Tijdschrift voor Onderwijsresearch, 18 (1993), nr. 1, pp. 26-41

Multiple correspondentie-analyse van school- en beroepsloopbanen

Toon W. Taris en Peter G.M. van der Heijden*

ABSTRACT

This article addresses the use of multiple correspondence analysis (MCA) in the analysis of sequential data, such as event histories. First the principles of MCA are discussed. Subsequently a typical data set concerning the educational and employment careers of 265 young adults is analyzed. Then a simple four-class career typology is constructed. Graphical representations of the careers within each group indicate that these indeed differ strongly, providing evidence of the validity of the approach taken here. The paper concludes that MCA is a tool that enables the scientist to construct viable classifications of careers, which yields rather much understanding in the reasons why careers are grouped in a specific way.

1. INLEIDING

Het verloop van school- en beroepsloopbanen is voor onderwijsonderzoekers al geruime tijd een dankbaar object van studie. Longitudinaal onderzoek op dit terrein levert vaak data in de vorm van *event histories* op: sequenties van kwalitatieve toestanden die een subject gedurende de geobserveerde periode bezette, alsmede de tijdstippen waarop overgangen van de ene naar de andere toestand plaatsvonden. Deze toestanden zijn over het algemeen de categorieën van een nominale variabele, bijvoorbeeld 'werkend', 'werkloos' en 'niet beschikbaar voor de arbeidsmarkt'.

In de beschrijving en verklaring van het verloop van dergelijke loopbanen zijn grofweg twee benaderingen te onderscheiden. De eerste benadering richt zich op de *afzonderlijke overgangen* van de ene toestand naar de andere; bijvoorbeeld overgangen van werkloosheid naar werk, van alleenstaand naar samenwonend, of van werkzaam in een bepaalde baan naar niet meer werkend in die baan. De sequentie van toestandsovergangen van een bepaald subject wordt aldus opgedeeld in de afzonderlijke episoden, waarin het subject in één bepaalde toestand verkeerde (bijvoorbeeld de tijd dat men werkloos was). Op dit moment lijkt dit de dominante benadering, wat wellicht samenhangt met het feit dat er geavanceerde statistische technieken beschikbaar zijn voor de analyse van dit type gegevens (Allison, 1984; Blossfeld, Hamerle & Mayer, 1989; Tuma & Hannan, 1984, geven overzichten van deze technieken, die nauw gerelateerd zijn aan overlevingsduuranalyses uit de biologie). Deze technieken zijn bij uitstek geschikt voor de identificatie van factoren die het plaatsvinden van een bepaalde overgang beïnvloeden.

De tweede benadering richt zich op de beschrijving en verklaring van het verloop van de loopbanen gedurende een bepaald *interval*. In dit interval kunnen meerdere overgangen plaatsvinden, die in principe allemaal in de analyse betrokken worden. Centraal in deze benadering staat het vergelijken en classificeren van deze loopbanen, en het gebruiken van deze classificaties in vervolg-analyses (Abbott & Hrycak, 1990). Daarbij wordt aangenomen dat het resultaat niet goed is te modelleren als het resultaat van een of ander onderliggend stochastisch mechanisme (bijvoorbeeld een eerste-orde Markovproces, Van de Pol, 1989). Een belangrijk probleem hierbij is dat het aantal theoretisch mogelijke loopbaanpatronen over het algemeen zeer groot is: als s het aantal mogelijke toestanden (bijvoorbeeld beroepen, schooltypen, enzovoort) is, en t het aantal tijdstippen waarop deze toestanden zijn waargenomen gedurende een bepaalde periode, dan bedraagt het aantal theoretisch mogelijke loopbanen s^t. Als er gedurende 6 tijdstippen geobserveerd wordt en er 4 toestanden onderscheiden worden, dan zijn er in principe $4^6 = 4096$ verschillende loopbanen denkbaar. Het aantal empirisch in de steekproef voorkomende sequenties zal meestal kleiner zijn, maar toch al snel het bevattingsvermogen van de onderzoeker overschrijden. Daarmee is het een belangrijke vraag hoe het grote aantal mogelijke patronen gereduceerd kan worden tot een beperkt aantal interpreteerbare en analytisch bruikbare patronen. In dit artikel wordt hiervoor een methode gepresenteerd. Deze techniek, multiple correspondentie-analyse, is geschikt voor de analyse van twee-dimensionale tabellen met rijen en kolommen, en niet-negatieve getallen in de cellen. Deze twee-dimensionale matrix wordt zo goed mogelijk benaderd door een matrix van een lagere rang (dat wil zeggen, met minder rijen en/of kolommen), waardoor de belangrijkste aspecten van de data in een laag-dimensionele ruimte kunnen worden bestudeerd en een beter inzicht in de gegevens verkregen kan worden. De rijen en kolommen kunnen worden afgebeeld in een grafische representatie van deze laag-dimensionele ruimte. Indien in een dergelijke representatie twee rijen (kolommen) dicht bij elkaar liggen, dan lijken deze sterk op elkaar; in het andere geval zijn de verschillen juist erg groot. Multiple correspondentie-analyse is geschikt voor de analyse van veel verschillende typen data (vergelijk bijvoorbeeld Gifi, 1990); doel van dit artikel is te laten zien hoe MCA specifiek bij de analyse van loopbanen van nut kan zijn. Ook deze kunnen immers worden weergegeven in de vorm van een twee-dimensionele tabel, waarbij elke rij een sequentie van bijvoorbeeld een persoon gedurende een bepaalde periode representeert, en een kolom een combinatie van een toestand waargenomen op een tijdstip – bijvoorbeeld een bepaald beroep waargenomen op tijdstip 2 – weergeeft.

In de volgende paragraaf worden eerst de principes van MCA kort besproken. Daarna wordt een toepassing van deze techniek gepresenteerd op de werk- en schoolloopbanen van jongeren over een periode van 2 jaar; aan de hand hiervan zal dieper worden ingegaan op multiple correspondentie-analyse van loopbanen.

2. CORRESPONDENTIE-ANALYSE VAN LOOPBANEN

Op de rijen van Tabel 1a zijn de loopbanen van 5 personen gegeven. De kolommen geven de 4 mogelijke toestanden weer waarin elk persoon zich kan bevinden, respectievelijk 'onderwijsvolgend', 'werkloos', 'verricht niet bij opleiding passende arbeid', en 'verricht wel bij de opleiding passende arbeid'. De cellen van Tabel 1a geven het aantal weken dat iemand zich in een bepaalde toestand heeft bevonden. De eerste persoon heeft bijvoorbeeld gedurende de twee jaar dat deze is geobserveerd geen onderwijs gevolgd, is in totaal 40 weken werkloos geweest, en heeft achtereenvolgens 12 weken niet en 52 weken wel bij de opleiding passende arbeid verricht.

Om de loopbanen van deze vijf personen in deze twee jaar te vergelijken kan gekeken worden naar de proportie tijd die zij ieder in elk van de vier toestanden hebben doorgebracht. Deze proporties zijn gegeven in Tabel 1b. Te zien is bijvoorbeeld dat de loopbaan van de eerste persoon enigszins lijkt op die van de derde persoon, maar sterk verschilt van die van de vierde persoon.

Door de vijf personen werden in totaal 126 weken in de eerste toestand zijn doorgebracht, 77 in de tweede toestand, 103 in de derde, en 214 in de vierde. De proporties in Tabel 1c geven aan in hoeverre elk van de personen heeft bijgedragen aan elk van de toestanden. Te zien is bijvoorbeeld dat de vierde persoon voor 56% heeft bijgedragen aan de 126 weken van de eerste toestand; de eerste persoon droeg voor 52% bij aan de tweede toestand, enzovoort.

Kleine tabellen zoals Tabel 1a zijn gemakkelijk met het oog te bestuderen, maar dat wordt

*Toon Taris is verbonden aan de Vrije Universiteit, Vakgroep Arbeids- en Organisatiepsychologie, De Boelelaan 1081, 1081 HV Amsterdam, e-mail V70UTOON@HASARA11.
Peter van der Heijden is werkzaam bij de werkgroep Methoden en de Vakgroep Sociologie van de FSW van de Rijksuniversiteit Utrecht.
Adres: T.W. Taris, Vrije Universiteit, Vakgroep Arbeids- en Organisatiepsychologie, De Boelelaan 1081, 1081 HV Amsterdam.

Tabel 1. De loopbanen van 5 hypothetische respondenten gedurende 2 jaar.

	Tabel 1a: aantal weken gespendeerd in diverse toestanden.					Tabel 1b: proportie tijd gespendeerd in diverse toestanden.					Tabel 1c: bijdrage van elk van de respondenten aan elke toestand.				
toestand	f	g	h	i	tot	f	g	h	i	tot	f	g	h	i	tot
resp a	0	40	12	52	104	.00	.38	.12	.50	1.00	.00	.52	.12	.24	.20
resp b	14	10	40	40	104	.13	.10	.38	.38	1.00	.11	.13	.39	.19	.20
resp c	0	0	24	80	104	.00	.00	.23	.77	1.00	.00	.00	.23	.37	.20
resp d	70	17	17	0	104	.67	.16	.16	.00	1.00	.56	.22	.17	.00	.20
resp e	42	10	10	42	104	.40	.10	.10	.40	1.00	.33	.13	.10	.20	.20
tot	126	77	103	214		.24	.15	.20	.41	1.00	1.00	1.00	1.00	1.00	1.00

Nb. f = onderwijsvolgend, g = werkloos, h = verricht niet bij de opleiding passende arbeid, i = verricht wel bij de opleiding passende arbeid.

moeilijker als het aantal personen of het aantal toestanden groter wordt; in dat geval wordt correspondentie-analyse behulpzaam. Stel dat de loopbanen van de vijf personen vergeleken worden. De proporties in Tabel 1b geven dan duidelijk weer hoe een loopbaan van een persoon door de tijd heen is opgebouwd. In correspondentie-analyse worden de proporties van Tabel 1b als coördinaten in een hoogdimensionele ruimte gebruikt. De vijf personen worden elk door 1 punt weergegeven in een in dit geval vier-dimensionele ruimte waarbij de proporties als coördinaten dienen. Zo heeft de eerste persoon de coördinaten (.00, .38, .12, .50), de tweede persoon (.13, .10, .38, .38), enzovoort. De afstanden tussen de punten geven nu een indicatie van de overeenkomst in de desbetreffende loopbanen: personen die dicht bij elkaar liggen volgden loopbanen die op elkaar lijken, en personen die ver uiteen liggen hebben loopbanen die erg verschillen.

In principe kunnen loopbaanproporties het meest verschillen op de toestand waarin in totaal de meeste weken terecht zijn gekomen. In dit voorbeeld is dit toestand 4, waarin 41% van alle weken valt. Hiervoor wordt in correspondentie-analyse gecorrigeerd door de assen van de vier-dimensionele ruimte 'op te rekken', door de eenheidslengte van elke as te vermenigvuldigen met $(1/p_j)^{1/2}$, waarbij p_j de proportie weken is die in deze toestand zitten. Voor de vier toestanden in het voorbeeld zijn deze oprekfactoren respectievelijk (2.03, 2.60, 2.25, 1.56).

Voor een meer formele definitie van de afstanden tussen de loopbanen moet eerst wat notatie ingevoerd worden. De elementen van de matrix in Tabel 1a worden genoteerd als x_{ij}, waarbij i (i=1, ..., I) de rijen indiceert en j (j=1, ..., J) de kolommen. De elementen van de matrix in Tabel 1b worden aangegeven met $p_{ji} \equiv x_{ij}/x_{i+}$, waarbij $x_{i+} = \Sigma_j x_{ij}$. Analoog hieraan worden de elementen van de matrix in Tabel 1c als $p_{ij} \equiv x_{ij}/x_{+j}$ genoteerd. De *chi-kwadraatafstand* $\delta(i,i^*)$ tussen de rij i en i^* in de vier-dimensionele ruimte is gedefinieerd als

$$\delta^2(i,i^*) = \sum_{j=1}^{J} \frac{(p_{ji} - p_{ji^*})^2}{p_j} \qquad (1)$$

Voor elke toestand wordt dus het verschil gekwadrateerd, en gedeeld door p_j. Vervolgens worden al deze verschillen opgeteld. Doordat voor elk paar loopbanen deze chi-kwadraatafstand berekend wordt ontstaat aldus een afstandsmatrix.

Het *zwaartepunt* representeert de 'gemiddelde' loopbaan, en heeft in het voorbeeld de coördinaten (.24, .15, .20, .41); deze zijn gelijk aan de proporties p_j. De afstand tussen rij i en de oorsprong geeft dan aan hoezeer rij i afwijkt van het gemiddelde. Indien rij i en rij i^* in dezelfde richting afwijken van de oorsprong, dan wijkt hun loopbaan op vergelijkbare wijze af van de gemiddelde loopbaan.

De afstanden tussen de punten in de vier-dimensionele ruimte zijn nu helder gedefinieerd. Maar hoe moet deze ruimte bestudeerd worden? Het is niet gemakkelijk om afstanden te bekijken in ruimten met meer dan twee dimensies. Er is dus alle reden toe om te zoeken naar een manier om deze bestudering te vergemakkelijken. In correspondentie-analyse gebeurt dit door de ruimte zo te roteren dat achtereenvolgende dimensies in de geroteerde ruimte zoveel mogelijk laten zien van de afstanden in de vier-dimensionele ruimte. Voor elke volgende dimensie α is het te maximaliseren criterium

$$\lambda_\alpha^2 \equiv \Sigma_i p_i r_{i\alpha}^2 \qquad (2)$$

waarbij $r_{i\alpha}$ de coördinaat van rij i op de nieuwe (d.w.z. geroteerde) dimensie α is. De rijproportie p_i in deze vergelijking geeft aan dat de betergevulde rijen een grotere rol spelen in de bepaling van de uiteindelijke oplossing. Omdat voor het voorbeeld alle punten in een drie-dimensionele deelruimte van de vier-dimensionele ruimte liggen, zijn alleen λ_1^2, λ_2^2 en λ_3^2 groter dan nul. Hun som $\lambda_1^2 + \lambda_2^2 + \lambda_3^2$ wordt *inertie* genoemd. De inertie is een maat voor de totale (gewogen) afstand in de ruimte; $\lambda_1^2 / (\lambda_1^2 + \lambda_2^2 + \lambda_3^2)$ is bijvoorbeeld de proportie van de inertie die is afgebeeld in de eerste dimensie. Voor het voorbeeld zijn de eerste drie waarden λ_α^2, achtereenvolgens .41, .13 en .05, en deze drie dimensies laten respectievelijk 67%, 23% en 9% van de totale inertie zien. Als alleen de eerste twee dimensies bestudeerd worden, dan wordt 91% van de totale inertie bestudeerd.

Een twee-dimensionele afbeelding van de rijpunten is te vinden in figuur 1a. Daar is te zien dat de loopbanen b, c, d en e op een lijn liggen die ongeveer evenwijdig loopt aan de horizontale as van de eerste dimensie, en dat vooral loopbaan a hoog scoort op de tweede dimensie. Loopbaan d lijkt het meest op loopbaan e, en het minst op loopbaan a en c. Om te weten waarom dit zo is, moet de relatie tussen de loopbanen en de toestanden onderzocht worden. Dit gebeurt tegelijkertijd in dezelfde correspondentie-analyse.

Voor de kolommen van de matrix wordt een vergelijkbare procedure als voor de kolommen gevolgd; uitgangspunt zijn nu echter de proporties in Tabel 1c. Elke toestand is als een punt af te beelden in een vijf-dimensionele ruimte, waarbij de proporties per kolom als coördinaten gebruikt worden. De vier toestanden liggen gezamenlijk in een drie-dimensionele deelruimte van de vijf-dimensionele ruimte. Het zwaartepunt heeft als coördinaten het gewogen gemiddelde van de vier toestanden, namelijk (.20, .20, .20, .20, .20). De assen van de vijf-dimensionele ruimte worden opgerekt met een factor $(1/p_i)^{1/2}$. Voor het voorbeeld betekent dit dat elke dimensie op dezelfde wijze wordt opgerekt, omdat p_i = .20 voor elke i. De chi-kwadraatafstand tussen toestand j en j^* wordt analoog aan die voor de afstanden tussen de rijen berekend, en ook de manier van roteren is gelijk aan die voor de rijen.

De oplossing voor de kolompunten is gegeven in figuur 1b. Daar is te zien dat de personen vergelijkbaar scoren op toestand h en i, en nogal verschillend op f en g; deze liggen immers ver van de oorsprong. Een beter begrip van deze oplossing ontstaat door deze oplossing in figuur 1b te relateren aan de oplossing voor de rijen in figuur 1a. De rijcoördinaten $r_{i\alpha}$ en de kolomcoördinaten $c_{j\alpha}$ zijn als volgt gerelateerd:

$$c_{j\alpha} = \lambda_\alpha^{-1} \sum_{i=1}^{I} p_{ji} r_{i\alpha} \qquad (3a)$$

en

$$r_{i\alpha} = \lambda_\alpha^{-1} \sum_{j=1}^{I} p_{ij} c_{j\alpha} \qquad (3b)$$

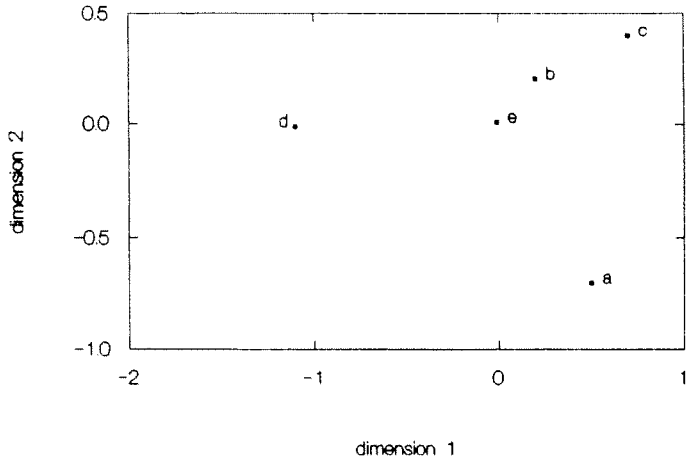


Fig. 1a. Weergave van de rijen in een twee-dimensionele ruimte.

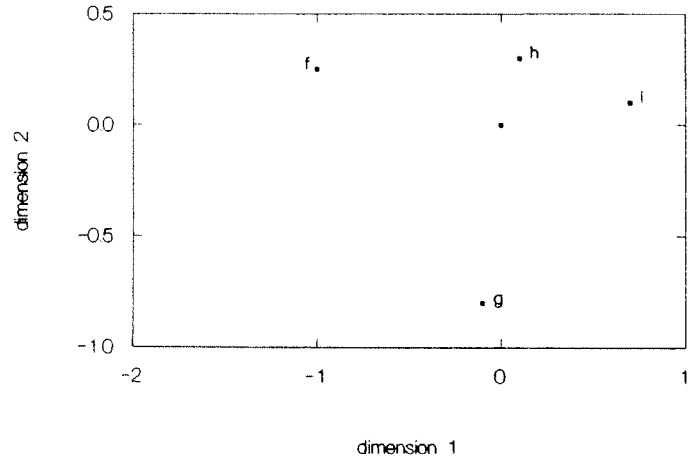


Fig. 1b. Weergave van de kolompunten in een twee-dimensionele ruimte.

Fig. 1. Afbeeldingen van respectievelijk de rij-punten (loopbanen) en kolom-punten (categorieën) van de gegevens in Tabel 1.

De eerste vergelijking laat zien dat, op een schaalfactor λ_α^{-1} na, de kolomcoördinaat $c_{j\alpha}$ voor kolom j op dimensie α te berekenen is als het gewogen gemiddelde van alle rijcoördinaten op dimensie α, waarbij de proporties p_{jli} als gewicht worden gebruikt. Op een schaalfactor na liggen de kolompunten dus in het gewogen gemiddelde van de rijpunten. De tweede vergelijking laat zien dat de rijpunten tegelijkertijd in het gewogen gemiddelde van de kolompunten liggen. Verder is het van belang te weten dat $\Sigma_i\, p_i r_{i\alpha} = 0$ en $\Sigma_j\, p_j c_{j\alpha} = 0$; indien het gemiddelde profiel gebruikt wordt voor het berekenen van een gewogen gemiddelde, dan geeft dit de coördinaten van de oorsprong. De interpretatie van de relatie tussen de oplossing voor de rijpunten

en die voor de kolompunten is nu eenvoudig. Ruwweg geldt dat, indien $p_{jli} > p_j$ (en dan geldt tegelijkertijd dat $p_{ilj} > p_i$, want uit $p_{jli} = p_{ij}/p_i > p_j$ volgt dat $p_{ilj} = p_{ij}/p_j > p_i$), dat rijpunt i dicht bij kolompunt j ligt. Dit is na te gaan in de Figuren 1a en 1b, gecombineerd met de Tabellen 1b en 1c. In Tabel 1b is te zien dat loopbaan a als speciaal kenmerk heeft dat de tweede toestand een veel grotere proportie heeft dan gemiddeld, namelijk .38 versus .15 (en in Tabel 1c is te zien dat .52 veel groter is dan .20). Daarom wijken loopbaan a en b op dezelfde wijze af van de oorsprong voor zover het de tweede dimensie betreft. Op dezelfde manier geldt dat de vierde loopbaan d een veel hogere proportie heeft op toestand 1, namelijk .67 versus .24.

Tot zover deze inleiding in correspondentie-analyse van loopbanen. Resumerend: correspondentie-analyse geeft een grafische representatie van loopbanen zoals die zijn samengevat zoals in Tabel 1a. Deze representatie laat zien in hoeverre loopbanen op elkaar lijken, en op welke punten ze overeenstemmen of juist verschillen. Alhoewel op deze manier al veel inzicht in de overeenkomsten en verschillen tussen loopbanen kan worden verkregen, zullen nu enkele manieren besproken worden om de informatie die aanwezig is in loopbaangegevens nog beter te benutten.

Verfijningen

In Tabel 1a zijn de gegevens van vijf loopbanen in twee jaar tijd samengevat. Indien er een duidelijke volgorde is in de toestanden en vrijwel iedereen dezelfde volgorde doorloopt, dan gaat hier geen informatie verloren. Als de volgorde in de tijd bijvoorbeeld altijd 'onderwijsvolgend', 'werkloos', 'niet bij opleiding passende arbeid', en 'wel bij de opleiding passende arbeid' is, dan kan uit de gegevens van Tabel 1a de loopbaan van elk van de vijf personen exact gereconstrueerd worden. In de praktijk is het verloop van loopbanen echter ingewikkelder; de volgorde van toestanden ligt niet vast, en men kan bijvoorbeeld korte perioden van werkloosheid afwisselen met perioden van passende en niet-passende arbeid. Dit kan niet gereconstrueerd worden aan de hand van Tabel 1a. Deze informatie is echter vaak wel beschikbaar, en in dat geval kan een meer verfijnde tabel dan Tabel 1a worden samengesteld. In Tabel 2 zijn bijvoorbeeld de gegevens van de vijf loopbanen uitgesplitst in twee jaren, die elk sommeren tot 52 weken. Als deze twee deeltabellen opgeteld worden, dan geeft dit Tabel 1a. Op de data in Tabel 2 is een correspondentie-analyse uit te voeren. Dit zal een oplossing opleveren met 5 loopbanen en 2x4 jaar-toestand combinaties.

Dit opsplitsen van de tabel is steeds verder door te voeren; uiteindelijk is het mogelijk een tabel van 5 loopbanen en 104x4 week-toestand combinaties te construeren. Deze matrix van 5x(104x4) wordt wel een super-indicatormatrix genoemd (Gifi, 1990), en correspondentie-analyse van een super-indicatormatrix staat ook wel bekend als *multiple* correspondentie-analy-

Tabel 2. Uitsplitsing van de gegevens in Tabel 1a naar jaar van observatie.

	Jaar 1				jaar 2				
toestand	f	g	h	i	f	g	h	i	totaal
resp a	0	30	12	10	0	10	0	42	104
resp b	14	4	30	4	0	6	10	36	104
resp c	0	0	24	28	0	0	0	52	104
resp d	52	0	0	0	18	17	17	0	104
resp e	42	10	0	0	0	0	10	42	104
	108	44	66	42	18	33	37	172	

Nb. f = onderwijsvolgend, g = werkloos, h = verricht niet bij de opleiding passende arbeid, i = verricht wel bij de opleiding passende arbeid.

se of HOMALS (HOMogeniteitsanalyse via Alternating Least Squares). Correspondentie-analyse van de 5x(104x4)-matrix zal echter nogal instabiele resultaten te zien geven. Het minder verfijnd representeren van de data, door bijvoorbeeld gebruik te maken van de 5x(2x4)-matrix zoals gegeven in Tabel 2, is een manier om tot een stabielere oplossing te komen. In dit licht is het van belang dat de correspondentie-analyse van de matrix van 5x(2x4) gelijk is aan de correspondentie-analyse van de matrix van 5x(104x4) met gelijkheidsrestricties op de toestandsscores, namelijk de restrictie dat toestandsscores voor een bepaalde toestand *binnen* een jaar aan elkaar gelijk zijn (zie Van der Heijden en De Leeuw, 1989).

Ontbrekende gegevens

In de praktijk heeft men vaak te maken met ontbrekende gegevens. De observatie van een loopbaan kan bijvoorbeeld later beginnen of eerder ophouden dan de observatie van andere loopbanen. In panel-onderzoek is onder andere vaak sprake van uitval (sommige personen werken wel aan de eerste meting mee, maar niet aan de volgende metingen); in dat geval is voor deze observaties sprake van een kortere waarnemingsperiode dan voor de overigen. Er zijn verschillende manieren om om te gaan met ontbrekende gegevens bij correspondentie-analyse (overzichten worden gegeven in Meulman, 1982, en Van der Heijden en Escofier, 1988). Hier zullen er twee besproken worden aan de hand van Tabel 1a.

Ten eerste, als bijvoorbeeld van de eerste loopbaan voor minder weken bekend is wat de toestand is - zodat bijvoorbeeld de gegevens respectievelijk 0, 40, 6, 40 zijn - dan analyseert men de matrix waarbij voor deze rij dan ook 0, 40, 6, 40 is ingevuld; er wordt dus geen enkele correctie toegepast. De analyse wordt verder op geen enkele manier aangepast, hoewel deze loopbaan een wat kleiner gewicht p_i zal hebben bij het bepalen van de assenstand. Ook bij het verder verfijnen van de analyse wordt op dezelfde wijze te werk gegaan. Zo wordt bij de superindicatormatrix voor die weken waarvoor geen gegevens bekend zijn (0,0,0,0) ingevuld.

Ten tweede, uitgaande van de gegevens 0, 40, 6, 40 voor de eerste loopbaan zijn er ontbrekende observaties voor 18 weken. Het is dan mogelijk om voor alle loopbanen een extra kolom die het aantal missende observaties aangeeft te definiëren, en vervolgens voor de eerste loopbaan in deze kolom '18' in te vullen. Dit is een goed alternatief indien de ontbrekende gegevens van verschillende loopbanen op dezelfde manier werden veroorzaakt, bijvoorbeeld, zij zijn allen later begonnen of eerder geëindigd. Het is echter duidelijk dat op dit gebied meer onderzoek gewenst is, bijvoorbeeld om te onderzoeken hoe 'linkse censurering', 'rechtse censurering' en andere vormen van ontbrekende gegevens het best behandeld kunnen worden.

Loopbanen als sequenties van toestanden

Bij de hier voorgestelde procedure om loopbanen te kwantificeren moet worden opgemerkt dat de informatie dat bepaalde toestanden volgen op of voorafgaan aan andere niet expliciet in de analyse wordt meegenomen; de kolommen van de te analyseren tabel kunnen met elkaar verwisseld worden, zonder dat hierdoor de uiteindelijke resultaten veranderen. Indien twee individuen zich op alle geobserveerde tijdstippen in dezelfde toestand bevinden - dus: voor alle kolommen dezelfde score hebben - zijn zij identiek. Naarmate er meer verschillen ontstaan zullen de twee loopbanen meer verschillend gekwantificeerd worden. Neem het volgende, tamelijk extreme, voorbeeld:

a b c d e f g h	(1)
b c d e f g h i	(2)
b c d e e f g h	(3)
j k l m n o p q	(4)

De eerste twee sequenties hebben op geen enkel tijdstip een element gemeen. De twee sequenties zijn echter bijna identiek; ze bevatten vrijwel dezelfde elementen in exact dezelfde volgorde. 'Zelfde tijd, zelfde toestand'-technieken zullen deze twee sequenties als verschillend beschouwen; het feit dat de volgorde van de gemeenschappelijke elementen gelijk is telt niet mee.

Deze twee sequenties hebben echter allebei 4 elementen gemeenschappelijk met sequentie 3, en zullen daarom beiden dicht bij 3 – en dus ook dicht bij elkaar – komen te liggen. Sequentie 4 heeft geen enkel element met de 3 voorgaande gemeenschappelijk, zelfs niet indien de volgorde van de elementen meegenomen wordt, en zal daarom ver van de andere 3 elementen worden geplaatst.

Dit voorbeeld toont aan dat het loutere feit alleen dat twee sequenties geen elementen op dezelfde tijdstippen gemeenschappelijk hebben niet hoeft te betekenen dat deze sequenties als zeer verschillend worden gezien; dat hangt in sterke mate af van de andere loopbanen in de data. Als er een grote samenhang is tussen de opeenvolgende categorieën van sequenties (zoals in de eerste drie sequenties van ons voorbeeld), heeft dit tot gevolg dat deze sequenties als ongeveer gelijk zullen worden gekwantificeerd. Met andere woorden, het feit dat er een bepaalde volgorde in de elementen van de sequenties zit heeft wel degelijk invloed op de resultaten.

Tenslotte

Correspondentie-analyse van loopbanen, of meer in het algemeen van tijdbudgetten en eventhistories werd voor het eerste als zodanig voorgesteld door Saporta (1981, 1985) en Deville en Saporta (1980, 1983). Zij behandelen dit onderwerp als een vorm van harmonische analyse voor nominale tijdreeksen, en benadrukken geometrische representaties van dit type gegevens. De Leeuw, Van der Heijden en Kreft (1985) en de Leeuw en Kreft (1985) benadrukken eerder de interpretatie van homogeniteitsanalyse en de mogelijkheid die deze interpretatie biedt om te komen tot een kwantificatie van loopbanen. Dit is vooral zinvol indien de afstanden tussen de loopbanen zonder al te veel verlies van informatie in 1 dimensie afgebeeld kunnen worden.

3. TOEPASSING: MULTIPLE CORRESPONDENTIE-ANALYSE VAN SCHOOL/WERKLOOPBANEN

Vraagstelling en opzet

In het tweede deel van dit paper zal correspondentie-analyse gebruikt worden voor de analyse van de school/werkloopbanen van 265 Nederlandse jongeren gedurende de periode 1987-1989. De gegevens werden eerder geanalyseerd door Claes, Quintanilla & Whitely (1992), wier onderzoek als doel had te beschrijven via welke paden mensen in een bepaalde toestand (een vaste baan in hetzij de administratieve, hetzij de metaalsector) terechtkomen.

De rest van deze sectie is gesplitst in drie delen. In het eerste deel worden kort de gebruikte gegevens en de gevolgde procedure beschreven. In het tweede deel komen de mogelijkheden van correspondentie-analyse bij de analyse van loopbanen aan de orde. Getoond zal worden hoe, op basis van afbeeldingen van categorieën van variabelen en personen, inzicht in de data verkregen kan worden. Toetsingen van hypothesen over de structuur van de data blijven hier echter achterwege. In het derde deel wordt eerst op basis van de resultaten van de correspondentie-analyse een classificatie van loopbanen geconstrueerd. Daarna wordt nagegaan hoe binnen elk van deze clusters de loopbaan verloopt.

Methode: data en procedure

De gegevens werden gedurende de herfst van 1989 verzameld in het kader van een internationaal onderzoek betreffende de werksocialisatie van jongeren (WOSY International Research Group, 1989; Whitely, Peiró & Sarchielli, 1992). In dit artikel worden alleen de Nederlandse data gebruikt. De te analyseren steekproef bestaat uit 265 mensen die ten tijde van het interview ongeveer 6 maanden in hun toenmalige (eerste) baan werkzaam waren. De gemiddelde leeftijd van de respondenten bedroeg 19,7 jaar. De helft van de onderzoeksgroep was werkzaam in beroepen waarbij in principe intensief met computers omgegaan wordt (administratief medewerkers, leerling-programmeurs en dergelijke). De andere helft bestond uit personen werkzaam in de metaal (lassers, bankwerkers en dergelijke).

Gedurende het interview werd informatie verzameld over de meningen, waarden en het gedrag ten aanzien van werk, alsmede over de school/werkloopbaan over de 2 jaar voorafgaande aan het interview. Voor ieder kwartaal moest worden aangegeven wat men hoofdzakelijk had gedaan, wat in totaal dus 8 tijdspunten oplevert. De mogelijke categorieën op deze tijdspunten waren: 'tijdelijke baan', 'vaste baan', 'in opleiding in verband met huidige baan', 'in opleiding, niet in verband met huidige baan', 'werkloos', 'in militaire dienst', en 'anders' (bijvoorbeeld ziek). Van de 265 personen hadden er 23 missende waarden op één of meer tijdspunten. Hiervoor werd op de betreffende variabelen de waarde 0 ('ontbrekend') ingevuld. De 8 tijdstippen waren daarmee te coderen in een super-indicatormatrix, die met behulp van multiple correspondentie-analyse (via het SPSS-programma HOMALS) werd geanalyseerd. De eerste 5 eigenwaarden van de oplossing bedroegen respectievelijk .60, .52, .41, .28, en .25. Op basis van het 'knik'-criterium (Cattell, 1966) werd de analyse tot de eerste 3 dimensies beperkt.

Resultaten

Vervolgens werd bekeken of de categorieën van de variabelen (in de kolommen) verschillend gerelateerd zijn aan de rijen (de loopbanen). Een afbeelding, omwille van de interpreteerbaarheid beperkt tot slechts 2 dimensies, van de lokatie van de categorieën van de variabelen is gegeven in figuur 2.

Uit deze figuur blijkt dat enkele categorieën sterk onderscheiden tussen de kolommen, terwijl andere categorieën aanzienlijk minder belangrijk zijn. Globaal kan gezegd worden dat de latere categorieën (gemeten op tijdstip 7 en 8) dichter bij het nulpunt liggen dan de overige categorieën. Dit geeft aan dat deze twee tijdstippen minder belangrijk zijn in het maken van onderscheid tussen de rijen. Omdat de steekproef bestaat uit mensen die in principe 6 maanden in de huidige baan werkzaam waren, is dit geen verrassend resultaat.

Verder zijn er ook enkele clusters van categorieën te zien. Het hebben van een vaste baan op de tijdstippen 1 tot en met 6 ligt bijvoorbeeld links onder in Figuur 2, vlak bij elkaar; dit geeft aan dat diegenen die op tijdstip 1 een vaste baan hebben, op tijdstip 2 tot en met 6 zeer waarschijnlijk ook nog een vaste baan zullen hebben. Het in opleiding zijn, maar niet in verband met de huidige baan, ligt voor alle tijdstippen ongeveer bovenin Figuur 2; het in opleiding zijn in verband met huidige baan ligt rechts van het nulpunt; en het hebben van een tijdelijke baan ligt ruwweg in het midden van Figuur 2. Het in militaire dienst zijn, het werkloos zijn en de categorie 'anders' liggen voor de verschillende tijdstippen relatief ongestructureerd door elkaar, in de rechterbovenhoek van de figuur. Hieruit kan geconcludeerd worden dat men relatief gemakkelijk wisselt tussen deze drie toestanden.

In Figuur 3 is de configuratie van de rij-punten (loopbanen) gegeven. Voor onze doeleinden is dit de belangrijkste figuur, omdat hieruit afgelezen kan worden welke personen een vergelijkbare loopbaan hebben gevolgd.

Figuur 3 laat zien dat er enkele clusters van rij-punten kunnen worden onderscheiden. Een groot aantal rij-punten is geconcentreerd in de rechterhelft van Figuur 3, wat aangeeft dat deze loopbanen sterk op elkaar lijken. Daarnaast kan een klein cluster linksonder in Figuur 3 onderscheiden worden; deze punten liggen zeer dicht bij elkaar, maar ver uit de buurt van het nulpunt en de meeste andere punten. Bovenin is ook een cluster zichtbaar, dat weliswaar intern niet zo homogeen is als het cluster rechts, maar ook niet bij het grote cluster lijkt te horen. Tussen deze relatief sterke clusters zweven nog enkele andere punten. De bij deze punten behorende mensen hebben blijkbaar weinig voorkomende en onderling sterk verschillende loopbanen gevolgd.

Classificatie van loopbanen

Globaal geeft Figuur 3 dus de indruk dat het zinvol zou kunnen zijn om zo'n 4 clusters te onderscheiden. Waar precies de grenzen van deze clusters liggen is echter niet helemaal duidelijk; het is niet mogelijk om deze clusters nu reeds te onderscheiden. Daarom is eerst op basis van de via correspondentie-analyse verkregen waarden voor de personen op de eerste drie dimensies een cluster-analyse uitgevoerd, zodat er besloten kan worden welke punten nog wel, en welke punten niet in een bepaald cluster thuishoren.[1] Uiteindelijk werden er op basis van een

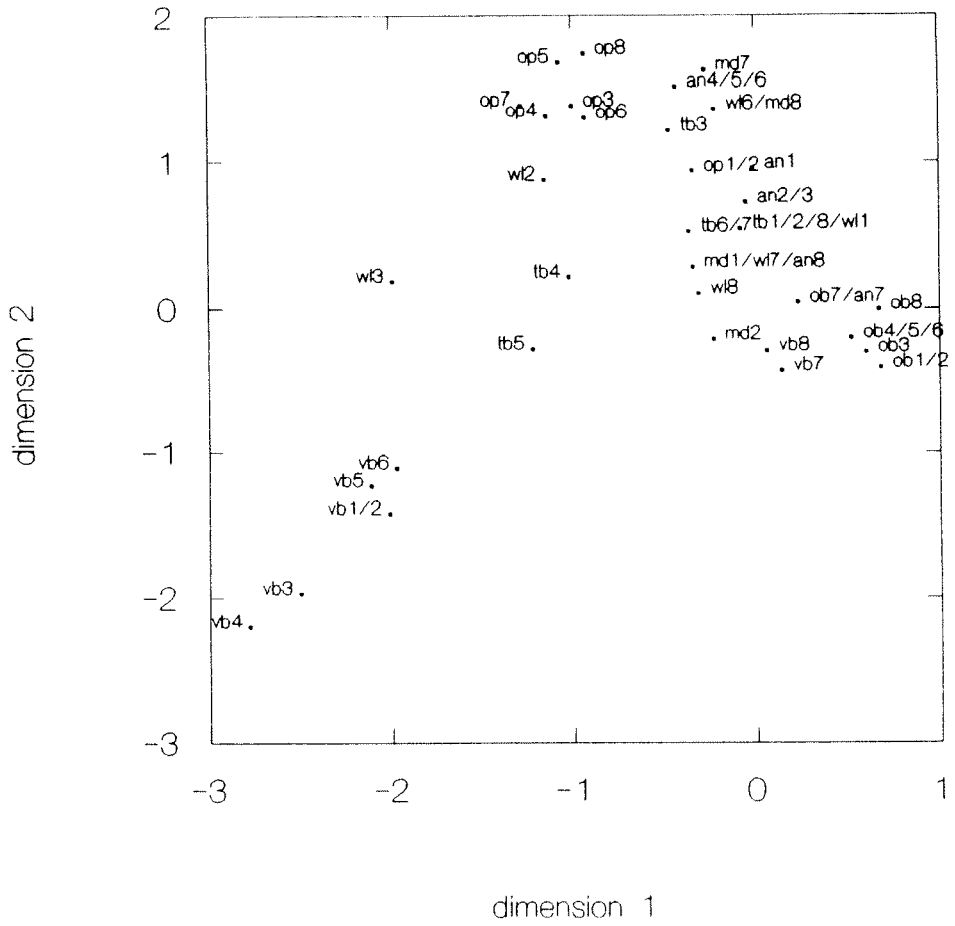


Fig. 2. Afbeelding van de categorieën van de variabelen in een twee-dimensionale ruimte, op basis van via MCA verkregen schaalwaarden.
Nb. vbi = vaste baan op tijdstip i, op = in opleiding (niet in verband met baan), ob = in opleiding (in verband met baan), md = in militaire dienst, wl = werkloos, vb = vaste baan, tb = tijdelijke baan, an = anders. Zeer dicht bij elkaar liggende punten zijn ter wille van de leesbaarheid samengevoegd.

nearest-neighbour cluster-analyse van de gekwadrateerde euclidische afstanden zoals berekend op basis van de eerste drie HOMALS-dimensies 4 verschillende clusters van loopbanen geconstrueerd, die zowel empirisch – op basis van de toename van de error sum of squares in de cluster-analyse – als inhoudelijk van elkaar te onderscheiden leken te kunnen worden. In Figuur 3 is ruwweg aangegeven welke mensen uiteindelijk in een bepaald cluster ingedeeld werden. In Figuur 4 is de indeling in clusters inzichtelijk gemaakt, door in diagrammen voor elk van de clusters aan te geven wat de personen op de verschillende tijdstippen deden. Tabel 3 tenslotte geeft de gemiddelden en standaardafwijkingen van de 4 clusters op elk van de 3 MCA-dimensies.

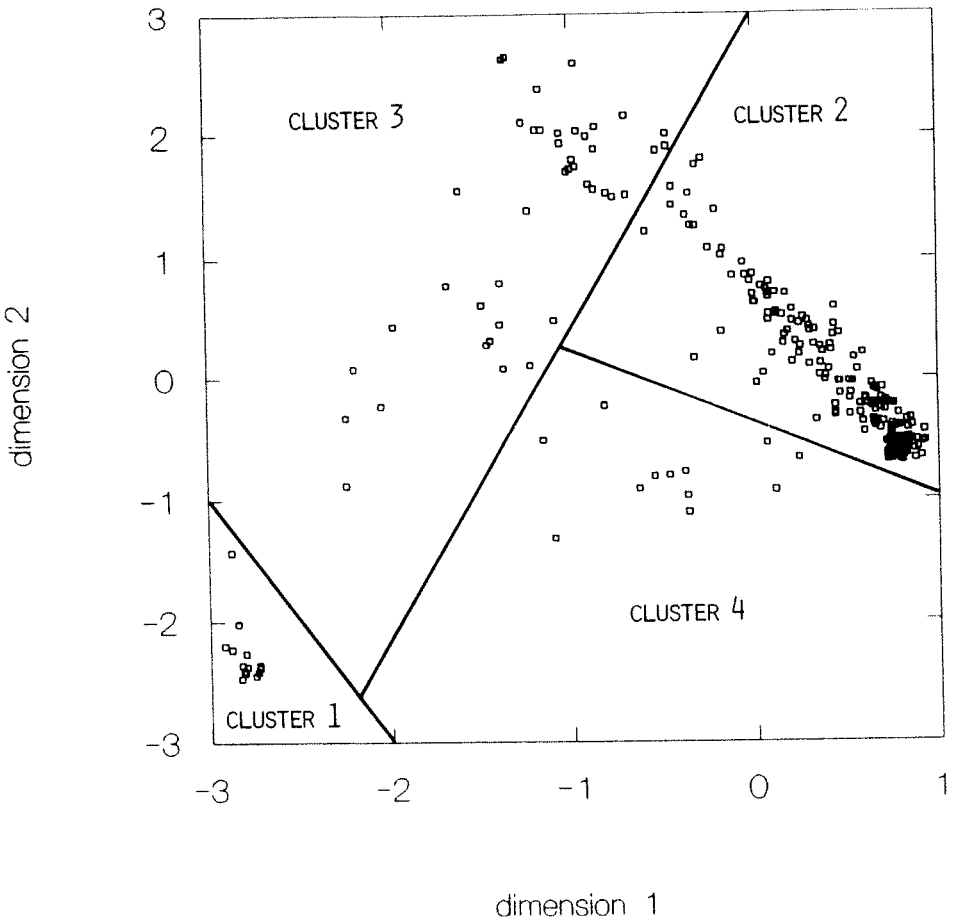


Fig. 3: Afbeelding van de 265 rijpunten (loopbanen) in een twee-dimensionale ruimte, op basis van de via MCA verkregen schaalwaarden.

Tabel 3 bevestigt globaal de indruk van figuur 3. Het is te zien dat de clusters wat betreft hun gemiddelden op de eerste drie MCA-dimensies vooral verschillen op de eerste twee dimensies, terwijl de derde dimensie eigenlijk weinig onderscheid maakt tussen de groepen. Met betrekking tot de standaardafwijkingen kan opgemerkt worden dat cluster 1 een relatief zeer homogeen cluster betreft. Met name cluster 3 is intern tamelijk los, zoals blijkt uit de standaardafwijkingen die op alle drie de dimensies relatief erg groot zijn.

In Figuur 4a zijn de marginale verdelingen van alle personen op de verschillende tijdstippen weergegeven. Op tijdstip 1 (T1) zijn bijvoorbeeld 38 mensen werkzaam in een vaste baan; 125 mensen bevinden zich in opleiding, in verband met hun huidige baan; een andere groot deel (83 personen) is wel in opleiding, maar niet in verband met hun huidige baan; de overige categorieën zijn klein en onbetekenend. Op tijdstip 8 (T8) heeft het grootste deel van de mensen (156 personen) een vaste baan, een ander deel (46 personen) een tijdelijke baan; de andere categorieën zijn weer relatief klein. De globale indruk van deze figuur is dat de steekproef overgaat van

Tabel 3. Gemiddelden en standaardafwijkingen van de vier onderscheiden clusters op de eerste drie MCA-dimensies.

	dimensie 1		dimensie 2		dimensie 3	
	gem.	SD	gem.	SD	gem.	SD
cluster 1	-2.925	.019	-2.299	.115	.068	.065
cluster 2	.508	.257	-.120	.432	-.010	.522
cluster 3	-1.199	.584	1.509	1.011	-.025	2.230
cluster 4	-.438	.367	-.801	.252	-.006	.249

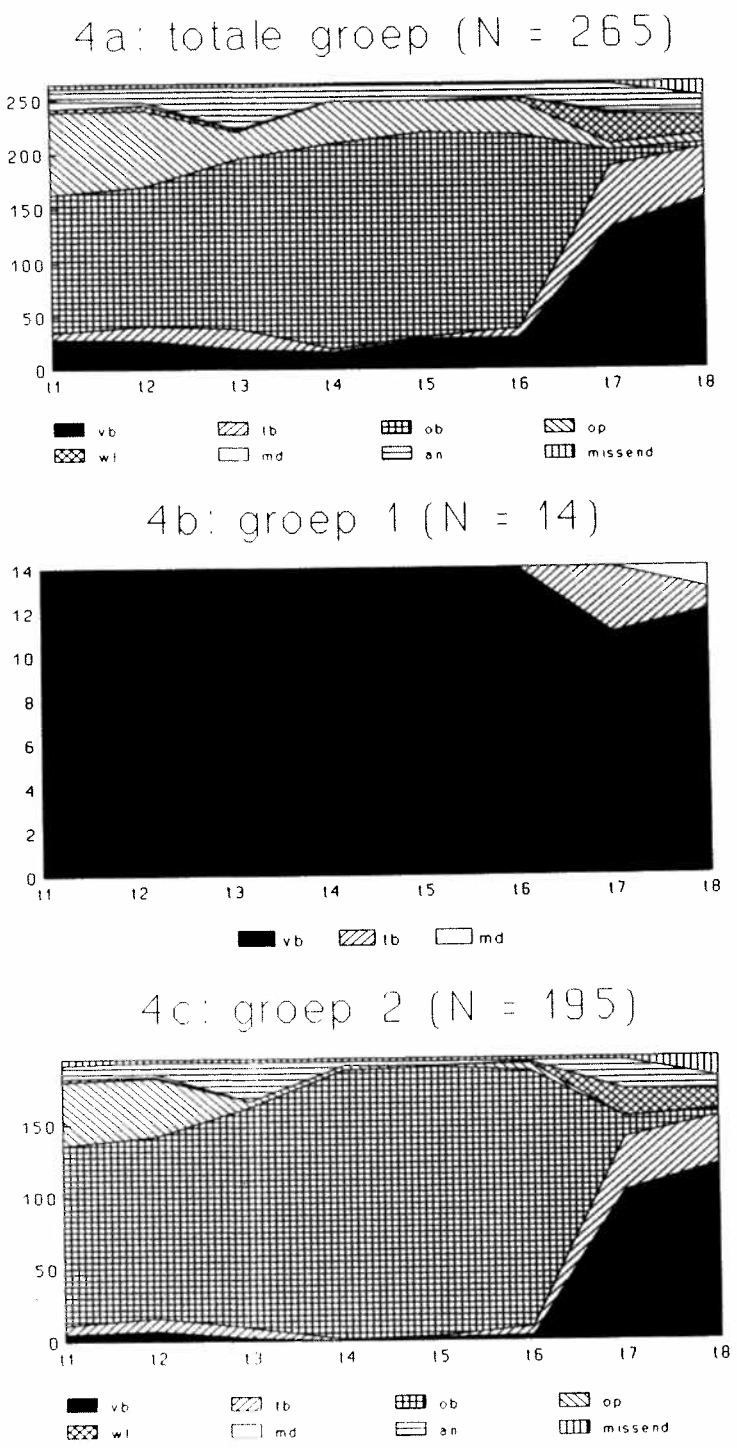
'in opleiding' naar 'werkend'. Figuur 4a geeft een weergave van de toestand van de totale groep op elk van de verschillende tijdstippen; dit is in feite te beschouwen als een representatie van de marginaal van de te analyseren kruistabel (zie de paragraaf 'verfijningen').

Deze grote groep personen werd op basis van de clusteranalyse opgedeeld in 4 verschillende groepen. Er is een kleine groep die gedurende de geobserveerde periode vrijwel constant een vaste baan had (figuur 4b); een tweede, zeer grote groep (N = 195, figuur 4c) blijkt verantwoordelijk voor de in het totaalbeeld optredende indruk van overgang van opleiding (in verband met de huidige baan) naar werk. Dit is het patroon dat gezien de opzet van het onderzoek verwacht mocht worden. Veel van de personen in deze steekproef komen immers via het leerlingstelsel aan hun baan. Er is een derde groep van redelijke omvang (N = 44, figuur 4d), die van groep 2 onderscheiden kan worden doordat ze in de middelste perioden in opleiding, *niet* in verband met de huidige baan zijn. Tenslotte is er nog een klein groepje (N = 12, figuur 4e) dat zich van de rest onderscheidt doordat zij relatief vaak in de categorie 'anders' vallen. Vermoed wordt dat het hier gaat om mensen die bijvoorbeeld langdurig ziek zijn geweest. De totale groep van 265 personen is dus op te delen in een grote groep, die van opleiding (in verband met de huidige baan) naar een - al dan niet vaste - baan gaat; daarnaast zijn er 3 andere, kwantitatief minder belangrijke patronen te onderscheiden.

4. SAMENVATTING, CONCLUSIES EN DISCUSSIE: CORRESPONDENTIE-ANALYSE VOOR DE ANALYSE VAN LOOPBANEN?

In dit artikel werd een techniek beschreven die gebruikt kan worden voor de analyse van loopbanen. Allereerst werden de principes van deze techniek, multiple correspondentie-analyse, uiteen gezet. Multiple correspondentie-analyse is geschikt voor de analyse van nominale variabelen die weergegeven zijn in een tweedimensionale kruistabel. Sequenties van toestandsovergangen (loopbanen) kunnen ook in een dergelijke matrix gecodeerd worden. Vervolgens werd, met behulp van gegevens over de opleiding- en werkloopbaan van 265 Nederlandse jongeren, getoond hoe correspondentie-analyse gebruikt kan worden om inzicht te verkrijgen in het verloop van loopbanen. Allereerst werden op basis van via correspondentie-analyse verkregen afstanden tussen categorieën en personen twee-dimensionale afbeeldingen geconstrueerd. Daarna werd, op basis van de via correspondentie-analyse verkregen afstanden tussen personen, een clusteranalyse uitgevoerd, die leidde tot een variabele 'loopbaantype'. Deze kan in vervolganalyses gebruikt worden als te verklaren, dan wel als verklarende variabele.

Uit deze exercities bleek dat via correspondentie-analyse inzicht verkregen kan worden in de structuur van sequentiële data. Duidelijk werd welke variabelen en categorieën het meest bijdroegen aan het onderscheid tussen de loopbanen van de personen. Geconcludeerd kan worden dat correspondentie-analyse ook voor de in loopbanen geïnteresseerde onderzoeker een interes-



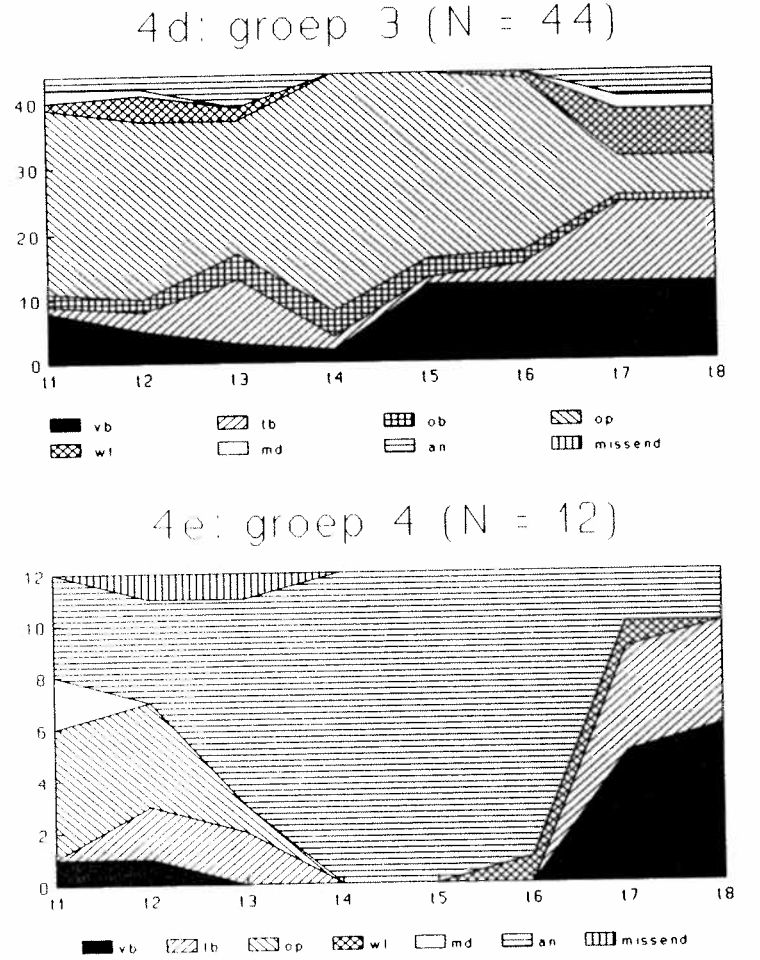


Fig. 4. De loopbaanpatronen van respectievelijk de totale steekproef (4a) en de 4 via MCA daaruit afgesplitste deelgroepen (4b-e). Nb. vb = vaste baan, tb = tijdelijke baan, wl = werkloos, op = opleiding, niet in verband met baan, ob = opleiding in verband met baan, md = in militaire dienst, an = overig, missend = geen score voor desbetreffende periode bekend.

sante techniek kan zijn, waarbij vooral veel inzicht verkregen kan worden in de mate waarin bepaalde variabelen/categorieën van variabelen bijdragen aan het onderscheid tussen personen.

Er zijn ook enkele nadelen te noemen van de correspondentie-analyse van loopbaangegevens. Allereerst is er het probleem van de wijze waarop er in MCA met missende gegevens omgegaan wordt. Als de score op een bepaald tijdspunt voor een bepaalde persoon ontbreekt, kan er een aantal oplossingen getroffen worden; het is echter niet duidelijk in hoeverre deze de resultaten systematisch beïnvloeden. Dit probleem is voor het besproken type data tamelijk belangrijk omdat de gegevens vaak verzameld worden via een panel-design, waarbij de infor-

matie over tussenliggende perioden meestal via retrospectieve vragen verzameld zal worden. Bij een grotere periode tussen de metingen van het panel zal het voor deze persoon echter steeds moeilijker worden voor alle tijdspunten aan te geven, wat hij/zij op het betreffende punt deed. Dat kan als gevolg hebben dat er relatief veel missende waarden in zulke datasets zullen voorkomen. Er zijn meerdere oplossingen voor dit probleem denkbaar, maar het is nog niet duidelijk in hoeverre deze de analyses systematisch beïnvloeden. Hierbij moet opgemerkt worden dat dit type problemen ook voor andere technieken speelt, en zeker niet uniek voor MCA is.

Ten tweede het ordeningsprincipe: er worden niet zozeer *loopbanen* met elkaar vergeleken worden, maar alleen *toestanden* op specifieke tijdspunten. De kolommen van de te analyseren tabel kunnen met elkaar verwisseld worden zonder dat de resultaten veranderen; de informatie over de volgorde van toestanden wordt dus strikt genomen niet meegenomen. Bij de bespreking van de toepassing bleek overigens dat het sequentiële karakter van de gegevens wel degelijk de resultaten zal beïnvloeden: over het algemeen zal gelden dat dezelfde categorieën op verschillende tijdstippen dichter bij elkaar zullen liggen dan verschillende categorieën op dezelfde tijdstippen, en zullen gelijksoortige sequenties – zelfs zonder dat ze elementen op bepaalde tijdstippen gemeenschappelijk hebben – als ongeveer gelijk worden gekwantificeerd.

Het is de vraag in hoeverre deze kwestie potentiële gebruikers ervan moet weerhouden om MCA te gebruiken bij de analyse van hun gegevens. Zoals gezegd is MCA zeer wel in staat verschillende loopbaantypen te onderscheiden; de in dit paper gegeven toepassing illustreert deze uitspraak. Daarnaast zijn er nauwelijks praktisch toepasbare alternatieven die dit nadeel niet hebben. Een andere 'zelfde tijd, zelfde toestand'-technieken als directe clustering van de toestanden via dummy-variabelen (voor het voorbeeld hierboven zouden dat 56 variabelen worden: 7 dimensies voor de toestanden * 8 tijdstippen) betrekt evenmin de volgorde van toestanden expliciet in de analyse, en geeft aanzienlijk minder inzicht in de structuur van de gegevens. De door Abbott & Hrycak (1990) voorgestelde 'optimal matching'-procedure (die met name de volgorde van toestanden vergelijkt om tot afstanden tussen loopbanen te komen) lijkt vooralsnog niet veel meer dan een aansprekend idee. Hier ontbreekt echter de ruimte om in te gaan op de voor- en nadelen van deze alternatieve benaderingen; voor een uitgebreid overzicht van deze en andere benaderingen van de analyse van loopbaangegevens wordt verwezen naar Taris (in voorbereiding).

Als laatste nadeel moet nog genoemd worden dat (multiple) correspondentie-analyse een tamelijk exploratieve methode is; er worden geen betrouwbaarheidsintervallen met betrekking tot de locatie van de rij- dan wel kolompunten geconstrueerd. Inzicht in de stabiliteit van de oplossing kan echter verkregen worden via bootstrap- of jackknife-analyses (Efron, 1981; Gifi, 1990).

DANKWOORD

Deze publicatie is gebaseerd op data die verzameld werden in het kader van een internationaal onderzoek betreffende de werksocialisatie van jongeren (WOSY). Het Nederlandse WOSY-project maakt deel uit van het onderzoeksprogramma 'Het proces van socialisatie van jong-volwassenen'. Dit onderzoeksprogramma wordt uitgevoerd aan de Vrije Universiteit te Amsterdam door de vakgroepen Methoden en Technieken van Sociaal-Wetenschappelijk Onderzoek (faculteit SCW) en de vakgroep Arbeids- en Organisatiepsychologie (faculteit FPP). De auteurs danken de leden van de internationale WOSY-projectgroep voor het mogen gebruiken van hun data, Jan van Gastel voor zijn aandeel in de verzameling van de Nederlandse data, en twee anonieme beoordelaars van dit tijdschrift voor hun opbouwende kritiek op een eerdere versie van dit stuk.

NOOT

1. In Figuur 2 en 3 werden de lokatie van de rij- en kolompunten op de eerste 2 dimensies getoond. Omdat de derde dimensie gezien diens eigenwaarde wel degelijk van belang is werd besloten ook deze bij de clusteranalyse te betrekken. Opgemerkt kan nog worden dat het empirisch weinig bleek uit te maken of de oplossing op basis van 2, dan wel 3 dimensies werd uitgevoerd. Dit is niet zo heel verwonderlijk gezien het feit dat elke volgende dimensie minder informatie bevat dan de voorgaande; zie ook Tabel 3.

LITERATUUR

Abbott, A. & Hrycak, A. (1990). Measuring resemblance in sequence data: An optimal matching analysis of musicians' careers. *American Journal of Sociology, 96,* 144-185.

Allison, P.D. (1984). *Event history analysis: Regression for the social sciences.* Beverly Hills: Sage.

Blossfeld, H.P., Hamerle, A. & Mayer, K.U. (1989). *Event history analysis: Statistical theory and application in the social sciences.* New York: Lawrence Erlbaum.

Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioural Research, 1,* 245-276.

Claes, R., Quintanilla, A.R. & Whitely, W. (1992). Career preparation patterns. *Revue Internationale du Psychologie Sociale, 5, 1,* 37-60.

Deville, J.-C. & Saporta, G. (1980). Analyse harmonique qualitative. In E. Diday (ed.), *Data analysis and informatics.* Amsterdam: North-Holland.

Deville, J.-C. & Saporta, G. (1983). Correspondence analysis, with an extension towards nominal time series. *Journal of Econometrics, 22,* 169-189.

Efron, B. (1981). Nonparametric estimates of standard error: the jackknife, the bootstrap, and other methods. *Biometrika, 68,* 589-599.

Gifi, A. (1990). *Non-linear multivariate analysis.* New York: John Wiley & Sons.

Heijden, P.G.M. van der & Escofier, B. (1988). *Multiple corresponding analysis with missing data.* Rennes: Institut National de Recherche en Informatique et en Automatique.

Heijden, P.G.M. van der & Leeuw, J. de (1989). Correspondence analysis, with special attention to the analysis of panel data and event history data. In C.C. Clogg (ed.), *Sociological Methodology 1989.* Oxford: Basil Blackwell.

Leeuw, J. de, Heijden, P.G.M. van der, & Kreft, I. (1985). Homogeneity analysis of event history data. *Methods of Operations Research, 50,* 299-316.

Leeuw, J. de & Kreft, I. (1985). *De definitie en kwantifikatie van de schoolloopbaan.* Leiden: Vakgroep Datatheorie, RR-85-20.

Meulman, J. (1982). *Homogeneity analysis of incomplete data.* Leiden: DSWO-Press.

Van de Pol, F.J.R. (1989). *Issues of design and analysis of panels.* Amsterdam: Sociometric Research Foundation.

Saporta, G. (1981). *Méthodes exploratoires d'analyse de données temporelles.* Ongepubliceerde doctoraalscriptie. Parijs: Université P. et M. Curie.

Saporta, G. (1985). Data analysis for numerical and categorical individual time-series. *Applied Stochastic Models and Data Analysis, 1,* 109-119.

Taris, T.W. (in voorbereiding). *Analysis of career data from a socialization perspective.* Amsterdam, vakgroep Arbeids- en Organisatiepsychologie (academisch proefschrift).

Tuma, N.B. & Hannan, M.T. (1984). *Social dynamics.* New York: Lawrence Erlbaum.

Whitely, W., Peiró, J.M. & Sarchielli, G. (1992). WOSY theoretical framework, research methodology and potential implications. *Revue Internationale du Psychologie Sociale, 5, 1,* 9-36.

WOSY International Research Group (1989). Socialización del joven: Un estudio transnational (Work socialization of youth: A transnational study). Papeles del Psicológico, 39/40, 32-35.

Manuscript ontvangen 27-1-1993
Definitieve versie ontvangen 30-3-1993