van je keuze? deze vraag heeft du 21 alternatieven, ruwe frequenties en percentages staan in de laatste kolom van bovenstaande tabel.

APPENDIX 2: Praktische tips voor HOMALS-gebruikers.

1:
Nummer en/of dateer elke output die je bewaart. Op deze manier kan je later terugvinden hoe de totale analyse verlopen is.

2:
Zet op elke output die je bewaart wat je op het eerste gezicht opvalt. Zet er ook bij op welke gronden je besluit om een nadere of andere analyse te doen over bepaalde variabelen. Immers ook bij HOMALS is het aantal verschillende analyses die je op een grote dataset kunt uitvoeren zéér groot.

3:
Beter dan bij twee wordt gesuggereerd is het volgende: houdt een logboek bij. Bij volgende analyses, waarbij je steeds minder naief wordt met betrekking tot de structuur van de data, ben je al snel vergeten welke stappen je gezet hebt, en waarom. Met andere woorden: je vergeet snel met welke hypothesen je de analyse begon als de resultaten anders uitvallen dan je verwachtte. Dit bijhouden van een logboek heeft twee functies. In de eerste plaats vermijd je het uitdraaien van veel op elkaar lijkende analyses, dat wil zeggen analyses die niet nodig zouden zijn als je de oude output beter (en met je scherpere oog vanwege de verkregen kennis van de data) bekeken had. Je vindt in 'oude' output vaak weer nieuwe resultaten. In de tweede plaats is het voor latere rapportage van de analyse vaak nodig je te herinneren waarom en waartoe je bepaalde dingen hebt gedaan. Immers analyses die weinig opbrengen ben je geneigd snel te vergeten. Vaak komen echter vragen over variabelen die je 'vergeten' bent.

4:
Als één of meerdere variabelen een lage tot zéér lage discriminatiemaat hebben, dan hebben ze ook weinig tot niets aan de gevonden oplossing bijgedragen. Dit betekent dat een nieuwe analyse met alleen die variabelen die wel een hoge discriminatiemaat hebben, dus met weglating van de variabelen met een lage discriminatiemaat, geen wezenlijk andere oplossing zal geven. Nieuwe analyses zijn in zo'n geval overbodig. Tenzij met de variabelen die een lage discriminatiemaat hebben.

Het is aan te raden die nog eens apart te bekijken, eventueel in combinatie met andere variabelen.

5:
Een hogere discriminatiemaat is niet altijd een absolute maat van belangrijkheid voor de desbetreffende variabele. Een discriminatiemaat heeft een verschillend maximum voor verschillende variabelen. Dit maximum wordt bepaald door het aantal categorieën van de variabele. Een variabele met 21 categorieën (zoals motivatievariabele 25 in dit onderzoek) heeft gemiddeld een 5 maal zo grote discriminatiemaat als een variabele met 4 categorieën. Zoals al eerder in dit paper werd opgemerkt is HOMALS het beste te interpreteren voor variabelen met een ongeveer gelijk aantal categorieën.

6:
Neem bij grote steekproeven of populaties (zoals in ons onderzoek) geen gepartitioneerde object scores plots. Je krijgt een ongelooflijk dik pak output (je moet daarom erg veel 'lines' opgeven), waar je in feite weinig aan hebt. Vraag in een dergelijk geval alleen de plots van de 'unlabeled object scores' en de categorie kwantificaties.

7:
Hercodeer variabelen met de categorieën die niet discrimineren. Neem deze categorieën bij elkaar. Doe hetzelfde met categorieën die een zelfde categoriekwantificatie hebben. Deze handelingen veranderen niets aan de structuur en samenhang van de variabelen. Een voordeel van deze handelswijze is echter dat je hierdoor beter gevulde categorieën krijgt. Het gevaar van 'lege' categorieën werd reeds eerder behandeld in dit paper.

8:
Als twee of meer achtergrondsvariabelen hoog correleren is het aan te raden er slechts één van actief in de analyse mee te laten doen. Je kan de andere achtergrondsvariabelen passief mee laten doen, of weglaten. Dit laatste is bij grote bestanden, zoals het onze, de enige praktische optie. HOMALS geeft niet, zoals dat wel het geval is bij PRIMALS, de categoriekwantificaties van de passieve variabelen. Men moet in dat geval afgaan op de object-score plots. Maar zoals we al

aangaven onder nummer 6 is dit bij een groot aantal respondenten ondoenlijk. In de toekomst schijnt HOMALS ook categoriekwantificaties over de passieve variabelen in de analyse aan te geven, zodat ook voor grote bestanden passieve variabelen mee kunnen doen.

9:
Als een grote set variabelen iets gezamenlijks meet, gebruiken we bij voorkeur PRIMALS. In ons onderzoek zijn dat de 21 motivatie-variabelen. Bij scheve variabelen moet men voorzichtig zijn met de interpretatie, omdat de PRIMALS kwantificaties en correlaties dan wel eens alleen scheefheid oftewel antwoordtendentie kunnen weergeven.

10:
Bij een scheve verdeling van een variabele - de beide uiterste categorieën zijn zeer ongelijk gevuld - zal men bij HOMALS vaak geen hoefijzer vinden, zelfs al is de onderliggende structuur ééndimensionaal. Dat komt door de eigenschap van HOMALS om een categorie die veel respondenten bevat bij het centrum (nulpunt) te plaatsen, en categorieën met weinig respondenten naar de periferie van de plot. Een één-dimensionale structuur leidt in zo'n geval tot een J-vormige plot van categoriekwantificaties voor de variabele, waarbij de J natuurlijk ook gespiegeld kan zijn.