Marktonderzoek kwartaalschrift

1971-2/3

Multivariate analyse

14

Marktonderzoek kwartaalschrift

Orgaan van de Nederlandse Vereniging van Marktonderzoekers
Secretariaat: p/a Organisatie-Bureau Wissenraet
Van Eeghenstraat 86, Amsterdam-z. Tel. 020-739551

4e jaargang, nummer 2/3, september/oktober 1971

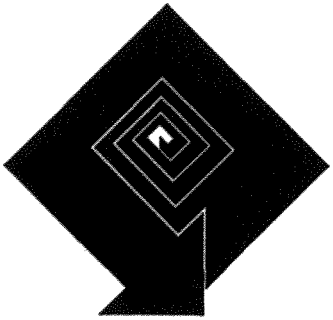
Redactie F. Candèl
M.T.G. Meulenberg
G.J. Randoe
G.M. van Veldhoven
A. van der Zwan

Redactie-secretariaat p/a Universitaire Pers Rotterdam
Heemraadssingel 112, Rotterdam

Uitgave, administratie en advertentie-exploitatie Universitaire Pers Rotterdam
Heemraadssingel 112
Rotterdam-3
Tel. 010-235563
Bankrel.: Pierson, Heldring & Pierson/Rotterdam

Abonnementen *f* 26,— (Bfr. 360) per jaargang
studenten *f* 13,— (Bfr. 180)
losse nummers *f* 7,50 (Bfr. 100)
incl. BTW

Auteursrecht voorbehouden



Van de redactie

Marktonderzoek richt zich vaak op het vinden van samenhangen tussen een groot aantal variabelen, waarover gegevens beschikbaar zijn. Multivariate statistische onderzoektechnieken zijn hierbij een belangrijk hulpmiddel. Hun betekenis groeit als gevolg van de toenemende belangstelling voor de gedragswetenschappelijke aspekten van het marktgebeuren. Het ligt derhalve in de rede om aan deze technieken speciale aandacht te besteden.

In dit nummer wordt zowel op de mogelijkheden als op de beperkingen van diverse multivariate statistische onderzoektechnieken ingegaan. Roskam geeft een zeer instruktief overzicht van een aantal multivariate technieken. Corsten wijdt een kritische beschouwing aan de toepassing van faktoranalyse. Randoe, Van Kooten en Bijnen laten zien hoe multivariate statistische technieken in het marktonderzoek kunnen worden toegepast.

De redactie hoopt dat de bijdragen in dit nummer de belangstelling voor en de verantwoorde toepassing van multivariate statistische technieken in het marktonderzoek zullen bevorderen.

M. T. G. Meulenberg

Inhoud

1 Van de redactie

2 Inhoud

3 Een overzicht van multivariate analysemethoden/E. E. Ch. I. Roskam

27 Wat is factoranalyse?/L. C. A. Corsten

38 Cluster-analyse/E. J. Bijnen

62 Vingeroefeningen in klassificatietechniek. Een toepassing van taxonometrische klassificatiemethoden op enquêteresultaten/ G. J. Randoe en T. H. M. van Kooten

97 Bibliografie multivariate-analyse

102 Literatuurbesprekingen
Research for marketing decisions (P. E. Green and D. S. Tull)/M. T. G. Meulenberg
Multivariate data analysis (W. W. Cooley and P. R. Lohnes)/E. E. Ch. I. Roskam

107 Uit de marktonderzoekwereld

109 Uit de buitenlandse tijdschriften

113 Medewerkers

Een overzicht van multivariate analysemethoden

E. E. Ch. I. Roskam

Een object van onderzoek moet meestal gekarakteriseerd worden door meer dan één meetbaar aspect.

Een aspect is een kenmerk dat verschillende objecten in verschillende mate bezitten. Ook als men in één bepaald aspect is geïnteresseerd, zal het vaak nuttig zijn dit aspect langs verschillende wegen te benaderen, en dat wil zeggen dat men meerdere aspecten observeert om daaruit het éne belangrijke af te leiden. Vaak ook tracht men uit een groot aantal kenmerken een kleiner aantal relevante kenmerken af te zonderen, of wil men weten welke aspecten onder bepaalde natuurlijke of experimentele condities verschuivingen te zien geven.

Wat hier is aangeduid met het woord aspect wordt gewoonlijk variabele of dimensie genoemd. De blik is gericht op deze aspecten, niet op de objecten die de drager daarvan zijn, ook al kan men (groepen van) objecten karakteriseren naar de (gemiddelde) verschillen die zij in de bepaalde aspecten te zien geven.

De samenhang tussen twee aspecten wordt uitgedrukt in hun *covariantie* of *correlatie*. Hebben zij een grote correlatie, dan kan men concluderen dat zij beide herleidbaar zijn tot een gemeenschappelijke factor.

Zo'n factor kan men het beste opvatten als een latent aspect, niet zichtbaar in de feitelijke geobserveerde aspecten, maar daaruit afleidbaar. In een groot aantal aspecten kunnen wij nu trachten een eenvoudig en inzichtelijk patroon te vinden, dat ons antwoord kan geven op specifieke vragen. De vraagstelling zal moeten zijn geformuleerd in een vorm die aansluit bij de empirische gegevens, en leidt tot bepaalde systematische bewerkingen daarvan.

In dit artikel willen wij een overzicht geven van de voornaamste multivariate analysetechnieken:

hoofdassen- en factoranalyse;
multiple en partiële regressie; canonische correlatie;
discriminant-analyse en multivariate variantie-analyse.

Ter inleiding daarop geven we eerst enkele algemene begrippen uit de multivariate analysetechniek en de daarbij behorende algebraïsche notatie.

I. Inleiding

Ter aanduiding van variabelen, objecten en condities gebruiken we de volgende algemene en abstracte notatie:

$Q = \{q_j, j = 1, \ldots, m\}$	een verzameling kenmerken of aspecten
$P = \{p_i, i = 1, \ldots, n\}$	een verzameling (steekproef) objecten, individuen, of stimuli
$C = \{c_h, h = 1, \ldots, g\}$	een verzameling experimentele of natuurlijke condities

Voorts duiden we met x_{ijh} aan de getalscore (op een of andere schaal) van, zeg, persoon p_i op aspect q_j onder conditie c_h.

We zullen in eerste instantie afzien van verschillen in condities C; m.a.w. we nemen aan dat alle scores zijn verkregen onder, in principe, één en dezelfde conditie. We hebben dus bij elke p_i een reeks getallen x_{ij}, $j = 1, \ldots, m$; de getallen in deze reeks hebben een vaste volgorde, overeenkomstig de volgorde der gemeten kenmerken. Zo kunnen we een persoon karakteriseren met zijn scores op, achtereenvolgens, een intelligentietest (q_1), een attitudeschaal (q_2), een inkomstenschaal (q_3), een opleidingsschaal (q_4), etc. Kortheidshalve duiden we de reeks scores aan met x_i.

In de vector-algebra is men gewoon zo'n reeks als kolom (van boven naar beneden) te schrijven; schrijft men hem als rij (van links naar rechts) dan voegen we een $'$ toe: x_i'.

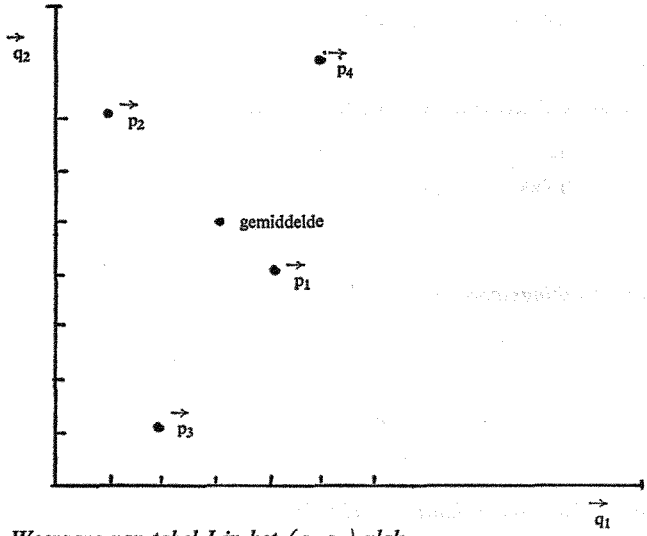
De getallen waaruit x_i bestaat heten de elementen of *componenten* van x_i, en x (zonder subscript) noemen we een *vector-variabele*. Anders gezegd: x is de algemene notatie voor een willekeurige observatie die bestaat uit een of meer componenten. Hoewel men gewoonlijk de kenmerken Q aanduidt als 'variabelen', is het juister om de observaties x aan te duiden als de variabelen: het zijn immers deze observaties, die variabele waarden op elk der kenmerken kunnen aannemen. De afzonderlijke aspecten heten dan de componenten van de variabele.

Een nuttig hulpmiddel in de beschouwing van variabelen met meerdere componenten is een voorstelling als punten in een meerdimensionele ruimte. We symboliseren de Q als onderling loodrechte assen $\vec{q}_1$, $\vec{q}_2$, etc. en de objecten P als punten $\vec{p}_1$, $\vec{p}_2$ etc., wier coördinaten op Q zijn gegeven in de getallenreeksen x_i.

Men merke op dat we een $\rightarrow$ toevoegen ter aanduiding van het feit dat we het over een meetkundige weergave hebben. De getallenreeks x kunnen we opvatten als de meting van p op Q, resp. als de coördinaten van $\vec{p}$ op $\vec{Q}$. Een voorbeeld is tabel I en figuur 1.

Tabel I. Eenvoudige data-matrix ($n = 4$, $m = 3$)

		q_1	q_2	q_3
p_1:	$x'_1 =$	4	4	9
P_2:	$x'_2 =$	1	7	4
p_3:	$x'_3 =$	2	1	6
p_4:	$x'_4 =$	5	8	5
gem:	$\bar{x}' =$	3	5	6



Figuur 1. Weergave van tabel I in het (q_1,q_2) vlak

In termen van de meetkundige voorstelling duiden we een multivariate variabele x aan als een meer-dimensionele variabele. De dimensionaliteit (= aantal dimensies) is m, het aantal componenten. Het gemiddelde van een meerdimensionele variabele krijgen we door het gemiddelde der afzonderlijke componenten te berekenen. Het levert ons weer een getallenreeks op die we aanduiden als de vector $\bar{x}$. Het punt met coördinaten $\bar{x}$ heet de *centroïd* van P. Zouden we onderscheid hebben gemaakt tussen verschillende observatiecondities C, dan zouden we voor elke conditie een centroïd $\bar{x}_h$ hebben kunnen berekenen (daarop komen we bij de bespreking van de multivariantie-analyse terug).

Het multidimensionele analogon van de variantie (= kwadraat van de standaard-deviatie) van een meerdimensionele variabele kan op verschillende wijzen gekozen worden. We bespreken allereerst de *variantie-covariantie-matrix*, die we aanduiden met C. De berekening van C is:

(1) $$C = \frac{1}{n}\sum_i (x_i - \bar{x})(x_i - \bar{x})$$

hetgeen wil zeggen dat per i alle produkten $(x_{ij} - \bar{x}_j)(\bar{x}_{ik} - \bar{x}_k)$ vormen, deze optellen over alle i en delen door n. Deze berekening is het voorbeeld van tabel I uitgevoerd in tabel II. Een element van de matrix C heeft de vorm:

(2) $$c_{jk} = \frac{1}{n}\sum_i (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

Als $j = k$ herkent men hierin de variantie van de j^{de} component. We definiëren als notatie van de standaarddeviatie (s_j):

(3) $$s_j = \sqrt{c_{jj}}$$

De correlatiematrix R krijgen we door te nemen:

(4) $$r_{jk} = \frac{c_{jk}}{\sqrt{c_{jj}\, c_{kk}}} = \frac{c_{jk}}{s_j \cdot s_k}$$

Tabel II

a) matrix van afwijkingsscores $(x_i - \bar{x})'$

	q_1	q_2	q_3
p_1	1	—1	3
p_2	—2	2	—2
p_3	—1	—4	0
p_4	2	3	—1

b) berekening variantie-covariantie-matrix C:

$$C = \frac{1}{n}\left\{\begin{pmatrix} 1 & -1 & 3 \\ -1 & 1 & -3 \\ 3 & -3 & 9 \end{pmatrix} + \begin{pmatrix} 4 & -4 & 4 \\ -4 & 4 & -4 \\ 4 & -4 & 4 \end{pmatrix} + \begin{pmatrix} 1 & 4 & 0 \\ 4 & 16 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 4 & 6 & -2 \\ 6 & 9 & -3 \\ -2 & -3 & 1 \end{pmatrix}\right\}$$

$$= \begin{pmatrix} 10/4 & 5/4 & 5/4 \\ 5/4 & 30/4 & -10/4 \\ 5/4 & -10/4 & 14/4 \end{pmatrix}$$

$$C = \begin{matrix} 2.5 & 1.25 & 1.25 \\ 1.25 & 7.50 & -2.50 \\ 1.25 & -2.50 & 3.50 \end{matrix}$$

c) correlatie-matrix R (de symmetrische onderhelft is weggelaten):

1.0	0.289	.423
	1.0	—.488
		1.0

De correlatie-matrix is de covariantie-matrix die we krijgen als we de afzonderlijke componenten herschalen zodat de variantie van elke component gelijk is aan 1. Deze herschaling is zinvol als de meet-eenheid van elke component willekeurig is, zoals meestal bij psychologische grootheden, of indien men welbewust de grootte-orde van elke component buiten beschouwing wil laten.

Omdat de covarianties worden uitgedrukt ten opzichte van het gemiddelde der componenten kunnen we in de meetkundige voorstelling zonder bezwaar de oorsprong van het assenstelsel verplaatsen naar de centroïd; daarmee brengen we dan de afwijkingsscores (tabel IIa) in beeld.

II. Hoofdassen-rotatie

De variantie-covariantie-matrix of de correlatie-matrix vertelt ons op zichzelf nog niet veel. De puntenwolk waarvan fig. 1 een tweedimensioneel voorbeeld liet zien, toont de dispersie van de observaties x in een m-dimensionele ruimte. Als alle componenten ongecorreleerd zijn heeft de puntenwolk in alle richtingen een dichtheid die uitsluitend bepaald wordt door de distributies der componenten. Als er echter samenhangen tussen de componenten bestaan, zal de dichtheid in bepaalde sectoren van de ruimte vrij gering zijn, omdat b.v. hoge intelligentie (q_1) en weinig schoolopleiding (q_2) relatief zeldzaam is. Bij een groot aantal componenten verwachten we in het algemeen zoveel onderlinge samenhangen dat niet m dimensies, maar veel minder dan m dimensies voldoende zijn om elke observatie bij benadering exact te representeren. De manier om dat vast te stellen is, te zoeken naar de richtingen (in de m-dimensionele ruimte) waarin de dispersie gering is, en de richtingen waarin de dispersie groot is. We voeren daartoe nieuwe componenten in: $A = \{\vec{a_t}, t = 1, \ldots m\}$, die we als een nieuw assenstelsel in onze puntenwolk leggen. De projecties van $\vec{p_i}$ op dit nieuwe assenstelsel kunnen we berekenen als we eerst de coördinaten van A ten opzichte van Q kennen. We definiëren:

$$
\begin{aligned}
& v_{jt} = \text{coördinaat van } \vec{a_t} \text{ op } \vec{q_j} \\
(5) \quad & x_{ij} = \text{projectie van } \vec{p_i} \text{ op } \vec{q_j} \\
& y_{it} = \text{projectie van } \vec{p_i} \text{ op } \vec{a_t}
\end{aligned}
$$

Als we nu de $\vec{Q}$-vectoren als loodrecht beschouwen en de eenheid op elke $\vec{a_t}$ bepalen door $\sum_j v_{jt}^2 = 1$ te stellen, krijgen we

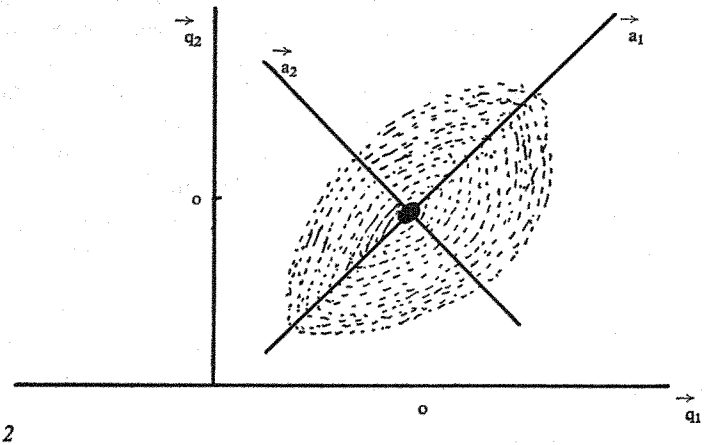
$$(6) \qquad y_{it} = \sum_j x_{ij} v_{jt}$$

Met andere woorden, de 'score' op a_t is een gewogen som van de scores op Q.

We kunnen de richting van $\vec{a_1}$ zodanig kiezen dat de variantie van $y_{.1}$ (dus de variantie over de individuen) zo groot mogelijk is. Men noemt deze $\vec{a_1}$ de eerste hoofdas van $\vec{P}$.

Zo kan men vervolgens een tweede $\vec{a_2}$ bepalen die loodrecht op de eerste staat. Men kan aan fig. 2 gemakkelijk zien dat dan de scores (= projecties) van de individuen op deze twee assen ongecorreleerd zijn. De tweede as wordt loodrecht op de eerste zo gekozen dat de variantie van $y_{.2}$ weer zo groot mogelijk is. Men gaat zo door totdat de variantie op een volgende as te verwaarlozen is.

Men vindt de coëfficiënten v_t door de *eigenvectoren* te berekenen van de covariantie-matrix C. De daarbij behorende *eigenwaarden* λ_t zijn gelijk aan de variantie van $y_{.t}$. Op de berekeningsmethode gaan we hier niet in. Bij Pawlik (1969a, p. 67) en Harman (1968) vindt men gedetailleerde beschrijvingen van deze, en andere, berekeningswijzen.



Figuur 2

Men noemt het assenstelsel $\vec{A}$ de *hoofdassen* van $X = (x_i', i = 1, \ldots, n)$. Afgebeeld op deze nieuwe assen is y_i de representatie van p_i. De componenten van y_i zijn op te vatten als de scores van p_i op nieuwe componenten, die zijn geconstrueerd als lineaire combinatie van de oorspronkelijke. Doordat de nieuwe componenten zodanig zijn gekozen dat zij ongecorreleerd zijn, kunnen we aan de varianties λ_t zien hoe groot de dispersie van de puntenwolk is in elk der richtingen (dimensies) $\vec{a_t}$, $t = 1, \ldots, m$. Meestal blijkt dan dat slechts in enkele dimensies sprake is van een beduidende spreiding; op de overige is de spreiding klein en wij mogen dan deze dimensies verwaarlozen (er bestaan daarvoor statistische criteria). Dit betekent dat wij de m-dimensionele puntenwolk projecteren in een r-dimensionele subruimte ($r < m$), en wel de r-dimensionele subruimte gevormd door de hoofdassen met grootste λ's.

Een enkelvoudige maat voor de spreiding van een m-dimensionele puntenwolk is het geometrisch gemiddelde van de eigenwaarden van de covariantiematrix (de m^{de} machtswortel van het voortgezet produkt van alle eigenwaarden).

III. Multidimensionele schaling

Men kan $\vec{Q}$ als een multidimensionele schaal beschouwen, waarop de P zijn uitgedrukt. Ook $\vec{A}$ is een multidimensionele schaal, die equivalent is met $\vec{Q}$, doch van $\vec{Q}$ verschilt door de transformatie V. De posities van de P op deze schalen zijn resp. x en y. Het is mogelijk om de P te schalen als men alleen beschikt over de gelijkenissen tussen de afzonderlijke p_i, p_j. Men moet dan uiteraard beschikken over een gelijkenissenmaat en deze interpreteren als een functie van de onbekende schaalwaarden y (of x). Een gangbare methode is om numerieke gegevens te verzamelen omtrent de *on*gelijkenis tussen alle paren (p_i, p_j), bv. met een 'rating-methode'. Men stelt dan de ongelijkenis gelijk aan de euclidische afstand d_{ij} tussen de punten die p_i en p_j representeren (zie figuur 1).

Zij nu $\vec{A}$ een coördinaten-stelsel voor de onbekende y. Men kan afleiden (zie Torgerson, 1958, p. 254) dat:

$$(7) \qquad \sum_t y_{it}y_{jt} = \tfrac{1}{2}\left(\frac{1}{n}\sum_k d_{ik}^2 + \frac{1}{n}\sum_k d_{kj}^2 - \frac{1}{n^2}\sum_{kl} d_{kl}^2 - d_{ij}^2\right)$$

waarin i, j, k, l = 1, ..., n de indices zijn voor de individuen. De elementen in het rechterlid vormen een matrix $B = YY'$. Van deze matrix bepaalt men de eigenvectoren w_t ($t = 1, \ldots$) en de eigenwaarden λ_t.

Men kan nu stellen:

$$(8) \qquad y_{it} = w_{it}\,\lambda_t^{\frac{1}{2}}$$

waarmee we een representatie van de P krijgen op een schaal $\vec{A}$ met ongecorreleerde componenten, uitgaande van gegeven ongelijkenissen d_{ij}. De variantie van P op elke dimensie $\vec{a_t}$ is weer λ_t.

Ook hier laten we dimensies met kleine λ_t weg. We projecteren daarmee de puntenwolk in een ruimte met zo klein mogelijke dimensionaliteit waarmee we toch nog de gegeven d_{ij} bij benadering exact kunnen presenteren. Hier doet zich ook het probleem voor dat we een interpretatie aan de dimensies willen geven. Deze is uiteraard slechts mogelijk indien we reeds iets weten over de 'objecten' p_i.

IV. Component-analyse

Met de naam *(principal) components analysis* wordt een analysetechniek aangeduid die nauw verwant is met de hoofdassen-rotatie van de vorige paragraaf. De componenten-analyse wordt ook vaak aangeduid als een vorm van factor-analyse, hoewel men tegen deze aanduiding ernstige theoretische bezwaren kan hebben (zie paragraaf V).

Componenten-analyse zal ons helpen om de dimensies die door hoofdassen-rotatie zijn verkregen te interpreteren. De gewichten v_t waarmee we de t^{de} hoofdas construeren vertellen namelijk nog niet veel over de samenhang tussen de hoofdas en de empirische componenten Q. Die samenhang kan blijken uit de correlaties tussen de scores van P op $\vec{a_t}$ en de scores van P op $\vec{Q}$. De matrix van deze correlaties (f_{jt}) wordt gevormd door

$$(9) \qquad f_{jt} = v_{jt}\,\lambda_t^{\frac{1}{2}}$$

als v_t een eigenvector van de correlatie-matrix van X is en λ_t de daarbij behorende eigenwaarde[1].

Omdat f_{jt} een relatie uitdrukt met gestandaardiseerde y, definiëren we

$$(10) \qquad z_{it} = y_{it}\,\lambda_t^{-\frac{1}{2}}$$

m.a.w. z is de gestandaardiseerde y (we delen door de standaarddeviaties). Deze standaardisatie is equivalent met het invoeren van nieuwe eenheden op de A-dimensies. Meetkundig is dat uitgedrukt in figuur 3. We rekken de puntenwolk van P als het ware uit zodat hij op elke dimensie $\vec{a_t}$ eenzelfde spreiding (gelijk aan 1) heeft.

In de figuur willen we de x_{ij} nog steeds terug vinden als de projectie van $\vec{p_i}$ op $\vec{q_j}$, doch dat kan alleen als we de $\vec{Q}$-vectoren anders tekenen, en wel zodanig dat de projecties van $\vec{Q}$ op $\vec{A}$ gelijk zijn aan de correlaties f_{it} tussen q_j en a_t. Zo ontstaat de bekende 'factor-structuur-representatie' (die eigenlijk ' componenten-structuur-representatie' zou moeten heten).

Men kan nu de volgende relaties afleiden:

$$(11) \qquad \frac{x_{ij} - \bar{x}_j}{s_j} = \sum_t (z_{it} - \bar{z}_t)\, f_{jt}$$

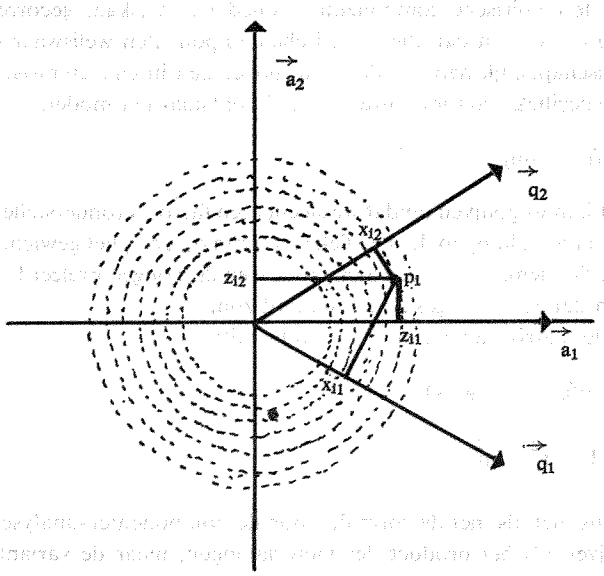
(waarin s_j de standaarddeviatie van $x_{.j}$ is). Meestal drukt men de voorgaande formule wat eenvoudig uit door te veronderstellen dat de x reeds gestandaardiseerd is, met gemiddelden gelijk aan nul, en standaarddeviaties gelijk aan 1. Men krijgt dan:

(12) $x_{ij} = \sum_t z_{it} f_{jt}$ of $X = ZF'$

Men ziet aan deze uitdrukking dat we x_{ij} opvatten als een gewogen combinatie van de standaardscores van een individu op elk der hoofdassen, waarbij de gewichten f_{jt} aangeven hoe sterk q_j 'beladen' is met a_t. Voorts is:

(13) $r_{jk} = \sum_t f_{jt} f_{kt}$ ofwel $R = FF'$

Met andere woorden: de correlatie tussen twee empirische componenten is uitgedrukt als de productsom van de gewichten of ladingen op de principiële componenten A.



Figuur 3

Een belangrijk praktisch probleem is de rotatie van het assenstelsel. Heeft men eenmaal een oplossing met een bepaald aantal dimensies, dan kan men het assenstelsel roteren. Men krijgt dan een afbeelding van de Q-componenten op een ander factorstelsel, dat evengoed bij de gegevens past. Bij voorkeur roteert men zodanig dat elke component zoveel mogelijk lage factor-ladingen heeft, en zo weinig mogelijk hoge ladingen. Men krijgt dan een eenvoudige structuur die beter interpreteerbaar is. Men kan een factor interpreteren aan de hand van de componenten die een hoge lading bij die factor hebben, en lage ladingen op alle andere factoren (zie verder Harman, 1968, Pawlik, 1968).

V. Factor-analyse

Factor-analyse lijkt bedriegelijk veel op componenten-analyse, maar wijkt daarvan af doordat de factor-analyse uitgaat van een zekere theorie, zij het een zeer globale, over de data.

Kijken we nog eens naar (12); we schreven daar dat de geobserveerde score van individu p_i op component q_j gelijk is aan de gewogen som van zijn score op de hoofdcomponenten a_t. We kunnen nu deze a_t opvatten als theoretische grootheden of factoren, waarmee we de samenhang tussen de geobserveerde scores willen gaan beschrijven. In het allereenvoudigste geval zouden we kunnen veronderstellen dat slechts één factor de scores op de empirische componenten bepaalt; zou dit zo zijn dan zouden alle empirische componenten volledig met elkaar gecorreleerd zijn. We nemen nu echter aan dat alle empirische componenten weliswaar slechts één factor gemeenschappelijk hebben, doch dat bovendien in elke empirische component nog een specifieke factor aanwezig is[2]. Zo ontstaat het model:

$$(14) \quad x_{ij} = z_i f_j + e_{ij} u_j$$

De t-subscript kan wegblijven omdat we slechts één factor veronderstellen; voorts is e_{ij} de score van individu p_j op de specifieke factor in q_j, en u_j het gewicht of aandeel daarvan in q_j. We nemen uiteraard aan dat de *e* en de *z* ongecorreleerd zijn. Voorts nemen we aan dat *x*, *e* en *z* gestandaardiseerd zijn.

De correlatie-matrix van *x* ziet er dan aldus uit:

$$(15) \quad r_{jk} = f_j f_k \qquad (j \neq k)$$

$$(16) \quad r_{jj} = 1 = f_j^2 + u_j^2$$

We kunnen dus, net als met de formule voor de componenten-analyse, de correlaties beschrijven als het product der factorladingen, maar de variantie (gestandaardiseerd zodat $r_{jj} = 1$) is opgebouwd uit een 'gemeenschappelijk' deel en een 'specifiek' deel.

De voorgaande theorie is afkomstig van Spearman (1904, 1914) en veralgemeniseerd door Thurstone (1931, 1946). Thurstone's 'common factor-analysis' gaat uit van:

$$(17) \quad x_{ij} = \sum_t z_{it} f_{jt} + e_{ij} u_j \qquad t = 1, \ldots, r < m$$

en, veronderstellen we dat de componenten van *z* ongecorreleerd zijn:

$$(18) \quad r_{jk} = \sum_t f_{jt} f_{kt} \qquad (j \neq k)$$

De r_{jj} zijn nu:

(19) $$r_{jj} = \sum_t f_{jt}^2 + u_j^2$$

De totale variantie ($r_{jj} = 1$) van een empirische component is dus opgebouwd uit twee delen:

(20) de communaliteit $h_j^2 = \sum_t f_{jt}^2$

(21) de uniciteit u_j^2.

Stellen we de Q-vectoren voor in een r-dimensionele ruimte, dan krijgen we een voorstelling waarbij de lengte van elke vector $\vec{q_j}$ gelijk is aan h_j (de wortel uit de communaliteit, d.i. de standaarddeviatie van het 'gemeenschappelijke deel' van q_j). Voorts is:

(22) cosinus $(\vec{q_j}, \vec{q_k}) = \dfrac{r_{jk}}{h_j h_k}$

De interpretatie van een multidimensionele variabele in termen van een (klein) aantal gemeenschappelijke factoren is zeer aantrekkelijk. We reduceren de scores op de empirische componenten tot de scores op een kleiner aantal theoretische componenten (A), en we staan toe dat elke empirische component iets specifieks bevat dat hij niet gemeen heeft met de andere.

De factor-ladingen f_{jt} geven, evenals bij de componenten-analyse, een soort profiel van elke q_j. De score der individuen op elke empirische component wordt herleid tot hun score (z_{it}) op de gemeenschappelijke factoren en het aandeel f_{jt} van elke gemeenschappelijke factor in de variantie op de empirische componenten, plus zijn score (e_{ij}) op de specifieke (of onsystematische) factor en het aandeel (u_j) daarvan.

Men heeft de factor-analyse voornamelijk in de psychologie gebruikt om een beter inzicht te krijgen in de factoren die een rol spelen in b.v. intellectuele prestaties. Men verzamelt scores van een steekproef van individuen op een multivariate intelligentie-test, en voert daarop een factor-analyse uit. Van belang is dat men uitgaat van een 'universe of content' (Guttman, 1953, p. 280), dat wil zeggen: de multivariate variabele x moet representatief zijn voor het domein waarin men is geïnteresseerd. Slechts dan kan een factor-analyse antwoord geven op de vraag welke factoren in dat domein van belang zijn. Bovendien moet uiteraard de steekproef van individuen representatief zijn, en niet een selectie waarin het 'universe of content' onvoldoende variantie vertoont.

Alternatieven voor factor-analyse

Naast de principale componenten-analyse en de Spearman-Thurstone factor-analyse zijn modellen ontwikkeld die berusten op andere interpretaties van de correlatie of covariantie-matrix.

Theoretisch belangwekkend is Guttman's (1953a) 'image analyses', waarvan Thurstone's 'common factor analysis' een speciaal geval is.

Practisch vooral van betekenis is Guttman's (1953 b) 'radex model', waarvan een illustratie te vinden is in Guttman (1964).

VI. Q-analyse en cluster-analyse

Vaak is men minder geïnteresseerd in de structuur van de componenten der variabelen, en meer in de structuur van de observatie-objecten (p_i, $i = 1, \ldots$) welks kenmerken in de multivariate scoring x_i zijn vastgelegd. Voor de hand liggend is de z.g. Q-analyse[3], dat is een factor-analyse op de gekantelde data-matrix. Men verwisselt eenvoudig objecten en componenten. Aan deze techniek kleven echter een aantal bezwaren, zodat toepassing daarvan zonder meer vaak tot onbevredigende resultaten leidt. Een bezwaar is dat grote verschillen in de gemiddelden der componenten leidt tot een niet-relevante verhoging van de correlaties tussen de objecten. De eerste factor die men dan vindt is vrijwel geheel een artefact. Men lost dit op door eerst de gemiddelden der componenten af te trekken, en dan de correlaties tussen de objecten te berekenen. Aangezien bij de berekening van deze correlaties nu ook de gemiddelde score van een object van zijn scores op alle componenten wordt afgetrokken is de resulterende covariantie-matrix dubbel-gecentreerd, met als gevolg dat elk object een negatieve lading krijgt op de unieke factoren der andere objecten (Delbeke, 1969, p. 64) als men tenminste meer componenten dan objecten heeft (anders kan men alleen maar een principale componenten-analyse uitvoeren).

Bij een dubbel-gecentreerde matrix zijn alle factoren (ook de unieke factoren) *bipolair* dat wil in dit geval zeggen dat de som der ladingen per factor steeds nul is.

Een ander probleem wordt gevormd door de scoringsrichting. Spiegelt men een component om zijn gemiddelde (d.w.z. men vervangt x_{ij} door $x_{ij} - 2(x_{ij} - \bar{x}_j)$) dan draait ook het teken van de correlaties met andere componenten om, en hetzelfde geldt voor de factorlading van een gespiegelde component. De scoringsrichting is vaak arbitrair (zeker bij attitudes en waarderingsoordelen) maar behoeft bij correlatieberekening en hoofdassen-analyse geen zorgen te baren.

Past men echter een factor-analyse toe op de gekantelde data-matrix, dan kan het omkeren van een scoringsrichting van een aantal componenten grote invloed hebben op de correlaties tussen de objecten (personen). In plaats van correlaties tussen objecten zou dan ook een afstandsmaat, die door zulke omdraaiingen niet beïnvloed wordt, gebruikt moeten worden. De simpelste is:

$$(23) \quad d_{ik} = \sqrt{\sum_j (x_{ij} - x_{kj})^2}, \qquad j = 1, \ldots, m$$

die echter geen rekening houdt met de meet-eenheid der componenten van x_i. Daarmee houdt wel rekening:

$$(24) \quad d_{ik} = \sqrt{\sum_j (x_{ij} - x_{kj})^2 / s_j} \qquad j = 1, \ldots, m$$

Een stap verder gaat de z.g. Mahalanobis-maat, die bovendien de covarianties tussen de componenten verdisconteert:

$$(25) \quad d_{ik} = \sum_{jl} (x_{ij} - x_{kj})\, c^{jl} (z_{il} - x_{kl}) \qquad j,l = 1, \ldots, m$$

waarin c^{jl} het (j, l) element in de inverse[4] van de covariantie-matrix C is.

Ook andere afstandsmaten zijn denkbaar, en een keuze is moeilijk. De volgende stap dient te bestaan uit toepassing van een procedure die groepjes (clusters) van objecten vormt op grond van hun onderlinge afstand. De meest rationele methoden zoeken naar een verdeling in groepen zodanig dat binnen elke groep de kwadraatsom der afstanden zo klein mogelijk is en tussen alle groepen de kwadraatsom der afstanden zo groot mogelijk is. Men gaat dan hiërarchisch te werk, b.v. door eerst twee objecten die de kleinste onderlinge afstand hebben samen te nemen, en zo vervolgens elk object samen te nemen met een ander object (of groepje van objecten).

Op dit gebied is de laatste jaren een hernieuwde belangstelling ontstaan, die mogelijk is geworden doordat slechts zeer grote en snelle computers in staat zijn deze procedures, die soms bestaan uit niet anders dan systematische trial-and-error, uit te voeren (zie verder Cattel, Coulter & Tsujoka, 1966, Fleiss & Zubin, 1969, Wishart, D. 1969, Boon van Ostade, 1971).

Heeft men dan een aantal clusters gevonden dan kan men een karakterisering van elk cluster krijgen door het gemiddelde x te bepalen van de objecten in dat cluster.

VII. Multiple regressie en partiële regressie

Soms wil men een enkelvoudige afhankelijke variabele relateren aan een multivariate onafhankelijke variabele. De eenvoudigste vorm waarin dat kan geschieden is door de afhankelijke variabele y te benaderen met een lineaire combinatie van de onafhankelijke variabele x:

$$(26) \quad y_i = \sum_j x_{ij} v_j + e_i$$

Hierin is v_j het regressie-gewicht van y op $x_{.j}$, en e_i is een 'residu'. Men tracht nu de

gewichten zodanig te bepalen dat de residuele variantie zo klein mogelijk is. Dit leidt tot de volgende oplossing:[4] is

$$(27) \qquad \beta = R_{xx}^{-1} r_{xy}$$

waarin β een kolom van standaardgewichten is ($\beta_j = (s_y/s_j)\, v_j$), toepasbaar op gestandaardiseerde scores. De matrix R is de correlatiematrix van x, en r_{xy} is de kolom van correlaties tussen y en de componenten van x.

De multiple correlatie (MR) tussen x en y is de correlatie tussen de gewogen som-scores $\sum_j x_{ij} v_j$ en y:

$$(28) \qquad MR^2 = \sum_{jk} \beta_j \beta_k r_{jk}$$

Van belang om op te merken is dat de multiple correlatie berekend op een steekproef van observaties meestal een overschatting geeft van de ware multiple correlatie, en wel des te meer naarmate men minder observaties en meer predictoren (= het aantal componenten van x) gebruikt (cf. Guilford, 1956, pp. 398–399; Herzberg, 1969).

Een meetkundige interpretatie van de multiple correlatie is de volgende. Men stelle zich de vector $\vec{q}_y$, die de afhankelijke variabele representeert, en de vectoren $\vec{q}_j$ ($j = 1, \ldots, m$) die de onafhankelijke variabele representeren, voor in een $m + 1$ dimensionele ruimte. Noem $\vec{q}_y^*$ de projectie van $\vec{q}_y$ op de ruimte die door $\vec{q}_j$ ($j = 1, \ldots, m$) wordt opgespannen. De vector $\vec{q}_y^*$ representeert dan de lineaire combinatie van de onafhankelijke variabelen die de afhankelijke variabele q_y zo goed mogelijk benadert. De cosinus van de hoek tussen $\vec{q}_y$ en $\vec{q}_y^*$ is dan gelijk aan de multiple correlatie.

Multiple regressie is vooral van practische betekenis voor het voorspellen van een criterium. Veel inzicht in de interne samenhang der betreffende componenten krijgt men er meestal niet door. Dat is vooral het gevolg van het feit dat de multiple regressie, in de bovengeschatte vorm, een *lineaire* methode is, zodat niet lineaire samenhangen verborgen blijven[5]. Ook het feit dat de mate (of vorm) van de samenhang tussen de j^de^ component van x en het criterium kan variëren met de score op een andere component van x, kan het beeld volledig vertroebelen[6]. Vaak kan men alleen door een zorgvuldige inspectie van de gegevens op het spoor komen van zulke verschijnselen.

In de tweede plaats is de multiple regressie-analyse van betekenis voor het afzonderen van de gemeenschappelijke en de 'eigen' delen van de (co-)varianties der componenten van x.

Van het afzonderen van de 'eigen' delen wordt gebruik gemaakt in de zg. covariantie-analyse, waarop we nu nader ingaan.

De term covariantie-analyse is verwarrend; bedoeld is dat men een samenhang

van de multivariate variabele y in relatie tot een bepaald stel van (experimentele) condities onderzoekt, doch daarbij de covariantie met x (een multivariate covariabele) wil controleren of elimineren, zodat we de 'eigen' covarianties van de componenten van y overhouden.

Wil men bijvoorbeeld de samenhang tussen de prestaties van scholieren in verschillende vakken onderzoeken, dan zal men willen afzien van de factor intelligentie die – vermoedelijk – een groot deel van de covariantie tussen schoolcijfers in verschillende vakken veroorzaakt.

In het algemeen zal men een grootheid die 'spurious' covariantie veroorzaakt willen elimineren. In een onderzoek[7] werden de frequenties van verschillende syntactische componenten in het proza van een auteur geteld. Het samengaan (covariëren) van bepaalde componenten kan als 'stijl-factor' worden geïnterpreteerd; het bleek nu echter dat verscheidene componenten sterke samenhang vertoonden met het aantal woorden per zin en met het aantal redekundige woordengroepen per zin. Deze bron van 'spurious' correlaties moest worden geëlimineerd.

De eliminatie-techniek berust op het berekenen van de residuele scores die overblijven na het berekenen van de multiple regressie van x op y (x is de covariabele die geëlimineerd moet worden). Volgens (26) is het residu $\hat{y}$:

$$\hat{y}_i = y_i - x_i' V \tag{29}$$

waarin V bestaat uit de kolommen die elk de regressie-gewichten van de onderscheiden componenten van y op x bevatten. De covariantie van $\hat{y}$ ziet er dan als volgt uit

$$C_{yy.xx} = C_{yy} - C_{yx} C_{xx}^{-1} C_{xy} \tag{30}$$

Om hieruit de residuele correlatie-matrix te krijgen moet dan nog elk element van $C_{yy.xx}$ gedeeld worden door de wortel uit het product van de residuele varianties; deze vindt men uiteraard op de diagonaal van $C_{yy.xx}$.

De residuele covariantie- of correlatie-matrix is meestal op zich zelf weinig zeggend: waar het om gaat is dat we daarop bv. een factor-analyse of een multivariate variantie-analyse kunnen doen, waarvan de uitkomsten nu gezuiverd zijn van de geëlimineerde covariabele x.

In plaats van 'residuele correlatie' spreekt men ook van 'partiële correlatie'.

De meetkundige voorstelling van partiële correlaties is de volgende. De componenten van y worden voorgesteld als vectoren $\vec{Q}_y$, en de componenten van x worden voorgesteld als vectoren $\vec{Q}_x$. Beide vectorbundels spannen een ruimte op van resp. m_y en m_x dimensies. Men stelle zich voorts voor een ruimte V waarvan alle dimensies loodrecht staan op de ruimte van $\vec{Q}_x$. Door nu $\vec{Q}_y$ in V te projecteren krijgen we vectoren $\vec{Q}_y^*$ die loodrecht staan op $\vec{Q}_x$. De projecties stellen componenten voor die met Q_x niet gecorreleerd zijn. De cosinussen van de hoeken tussen deze projecties van $\vec{Q}_y$ in V, loodrecht op $\vec{Q}_x$, zijn partiële correlaties $R_{yy.xx}$.

VIII. Canonische correlaties

Een canonische correlatie is een veralgemenisering van de multiple correlatie, en in zekere zin de tegenhanger van de matrix van partiële correlaties.

Voor de multiple correlatie zochten we gewichten V die we aan x_i konden toekennen om y_i te benaderen, waarbij x uit meerdere, en y uit één component bestaat. Als nu ook y uit meerdere componenten bestaat, proberen we gewichten V en W te vinden zodanig dat we twee grootheden $a^{(x)}$ en $a^{(y)}$ vinden als combinaties van de componenten van x, resp. y, die zo veel mogelijk overeenkomen. Zij

$$(31a) \quad z_i^{(x)} = \sum_j \left(\frac{x_{ij} - \bar{x}_j}{s_{jj}}\right) v_j \qquad j = 1, \ldots, m^{(x)} \text{ (componenten)}$$

$$i = 1, \ldots, N \text{ (observaties)}$$

$$(31b) \quad z_i^{(y)} = \sum_k \left(\frac{y_{ik} - \bar{y}_k}{s_k}\right) w_k \qquad k = 1, \ldots, m^{(y)} \text{ (componenten)}$$

en

$$\lambda = \text{correlatie tussen } z^{(x)}, z^{(y)}.$$

De getallen $z_i^{(x)}$ en $z_i^{(y)}$ zijn de scores van p_i op de canonische componenten $a^{(x)}$ en $a^{(y)}$.
Deze correlatie λ heet de canonische correlatie tussen x en y. De rekenkundige oplossing is tamelijk ingewikkeld. Men moet de grootste eigenwaarde en eigenvector w berekenen van:

$$(32) \quad (R_{yx} R_{xx}^{-1} R_{xy} - R_{yy}\, \lambda^2)\, w = o$$

en vindt vervolgens v uit:

$$(33) \quad \lambda^{-1} R_{xx} R_{xy}\, w = v$$

(hierin zijn x en v verwisselbaar met y en w).

De vergelijkingen (32) en (33) hebben meerdere oplossingen. Uiteraard is die oplossing welke de grootste λ heeft de gezochte maximale canonische correlatie. Ook de andere oplossingen zijn echter interessant. Elke λ_t met gewichten v_t en w_t is een canonische correlatie. De canonische dimensies $\vec{a}_t^{(x)}$ en $\vec{a}_t^{(y)}$ $(t = 1, \ldots, r)$ zijn verschillende lineaire combinaties van de componenten van x, resp. y. Deze zijn alle ongecorreleerd, behalve dan de paren $(\vec{a}_t^{(x)}, \vec{a}_t^{(y)})$ waartussen de correlatie gelijk is aan λ_t. Aldus heeft men bij elke multivariate variabele, x, resp. y, een stel ongecorreleerde dimensies vergelijkbaar met factoren, zodanig dat elk daarvan zo hoog mogelijk correleert met zijn tegenhanger uit de andere multivariate variabele.

De meetkundige voorstelling is als volgt. De componenten van x, resp. y worden

voorgesteld door vectoren $\vec{Q}_x$ en $\vec{Q}_y$. In de ruimte van $\vec{Q}_x$ wordt nu een dimensie $\vec{a}^{(x)}$ gezocht, en in de ruimte van $\vec{Q}_y$ wordt een dimensie $\vec{a}^{(y)}$ gezocht, zodanig dat de cosinus van de hoek tussen deze dimensies zo groot mogelijk is.

Meer algemeen: er wordt in $\vec{Q}_x$ een loodrecht assenstelsel $\vec{A}_x$, en in $\vec{Q}_y$ een loodrecht assenstelsel $\vec{A}_y$ gezocht, zodanig dat elke as van het éne stelsel bovendien loodrecht staat op alle assen op één na in het andere stelsel.

Om deze canonische dimensies te interpreteren dient men niet naar de gewichten v_t en w_t te kijken, maar naar de correlaties van elke component van x, resp. y met de canonische dimensies. Deze correlaties zijn vergelijkbaar met factor-ladingen; men vindt ze door

$$(34) \qquad f_{jt(x)} = \sum_i r_{ij}\, v_{it} \qquad i, j = 1, \ldots, m^{(x)}$$

$$\text{en} \qquad f_{kt(y)} = \sum_i r_{ik}\, w_{it} \qquad i, k = 1, \ldots, m^{(y)}$$

Als voorbeeld van de toepassing van canonische correlaties noemen we twee onderzoeken. In het ene ging het om de vraag of twee Nederlandse capaciteiten-tests als equivalent mochten worden beschouwd. De ene test bestond uit 8 subtests, de andere uit 10; de steekproef bestond uit 100 personen die zich hadden gemeld bij een beroepskeuzeadvies-bureau (de steekproef is eigenlijk te klein, en zeker niet representatief). De grootste canonische correlaties waren .88 en .76 (de procedure voor significantie-toetsing bespreken we in een volgende paragraaf). De conclusie lijkt dus gewettigd dat de beide tests een tweetal factoren gemeenschappelijk hebben. De berekende gewichten w en v maakten het mogelijk om de onderzochten een score toe te kennen op elk van deze factoren, waarbij de hoge canonische correlaties er op wezen dat de scoring even goed vanuit de éne als vanuit de andere test kon geschieden. Zij mochten althans ten aanzien van deze factoren als equivalent worden beschouwd. De interpretatie van de factoren was overigens niet geheel duidelijk. Men zou vermoedelijk de factoren hebben moeten roteren om tot een duidelijk beeld te komen. Opgemerkt zij hierbij dat rotatie alleen zinvol is als de canonische correlaties zeer hoog zijn; door de rotatie worden de grootste canonische correlaties weliswaar sub-maximaal, maar de kleinste kan nooit kleiner worden.

Een tweede voorbeeld[8] ontlenen we aan een analyse van een creativiteit-testserie. Aangezien de test pretendeert onafhankelijk te zijn van intelligentie, werden de canonische correlaties met een intelligentie-testserie bepaald, ten einde de hypothese te kunnen toetsen dat deze niet significant van nul zouden afwijken.

Hier is duidelijk te zien dat canonische correlatie wordt gehanteerd als het multivariate equivalent van een correlatie tussen twee één-dimensionele variabelen. Door canonische correlaties te berekenen werden de twee test-series in hun geheel met elkaar vergeleken.

IX. Multivariate variantie-analyse en discriminant-analyse

Variantie-analyse is wellicht één der meest gebruikte technieken voor statistische analyse van experimentele uitkomsten. In het eenvoudigste geval hebben we een ééndimensionele variabele b.v. 'progressieve attitude' geobserveerd bij een verzameling individuen, die op een a priori criterium verdeeld kunnen worden in een aantal groepen, b.v. naar nationaliteit, of naar opleidingsniveau. In het geval van een indeling naar opleidingsniveau zou men b.v. 5 niveaus kunnen onderscheiden, elk niveau een nummer geven, en de correlatie tussen opleidingsniveau en de gemeten attitude uitrekenen. Het bezwaar daarvan is dat een lineair verband wordt berekend met een score voor opleidingsniveau die hoogstens ordinale betekenis heeft. Nu is de berekening van een correlatie eigenlijk niets anders dan een berekening van de mate waarin men de attitude-score kan voorspellen uit de opleidingsniveau-score. Zonder lineariteit te veronderstellen kan men dan beter uitrekenen hoe groot de variantie der attitude-scores binnen elk opleidingsniveau afzonderlijk is, het (naar aantal gewogen) gemiddelde van deze varianties berekenen, en deze delen door de totale variantie der attitude-scores. Men heeft dan een maat voor de proportie residuele variantie als men het groepsgemiddelde zou gebruiken als beste voorspeller voor de attitudescore van iemand die tot een bepaalde opleidingsgroep behoort.

Men kan evengoed de variantie der gewogen groepsgemiddelden delen door de totale variantie en daarmee de mate van voorspelbare variantie uitdrukken. Deze laatste maat heet de correlatie-ratio; hij wordt gewoonlijk geschreven als η^2, en is equivalent met het kwadraat van een correlatie-coëfficiënt. In formule[9]

$$s^2_{\text{totaal}} = s^2_{\text{binnen groepen}} + s^2_{\text{tussen groepen}}$$

$$s^2_{\text{tussen groepen}} / s^2_{\text{totaal}} = r^2$$

Als x een meerdimensionele variabele is gaan we eerst nieuwe componenten in voeren – vergelijkbaar met de componenten die wij met hoofdassenrotatie invoerden – die ongecorreleerd zijn. In onderscheid met de hoofdassenrotatie gaan we nu echter de nieuwe componenten zo kiezen dat daarop de variantie-tussengroepen zo groot mogelijk is. De belangrijkste van deze dimensies (dus met de relatief grootste tussen-groepen-variantie) is tevens de dimensie waarop wij de groepen zo goed mogelijk kunnen discrimineren. Men kan bewijzen dat er bij g groepen, g–1 ongecorreleerde dimensies zijn waarop de groepen onderscheidbaar zijn. De overige m–(g — 1)-dimensies kunnen we dan eventueel bepalen volgens dezelfde criteria die bij hoofdassenrotatie werden gehanteerd. We krijgen als resultaat een voorstelling van alle observaties in een m dimensionele ruimte, waarvan g–1 dimensies zo zijn gekozen dat zij de groepen maximaal discrimineren.

Het formularium voor deze bewerking is tamelijk ingewikkeld. We berekenen eerst de matrix van productsommen binnen-groepen (W, 'within') en vervolgens de matrix van productsommen over alle observaties (T, 'total'):

$$(35a) \qquad W = \{w_{jk} = \sum_{i\,(h)} \sum (\bar{x}_{ij} - \bar{x}_{hj})(\bar{x}_{ik} - \bar{x}_{hk})\} \qquad \begin{array}{l} j, k = 1, \ldots, m \\ h = 1, \ldots, g \end{array}$$

$$(35b) \qquad T = \{t_{jk} = \sum_{i} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)\} \qquad \begin{array}{l} j, k = 1, \ldots, m \\ i = 1, \ldots, n_t \end{array}$$

waarin $\bar{x}_{jh}$ is het gemiddelde van de j^{de} component in de h^{de} groep, en $\bar{x}_j$ is het algemeen gemiddelde van de j^{de} component, en n_t is het aantal observaties in de totale groep.

Eventueel kunnen we de matrix van productsommen tussen-groepen (B, 'between') gemakkelijk vinden door:

$$(36) \qquad B = \{b_{jk} = t_{jk} - w_{jk}\}$$

De correlatie-ratio η^2_r op een discriminerende dimensie a_r vinden we door de geobserveerde componenten te wegen met gewichten v_{jr}; de score op een discriminerende dimensie wordt dan:

$$(37) \qquad y_{ir} = \sum_{j} (x_{ij} - \bar{x}_j)\, v_{jr}$$

en de correlatie-ratio wordt:

$$(38) \qquad \eta^2_r = \frac{\sum_h n_h\, \bar{y}^2_{hr}}{\sum_i y^2_{ir}}$$

(merk op dat door (37) het gemiddelde $\bar{y}_{.r}$ gelijk is aan nul). In termen van de productsommen-matrices is η^2_r gelijk aan:

$$(39) \qquad \eta^2_r = 1 - \frac{\sum_{jk} \sum_{i\,(h)} \sum v_{jr}(x_{ij} - \bar{x}_{hj})(\bar{x}_{ik} - \bar{x}_{hk})\, v_{kr}}{\sum_{jk} \sum_{i} v_{jr}(x_{ij} - x_j)(x_{ij} - x_k)\, v_{kr}} = 1 - \frac{v_r' \, W v_r}{v_r' \, T \, v_r} = \frac{v_r' \, B \, v_r}{v_r' \, T v_r}$$

Men vindt de getallenreeks v_r waarbij η^2_r maximaal wordt door de volgende vergelijking op te lossen:

$$(40) \qquad (B - \eta^2_r T)\, v_r = o \text{ oftewel } \sum_k (b_{jk} - \eta^2_r t_{jk})\, v_{kr} = 0$$

Dit is een zg. algemene eigenwaarde-eigenvector vergelijking die met standaardprocedures door een computer snel kan worden opgelost.

Men noemt v_1 een lineaire discriminant-functie. Er zijn g—1 van zulke functies omdat de centroïden van g groepen een ruimte van g—1 dimensies vormen (speciale gevallen uitgezonderd, met name indien drie centroïden op één lijn zouden liggen, of vier centroïden collineair zijn (de hoekpunten van een trapezium vormen)). Deze ruimte van g—1 dimensies noemt men de discriminant-ruimte. Door alle observaties x in de discriminant-ruimte te projecteren kan men een duidelijk beeld krijgen van de ligging van verschillende groepen of condities ten opzichte van elkaar. Van groot belang is daarbij dat men de discriminant-dimensies kan interpreteren. De discriminant-functie v_1 biedt op zich zelf weinig houvast voor de interpretatie. Het duidelijkst zijn de correlaties tussen de geobserveerde componenten en de discriminant-dimensies; daarmee kan men de discriminant-dimensies interpreteren in termen van de geobserveerde componenten, zoals men bij een factor-analyse ook doet.

Het is mogelijk en nuttig om met een significantietoets te beslissen of er inderdaad significante verschillen tussen de groepen bestaan, en welke discriminant-dimensies als niet-significant mogen worden verwaarloosd.

De zojuist besproken methode is niets anders dan de eenvoudigste vorm van multivariate variantie-analyse. Men toetst de hypothese dat de centroïden van een aantal groepen verschillend zijn, zoals men bij univariate variantie-analyse de hypothese toetst dat de gemiddelden der groepen verschillend zijn.

Multivariate variantie-analyse kan een veel nauwkeuriger inzicht geven in het effect van variaties in condities (groepen), dan een aantal univariate analyses van hetzelfde effect.

Op de toetsingsmethode gaan we nader in. De toetsingsgrootheid kan op verschillende manieren worden geschreven. Eén manier is[10]

$$(41) \qquad U = \prod_r (1 - \eta_r^2) \qquad\qquad r = 1, \ldots, g\text{–}1$$

waarin Π wil zeggen dat we het voortgezette product moeten nemen van termen die de residuele varianties aangeven op elk der dimensies. Men kan U een multivariate (gegeneraliseerde) residu-variantie noemen. De volgende grootheid heeft nu bij benadering een chi-kwadraat-distributie:

$$(42) \qquad \chi^2 \cong -h \; {}_e\log U, \text{ met } h = n_t - 1 - \tfrac{1}{2}(m + g)$$

en m(g–1) vrijheidsgraden

Als x uit slechts een component bestaat (dus het geval van een univariate variantie-analyse met één factor), is U direct verwant met de F-toets:

$$(43) \qquad \frac{1 - U_{(m=1)}}{U_{(m=1)}} \times \frac{n_t - g}{g - 1} = F_{(g-1),\,(n_t - g)}$$

Om de significantie van de tweede en volgende discriminant-dimensie te toetsen gebruikt men

$$(44) \qquad U_s = \prod_r (1 - \eta_r^2) \quad r = s, \ldots, g-1; \quad s = 2, \ldots, g-1.$$

De chi-kwadraat-benadering heeft dan (m — s + 1) (g — 1) vrijheidsgraden.

Het hierboven besproken procédé voor het toetsen van multivariate verschillen tussen groepen of condities kan gemakkelijk worden uitgebreid tot gevallen waarin van meer dan een groepsindeling sprake is, zoals bij een multi-factor variantie-analyse. We hebben bv. een indeling van subjecten naar leeftijdscategorieën (A) en naar inkomenscategorieën (B). We moeten nu echter wel voor elke variantiebron, met inbegrip van de interactie, een matrix van productsommen tussen groepen berekenen, en wel

$$(45) \qquad B_a = b_{jk}^{(a)} = \sum_h n_h (\bar{x}_{hj} - \bar{x}_j)(\bar{x}_{hk} - \bar{x}_k)$$

waarin $\bar{x}_{hj}$ het gemiddelde van de j^{de} component van x in de h^{de} groep van factor A is. Op analoge wijze berekent men de 'tussen' productsommen voor factor B. De 'binnen' productsommen-matrix is:

$$(46) \qquad W = w_{jk} = \sum_i \sum_c (x_{ij} - \bar{x}_{cj})(x_{ik} - \bar{x}_{ck})$$

waarin nu $\bar{x}_{cj}$ het gemiddelde is van de j^{de} component in de c^{de} 'cel' van de combinaties der factoren. Als men tenslotte de totale productsommen-matrix ook heeft berekend vindt men de interactiematrix het gemakkelijkst door te nemen

$$(47) \qquad B_{ab} = T - B_a - B_b - W$$

Om nu de analyse te voltooien moet men een B-matrix vergelijken met de W-matrix. Men vindt dan voor elke variantie-bron de η's door op te lossen:

$$(48) \qquad \{B_{.} - \eta_r^2 (B_{.} + W)\} \, v_{.r} = o$$

Men vindt dus voor elke variantie-bron een discriminant-ruimte, en de toetsing van de significantie van elke dimensie in elk van deze ruimten geschiedt analoog aan de in het voorgaande besproken methode. Overigens is men meestal alleen geïnteresseerd in de variantie-bronnen in het algemeen, zonder dat men het aantal significante dimensies wil weten. In dat geval past men alleen (41) toe, die ook kan worden berekend zonder elk der η's afzonderlijk te berekenen, en wel door de determinant-formule:

$$(49) \qquad U = \frac{|W|}{|W+B|}$$

Meer bijzonderheden kan men vinden bij Jones (1966) en Bock (1966), die ook uitgewerkte voorbeelden geven.

X. Besluit

Het bovenstaande is niet meer dan een overzicht van de hoofdzaken van multivariate methoden. In veel gevallen zal niet met één bepaalde analyse volstaan worden, en zal men, naar gelang de vraagstelling van een onderzoek meer dan één methode in combinatie moeten gebruiken.

Een aantal specifieke problemen hebben we niet besproken. Het belangrijkste is steeds de keuze van de variabelen, en de wijze waarop zij worden gemeten. Alle gangbare multivariate methoden vereisen data op het niveau van een intervalschaal, en zijn beperkt tot het onderzoek van lineaire verbanden. Men zij dan ook voorzichtig in het hanteren van deze methoden indien niet aan die beperkingen kan worden voldaan. Vaak zal men data die niet bij benadering normaal verdeeld zijn eerst moeten transformeren (zie b.v. Kirk, p. 63).

Verschillende meer gespecialiseerde methoden konden binnen het bestek van dit artikel evenmin ter sprake komen. Volledigheidshalve noemen we: Scheve rotatie van factor-matrices (zie Harman, 1968), rotatie van factor-matrices naar een criterium (Schönemann, 1966; Browne, 1967; Gruvaeus, 1970), drie-modale factoranalyse, waarbij niet alleen variabelen en subjecten, maar ook variaties in condities worden betrokken (Levine, 1965) en multi-trait-multi-method analyses (Campbell & Fiske, 1959; Boruch et al., 1970).

Toepassing van multivariate analyses is zonder computer vrijwel uitgesloten. Het Tijdschrift *Behavior Science* publiceert regelmatig beschrijvingen van computerprogramma's. In Nederland bezit het SISWO (Amsterdam) een documentatie van computerprogramma's.

Summary

This article reviews the most commonly used techniques for multivariate analysis and some of its applications in the behavioural sciences. The techniques discussed are principal axes analysis, factor analysis, multidimensional scaling, partial, multiple and canonical correlation technique, and multivariate analysis of variance. Essential formulas are presented at an introductory level.

Literatuur

Bock, R. D., Contributions of Multivariate Experimental Designs to Educational Research, in Cattel (Ed.), *Handbook of Multivariate Experimental Psychology*, Rand McNally, Chicago, 1966.
Boon van Ostade, A. H., Over type onderzoek, *Nederlands Tijdschrift voor psychologie*, 1971, *26*, 82–101.
Boruch, R. F., Larkin, J. D., Wolins, L. & McKinney, A. C. Alternative Methods of Analysis: multitrait-multimethod data, *Educational Psychol. Measurement*, 1970, *30*, 833–853.
Browne, M. W., On Oblique procrustes rotation, *Psychometrika*, 1967, *32*, 125–132.
Bijnen, E. J., Cluster-analyse. *Overzicht en Evaluatie van Technieken*, (Diss., Tilburg), 1969.

Campbell, D. T. & Fiske, D. W., Convergent and discriminant validation by the multitrait-multimethod matrix, *Psych. Bull.*, 1959, *56*, 81–105.
Cattel, R. B., Coulter M. A., and Tsujoka, B., The Taxonometric recognition of Types and Functional Emergents, in Cattel, R. B. (Ed) *Handbook of Multivariate Experimental Psychology*, Rand McNally, Chicago, 1966.
Cooley, W.W., Lohnes & P.R., *Multivariate Procedures for the Behavioral Sciences*, New York, 1962.

Delbeke, L., *The construction of preference spaces*, Leuven, 1969.
Digman, J. M., Interaction and non-linearity in Multivariate experiment, in Cattel, R. (Ed.) *Handbook of Multivariate Experimental Psychology*, Rand McNally, Chicago, 1966.

Fleiss, J. L., & Zubin, J., On the meth. and theory of clus, *Mul. var. Beh. Res. 69*, 235–250.

Guttman, L. *Image theory for the structure of quantitative variables*, *Psychometrika*, 1953, *18*, 277–296, (a).
Guttman, L., A new approach to factor-analysis: the radex, in: P. Lazerfeld, Ed., *Mathematical Thinking in the social sciences*, New York, Univ. of Columbia Press, 1953 (b).
Guttman, L., *The structure of intelligence tests*, Initional Conference in Testing problems, Educational Testing Service, Princeton N.J., Oct. 1964, p. 25–36.
Geer, v.d. J. P., *Multivariate Analyse*, Arnhem (Van Loghum Slaterus), 1967.
Guilford, J. P., *Fundamental statistics in Psychology and Education*, McGraw-Hill, 3d ed., 1956.
Gruvaeus, G. T., A general approach to procrustes pattern rotation, *Psychometrika*, 1970, *35*, 493–505.

Hamaker, H. C., A Multiple regression Analysis, *Statistica Neerlandia*, 1962, *16*, 31–56.
Harman, H. H., Modern Factor-Analysis, 2nd ed., 1968
Herzberg, P.A., *The parameters of cross-validation;* Psychometric Monographs, no. 16, 1969.
Horst, P., Matrix Algebra for Social Scientists, Holt, Rinehart and Winston, 1963.

Jones, L., Analysis of Variance in its Multivariate Developments, in Cattel (Ed.)., *Handbook of Multivariate Experimental Psychology*, Rand McNally, Chicago, 1966.

Levine, J., Three mode factor-analysis, *Psychol. Bull.*, 1965, *64*, 442–452.

Lienert, G. A., Auswertung mehrdimensionaler qualitativer Daten mittels Konfigurationsfrequenzanalyse. *Methodik der Information in der Medizin*, 1971.

Pawlik, K., *Dimensionen des Verhaltens*, Huber, Bern–Stuttgart, 1968.

Schoenemann, P. H., A generalized solution to the orthogonal procrustes problem, *Psychometrika*, 1966, *31*, 1–10.
Spearman, The proof and measurement of association between two things, *Amer. J. Psychol.*, 1904, *15*, 72–101.

Thurstone, L. L., Multiple factor Analysis, *Psychol. Review*, 1931, *38*, 406–427.
Torgerson, W. S., *Theory and methods of Scaling*, Wiley, 1958.

Wishart, D., Mode analysis, A generalization of Nearest Neighbour which reduces chaining effects, in *Numerical Taxonomy* (Ed. Cole), Academic Press, 1969.

Noten

1. Als de componenten van X gestandaardiseerd zijn op variantie = 1 en gemiddelde = 0, dan is $F = \frac{1}{n} X'YA^{-\frac{1}{2}} = \frac{1}{n} X'XVA^{-\frac{1}{2}} = RVA^{-\frac{1}{2}} = VA^{1}V'VA^{-\frac{1}{2}} = VA^{\frac{1}{2}}$.
2. Deze specifieke factor kan ook als onsystematische 'random' variatie worden opgevat.
3. De aanduiding 'Q-analyse' is van Cattel, en heeft niets te maken met het gebruik van het symbool Q in dit artikel.
4. De inverse C^{-1} van een matrix C heeft de eigenschap dat $\Sigma c^{jl} c_{li} = 1$ als $j = i$ en $= 0$ als $j \neq i$, zodat $C^{-1} C = I$ (een matrix met uitsluitend 1 op de diagonaal).

4bis Op de berekening van de inverse matrix R^{-1} in (27) gaan we niet in. Een methode om de regressie-gewichten te berekenen, is beschreven in Guilford, 1956, pp. 405–415.

5. Zie Hamaker (1962) voor een voorbeeld.
6. Zie b.v. Digman (1966) en Lienert (1971).
7. Uitgevoerd door J. van Maren, drs. neerl.
8. Onderzoek uitgevoerd door een projectgroep van de vakgroep functieleer van het psychologisch laboratorium van de Universiteit te Nijmegen.
9. $$S^2_{tot.} = \frac{\sum_{i} (x_{i(h)} - \bar{x}_h + \bar{x}_h - \bar{x}_{tot.})^2}{n_t}$$
$$= \frac{\sum_{h} \sum_{i(h)} (x_{i(h)} - \bar{x}_k)^2}{n_{tot.}} + \frac{\sum_{h} n_h (\bar{x}_h - \bar{x}_{tot.})^2}{n_{tot.}}$$
waarin $\bar{x}_h$ = gemiddelde in de h^{de} groep, $\bar{x}_{tot.}$ = algemeen gemiddelde. Als n_h = constant voor alle groepen, is de tussen groepen-variantie, eenvoudig:
$\sum_{h} (\bar{x}_h - \bar{x}_t)/g$, $\quad h = 1, \ldots g$ (g = aantal groepen).
10. Ook wel 'Wilks' Lambda' genoemd. Wilks' Λ is echter oorspronkelijk gedefinieerd als $U^{\frac{1}{2}n}$.

Wat is factoranalyse?

L. C. A. Corsten

1. Inleiding

Het is enige tientallen jaren lang gebruikelijk geweest bij toepassers van factoranalyse zich op een enigszins mysterieuze wijze over dit onderwerp uit te drukken, waardoor er sprake leek te zijn van een zich tussen wiskunde en filosofie afspelend gebeuren. Dit, gevoegd bij een hoogst onduidelijke wijze van naar voren brengen van de gedachten zowel als van toegepaste technieken, leidde tot een onmatige verering door hen die maar ten dele begrijpen wat er geschiedt en tot een krachtige onbarmhartige afwijzing door hen die geacht worden de wiskundige statistiek te beoefenen. De onmatige verering is nog toegenomen, nu computers een groot deel van het omvangrijke rekenwerk kunnen verrichten, waarbij uit het oog verloren wordt dat computers niet in staat zijn gezond verstand en inzicht te *vervangen*, maar wel te *ondersteunen*.

In beide extreme houdingen is de laatste tijd een verbetering ten goede mogelijk geworden nu de wiskundige achtergrond duidelijk, of althans veel duidelijker dan voorheen, geformuleerd kan worden, waardoor men de mogelijkheden èn beperkingen aan het licht kan brengen en onderzoeken. Door het gebruik van klare taal kunnen de overschattingen van de verdienstelijkheid van de methode worden uitgebannen en kan men een realistische, van speculatieve beschouwingen gespeende werkwijze presenteren. Een gevoel van bewondering voor de intuïtie en fantasie van de scheppers van de methode zou men overigens slechts ten onrechte onderdrukken. Bovendien blijkt, zoals bij elke stap in wetenschappelijk denken b.v. na het uitvoeren van een eenvoudige t-toets of de berekening van een correlatiecoëfficiënt, dat kennis van zaken en fantasie een blijvende noodzaak zijn om tot een bruikbare interpretatie te komen van de resultaten. Berekeningen zonder interpretatie leiden tot niets.

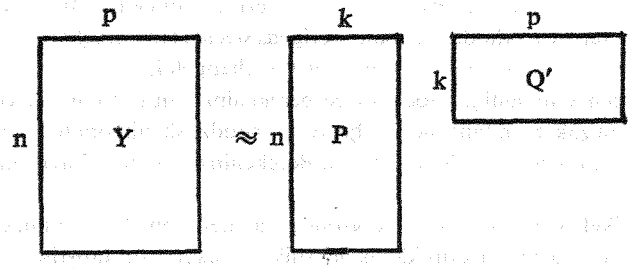
In dit artikel wordt iets over de grondgedachten van de factoranalyse medegedeeld. Het is daarbij onvermijdelijk gebruik te maken van statistische begrippen, zoals variantie, standaardafwijking, covariantie, correlatie e.d., en voorts van enige eigenschappen van vectorruimten, matrices en het begrip van eigenwaarden en

-vectoren van een matrix. Bewijzen van beweringen geven wij hier niet. Sommige zeer recente inzichten zijn verwerkt in deze uiteenzetting, die uitgaande van benaderingsmethoden uitmondt in wat met recht factoranalyse genoemd wordt. Door het volgen van de gedachtengang worden het bereikbare en het onbereikbare, naar wij hopen, tevens duidelijk.

2. Kleinste-kwadratenbenadering van een matrix

Ofschoon de auteur niets van marktonderzoek weet, wordt het voorbeeld gebruikt van gegevens, die betrekking hebben op verschillende gezinnen ($i = 1, \ldots, n$), al of niet uit een bepaalde sociale of sociaal-economische laag van een bevolking of bevolkingssegment afkomstig. Elk van de n gezinnen levert in een vaste volgorde p gegevens over zijn bestedingspatroon in een bepaald tijdvak, hetzij over het geheel (inclusief spaarbedragen, enz.), hetzij in detail over bepaalde sectoren b.v. verschillende onderdelen van primaire levensbehoeften of van een pakket zuivelartikelen of van mogelijke componenten van vrijetijdsbesteding, al of niet met inachtneming van het handelsmerk, enz. Elk van die p gegevens van een bepaalde soort krijgt een index j ($j = 1, \ldots, p$) en het j-de gegeven behorend bij gezin i wordt y_{ij} genoemd. Al deze np gegevens worden weergegeven in een matrix $(y_{ij}) = Y$, met n rijen (elke rij correspondeert met één gezin) en p kolommen (elke kolom betreft één soort gegeven, één van de p gemeten eigenschappen).

Aangezien het vrijwel niemand gegeven is in zulk een verzameling van getallen, zeker als n en p groot zijn, grote lijnen te ontdekken, kan men in het streven naar overzichtelijkheid een benadering van de matrix Y, en daarmede van de getallen y_{ij}, tot stand brengen door middel van het produkt van twee andere matrices, P en Q′. Hierin heeft P eveneens n rijen, maar slechts $k < p$ kolommen en Q, waarvan Q′ de getransponeerde of gespiegelde is, heeft p rijen en $k < p$ kolommen. Wij wensen m.a.w. Y te benaderen door PQ′, aldus uitgedrukt:



en dus $y_{ij} \approx \sum_{r=1}^{k} p_{ir}\, q_{ir}$.

De getallen in P en Q zijn vooralsnog onbekend; slechts het aantal k is gekozen. Het aantal getallen rechts, $k(n + p)$, is geringer dan het aantal links, np, en dat des te sterker naarmate k kleiner is. In het algemeen zal PQ', zelfs bij de meest geraffineerde keuze van P en Q niet gelijk zijn aan Y; vandaar het teken '$\approx$' voor 'wordt benaderd door'.

Wij wensen nu bij gegeven Y en k de benadering door PQ' zo goed mogelijk te doen zijn. Als maat f voor de kwaliteit van de benadering wordt gebruikt de som van de kwadraten van de np getallen in $Y - PQ'$. Deze f heeft o.a. het voordeel groter te worden naarmate de kwaliteit slechter is, en nul als de kwaliteit niet te verbeteren is, nl. als $Y = PQ'$. Wij zoeken dus P en Q zodat

$$\mathrm{tr}\,\{(Y - PQ')(Y - PQ')'\} = f$$

minimaal is. Opgemerkt wordt dat als men een best paar P en Q gevonden zou hebben er oneindig veel andere even goede paren zijn, nl.

$$P^* = PR' \text{ en } Q^* = QR^{-1}$$

waarin R een willekeurige niet-singuliere vierkante matrix van orde k is; immers $P^*Q^{*\prime} = PQ'$. Ondanks de vele mogelijkheden voor de beste P en Q blijkt er toch iets in het algemeen bij gegeven Y en k uniek te zijn, nl. het produkt PQ' zelf, en voorts de deelruimte van R^n opgespannen door de kolommen van P, evenals de deelruimte van R^p opgespannen door de kolommen van Q.

Men kan aantonen dat één (en daaruit ook elke andere) oplossing voor de benadering gevonden wordt als oplossing van de eigenwaarden-vergelijkingen:

$$YY'P = P\Lambda \quad (1)$$
$$Y'YQ = Q\Lambda \quad (2)$$

In de diagonale matrix Λ bevinden zich in volgorde van niet-afnemende grootte de k grootste eigenwaarden, elk zo vaak als haar multipliciteit bedraagt. De bijbehorende loodrechte kolommen van P en Q worden in hun lengte vastgelegd door

$$P'P = I \quad (3)$$

d.w.z. de kolommen van P zijn orthonormaal, en dientengevolge,

$$Q'Q = \Lambda \quad (4)$$

d.w.z. elke kolom van Q krijgt een lengte gelijk aan de wortel uit de bijbehorende eigenwaarde. Hieraan kan nog worden toegevoegd dat niet beide vergelijkingen (1) en (2) behoeven te worden opgelost, een taak meestal toevertrouwd aan een computer, maar slechts één van beide, want de andere volgt daaruit wegens:

$$Q = Y'P \quad (5)$$
$$P = YQ(Q'Q)^{-1} = YQ\Lambda^{-1} \quad (6)$$

3. De betekenis van de benadering

Door de benadering van Y worden nieuwe niet rechtstreeks gemeten eigenschappen, k in aantal, geïntroduceerd. Hun waarden staan in een vaste volgorde per gezin genoteerd in de overeenkomstige rij van P. Men zou hierbij eventueel kunnen denken aan b.v. mate van voorkeur voor goedkope voedingsmiddelen, statusbewustzijn, selectiviteit voor kwaliteit in het materiële dan wel in het culturele vlak, spaarzin enz.; maar in wezen gaat het om niet meer dan naamloze hersenspinsels. Elke waarneming y_{ij} wordt beschouwd als bij benadering een lineaire combinatie van deze nieuwe eigenschappen, nl.

$$\sum_{r=1}^{k} p_{ir}\, q_{jr}.$$

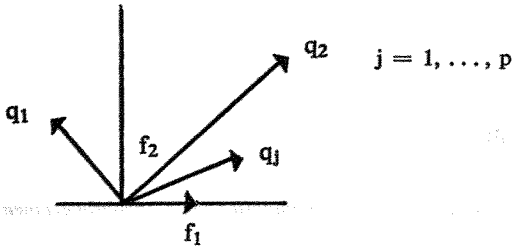
De nieuwe eigenschappen worden wel betiteld met de naam '*factoren*', de waarden ervan voor een bepaald gezin $p_{i1}, \ldots, p_{ik}$, de '*factorscores*' voor dat i-de gezin, genoteerd in de i-de rij van P, en de coëfficiënten $q_{j1}, \ldots, q_{jr}$ de '*factorladingen*' van de j-de oorspronkelijke eigenschap, genoteerd in de j-de rij van Q. Elke kolom van P bevat derhalve de factorwaarden of -scores van de n gezinnen voor één bepaalde factor. Elke kolom van Q bevat de factorladingen van één bepaalde factor in elk van de p oorspronkelijke eigenschappen. Het inwendig produkt van de rij van factorscores voor gezin i met de rij van factorladingen voor de j-de eigenschap geeft de benadering van y_{ij}.

Hiermede wordt de mogelijkheid geschapen (weliswaar bij benadering) inzicht te verkrijgen in de structuur van Y. Dit kan het beste geïllustreerd worden aan de hand van het bijzondere geval, waarin k, het aantal factoren, gelijk aan 2 is, hetgeen wij in de plaatjes zullen aannemen, maar de beschouwingswijze is ook geldig voor andere waarden van k.

In het vervolg wordt het inwendig produkt van kolom i van Y met kolom j van Y de covariantie van de i-de en de j-de eigenschap genoemd, waarbij de gebruikelijke deling door n als hinderlijk verwaarloosd wordt en waarbij ook in het midden wordt gelaten of de kolommen van Y gereduceerd (verminderd met het bijbehorend kolomgemiddelde) zijn of niet. Soms is dit gewenst, soms niet, maar de beschouwing wordt er in dit stadium niet anders door. De reductie kan ook afgedwongen worden door in de matrix Q in ieder geval een kolom van uitsluitend enen, of iets wat daarmee evenredig is, op te nemen.

Bij de keuze volgens de vergelijkingen (1) tot en met (6) bestaat de benadering van de j-de kolom van Y uit een lineaire combinatie (volgens de j-de rij van Q) van de kolommen van P. Wegens (5) bestaat die j-de rij van Q (factorladingen van eigenschap j) uit de coördinaten van de loodrechte projectie van diezelfde j-de kolom van Y op de door de k kolommen van P opgespannen ruimte, nl. de inwendige produkten van die kolom van Y met de k orthonormale kolommen in P. Door nu de k

kolommen van P (corresponderend met de k factoren en hun waarden) door een orthonormaal assenstelsel $f_1, \ldots, f_k$ voor te stellen (zie verg. (3)), wordt door een pijl met coördinaten $(q_{j1}, \ldots, q_{jk}) = q_j$ de benadering van de j-de kolom van Y (de j-de eigenschap), d.i. de j-de kolom van PQ′, exact voorgesteld in dat assenstelsel, en dat voor elke j.

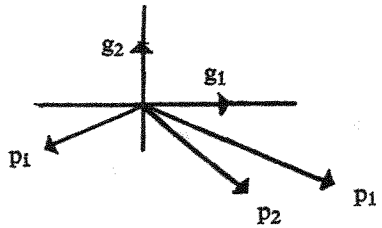


Figuur 1. De benadering q_j der kolommen van Y als lineaire combinaties van de factoren f (kolommen van P)

Hiermede wordt het complex van factoren, factorladingen en eigenschappen in beeld gebracht (fig. 1). Verder geldt voor de matrix van varianties en covarianties van de kolommen van de benadering PQ′ voor Y: QP′PQ′ = QQ′, d.w.z. het inwendig produkt tussen 2 pijlen van het type q_j is gelijk aan de overeenkomstige covariantie in de benadering van Y. De lengte van de pijl q_j is gelijk aan de standaardafwijking van de benadering van de j-de eigenschap en de cosinus van de hoek tussen 2 dergelijke pijlen die voor de bijbehorende correlatiecoëfficiënt. Men kan trouwens aantonen dat QQ′ een relatief nog betere benadering is van Y′Y, dan PQ′ dat is van Y. Uit zulk een voorstelling kan men o.a. zien welke eigenschappen of groepen daarvan sterk onderling gecorreleerd zijn en welke niet.

De benadering van y_{ij} door $\sum_{r=1}^{k} p_{ir}q_{jr}$ kan ook aldus worden gelezen. Er worden k 'standaardgezinnen' (eveneens hersenspinsels) gevormd, elk met hun waarden voor het p-tal eigenschappen. De waarden van de p eigenschappen staan voor elk standaardgezin genoteerd in de overeenkomstige kolom van Q. Nu wordt de i-de rij van Y (het i-de gezin) benaderd door de k standaardgezinnen volgens de i-de rij van P lineair te combineren. Die i-de rij van P (vroeger factorscores van gezin i genoemd), p_i, bevat volgens (6) tevens de coëfficiënten bij loodrecht projecteren van de i-de rij van Y op de ruimte opgespannen door (niet-orthonormale, maar orthogonale) kolommen van Q. Dus bestaat p_i uit de coëfficiënten van de loodrechte projectie van de i-de rij van Y (het i-de gezin) op de ruimte opgespannen door de k kolommen van Q (de k standaardgezinnen). Ondanks het feit dat deze standaardgezinnen lengten gegeven door $\Lambda^{\frac{1}{2}}$ hebben, stellen wij ze voor door een

orthonormaal stelsel $g_1, \ldots, g_k$ en vervolgens het i-de gezin door een pijl met de coördinaten $(p_{i1}, \ldots, p_{ik})$ van p_i, en dit voor elke i.



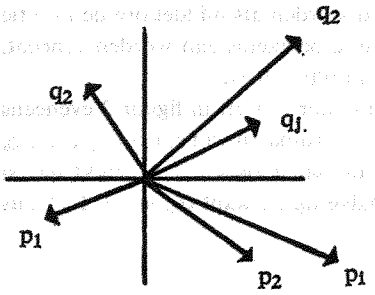
$i = 1, \ldots, n$

Figuur 2. De benadering p_i der rijen van Y als lineaire combinaties van standaardgezinnen g (kolommen van Q)

Door deze voorstelling (zie fig. 2) van het complex van standaardgezinnen, factorscores en gezinnen worden twee interessante dingen bereikt naast de reeds geboden mogelijkheid vast te stellen welke gezinnen sterk onderling verwant zijn en welke niet (groepsvorming!). Ten eerste wordt met de voorstelling van de kolommen van Q, $g_1, \ldots, g_k$, door een orthonormaal stelsel de pijl p_i niet de voorstelling van de i-de rij van $PQ' = YQ(Q'Q)^{-1}Q'$ maar van de i-de rij van
$YQ(Q'Q)^{-1}(Q'Q)^{-\frac{1}{2}}Q' = YQ(Q'Q)^{-\frac{1}{2}} \wedge^{-\frac{1}{2}} (Q'Q)^{-\frac{1}{2}} Q'$ d.w.z. van de loodrechte projectie van de i-de rij van Y op de door de (genormeerde) kolommen van Q opgespannen ruimte, gevolgd door vermenigvuldiging van de coördinant van die projectie in de richting van g_i met $\lambda_i^{-\frac{1}{2}}$ $(i = 1, \ldots, k)$. Terwijl de covariantiematrix van de kolommen van coördinaten (t.a.v. een orthonormale basis) in het eerste geval gelijk zou zijn aan die van $YQ(Q'Q)^{-\frac{1}{2}}$ d.i. $\wedge^{-\frac{1}{2}} Q'Y'YQ \wedge^{-\frac{1}{2}} = \wedge$, is dat nu die van de kolommen van $YQ(Q'Q)^{-\frac{1}{2}} \wedge^{-\frac{1}{2}}$ d.i. $\wedge^{-\frac{1}{2}} \wedge \wedge^{-\frac{1}{2}} = I$. Dit betekent dat de voorstelling van de gezinnen in een ruimte van dimensie k zowel t.o.v. het orthonormale stelsel $g_1, \ldots, g_k$, als t.o.v. elk ander orthonormaal stelsel in die ruimte (dus ook in fig. 2) varianties 1 en covarianties 0 heeft; m.a.w. er is een gestandaardiseerde projectie van de gezinnen tot stand gebracht.

In de tweede plaats is het mogelijk figuur 1 en figuur 2 op elkaar te leggenó, z dat f_i en g_i $(i = 1, \ldots, k)$ samenvallen, waardoor factoren en standaardgezinnen door dezelfde vectoren worden voorgesteld. Nu is het gewone inwendige produkt tussen pijl p_i en pijl q_j (lengte $p_i \times$ lengte $q_j \times$ cosinus van de ingesloten hoek) gelijk aan de benadering van y_{ij} zelf.

Men kan hiermede o.a. zichtbaar maken welke gezinnen t.a.v. een bepaalde gemeten eigenschap of een nieuwe eigenschap sterke of juist geringe variatie vertonen.

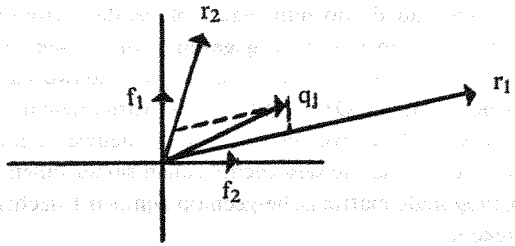


$i = 1 \ldots, n$
$j = 1 \ldots, p$

Figuur 3. Benadering van de matrix Y door inwendige produkten

4. De andere oplossingen P* en Q*

Bij beschouwing van de andere reeds vermelde mogelijkheden nl. $P^* = PR'$ en $Q^* = QR^{-1}$ blijkt dat in fig. 1 neer te komen op het vervangen van het orthonormaal assenstelsel $f_1, \ldots, f_k$ door een nieuw, meestal niet-orthogonaal assenstelsel $r_1, \ldots, r_k$ waarin de coördinaten van r_i t.o.v. $f_1, \ldots, f_k$ gegeven worden door de i-de rij van R ($i = 1, \ldots, k$), terwijl de q_j op hun plaats blijven (fig. 4).

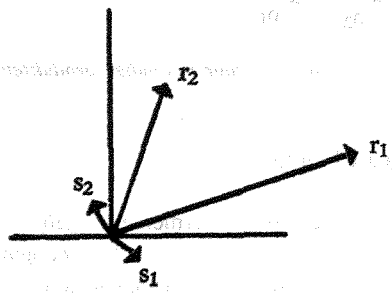


Figuur 4. Nieuwe factoren en nieuwe factorladingen voor de eigenschappen

Aldus worden nieuwe factoren, kolommen van $P^* = PR'$, voorgesteld door $r_1, \ldots, r_k$, geïntroduceerd met covariantiematrix $RP'PR' = RR'$, terwijl de nieuwe factorladingen, coördinaten van q_j t.o.v. $r_1, \ldots, r_k$, gegeven worden door de rijen van $Q^* = QR^{-1}$. Zulk een andere keuze van coördinaatassen bij gelijkblijvende configuratie der pijlen q_j kan worden gesuggereerd doordat men daarmee een k-tal bundels in de q_j goed meent te onderscheiden en zulk een nieuwe factor zelfs een naam meent te kunnen toekennen (zoals de genoemde selectiviteit, spaarzin, e.d.). Het is duidelijk dat hiervoor geen objectieve technieken kunnen bestaan; methoden zoals varimax (voor de introductie van een nieuw orthonormaal stelsel $r_1, \ldots, r_k$) en oblimax (voor de introductie van een nieuw niet-noodzakelijk orthonormaal stelsel)

en hun varianten kunnen hoogstens beschouwd worden als middel om de intuïtie een handje te helpen, maar er moet geen absolute betekenis aan worden gehecht, temeer daar het resultaat soms strijdig met de intuïtie blijkt.

De introductie van een nieuw assenstelsel in figuur 1 heeft in figuur 2 eveneens een nieuw assenstelsel $s_1, \ldots, s_k$ tengevolge. De coördinaten van s_i t.o.v. $g_1, \ldots, g_k$ worden gegeven door de i-de rij van $(R')^{-1}$ en, omdat $R'(R')^{-1} = I$, blijkt dat s_i loodrecht is op alle nieuwe assen in fig. 1 behalve op r_i, want $(r_i, s_i) = 1$. Aldus verschijnt fig. 5.



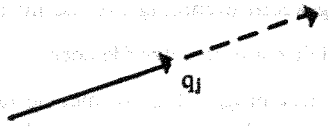
Figuur 5. Nieuwe assenstelsels $s_1, \ldots, s_k$ (voor fig. 2) en $r_i, \ldots, r_k$ (voor fig. 1) ten opzichte van elkaar

De nieuwe coördinaten van de op hun plaats blijvende vectoren p_i, dit zijn de nieuwe factorscores van gezin i, worden gegeven door de i-de rij van $P^* = PR'$. De nieuwe assen $s_1, \ldots, s_k$ zelf zijn de voorstelling van nieuwe standaardgezinnen, gegeven door de kolommen van $Q^* = QR^{-1}$. De covariantiematrix van de nieuwe factorscores is $RP'PR' = RR'$ evenals die van de nieuwe factoren. Opgemerkt wordt dat de twee nieuwe assenstelsels slechts zullen samenvallen als $R = (R')^{-1}$, d.w.z. als R een orthogonale matrix is, hetgeen op zijn beurt slechts een rotatie van de oude assen betekent.

Nieuwe assen, volgend uit andere oplossingen voor de benadering, laten derhalve de configuratie der p_i's en q_j's onderling en met elkaar ongemoeid; zij kunnen evenwel de interpretatie van de benadering vergemakkelijken.

5. Specifieke varianties

Men zal, ook in verband met het volgende, vaak willen weten in hoeverre de benadering van de varianties, dat zijn de diagonale elementen van $Y'Y$, door de kwadraten van q_j (fig. 1) geslaagd is. Men zou daartoe in de richting van q_j tevens een pijl beginnend in de oorsprong kunnen tekenen maar met een lengte gelijk aan de corresponderende standaardafwijking σ_j.



Een rechthoekige driehoek met de laatste pijl als schuine zijde en de eerste pijl als rechthoekszijde geeft een goed idee van die geslaagdheid. Het kwadraat van de andere rechthoekszijde $k_j^2 = \sigma_j^2 - q_j^2$ heet *specifieke variantie*; het is de component van de variantie σ_j^2 die bij de benadering moest worden verwaarloosd. De fractie q_j^2/σ_j^2 wordt wel communaliteit van eigenschap j genoemd en het complement $1 - q_j^2/\sigma_j^2$ de specificiteit, het in de benadering niet aan bod gekomen deel van de variantie.

6. Factoranalyse

Wat wij in de secties 2 t/m 5 vertoond hebben is nog geen factoranalyse, maar een uitbouw van wat *hoofdcomponenten-analyse* wordt genoemd. Er zijn namelijk twee nogal willekeurige elementen in de benaderingsmethode geïntroduceerd. De eerste daarvan is de keuze van het aantal k, hoewel men nog enigszins vrede zou kunnen hebben met de uitspraak dat men gemakshalve niet verder wenst te gaan dan een vooraf gekozen aantal factoren en dus ook standaardgezinnen. Bij zekere veronderstellingen omtrent het statistisch gedrag der waarnemingen kan men in principe ook wel toetsen of een gekozen k klein genoeg is, dan wel dient te worden vervangen door een grotere. Details blijven hier achterwege.

Een tweede veel ernstiger bezwaar is de keuze van het benaderingscriterium, nl. de kleinste-kwadratenmethode, ter bepaling van de ruimten voor de kolommen van Q en daarmee van die voor de kolommen van P. Volgens dit criterium doen alle p eigenschappen op gelijkwaardige wijze mee in die bepaling, of ze nu goed of slecht passen in de lager-dimensionale structuur. Het zou echter meer acceptabel zijn de eigenschappen met grote specificiteiten in mindere mate te laten meedoen dan die met geringe specificiteit. Daarbij komt het bezwaar dat de benadering afhangt van de min of meer toevallige keuze van de schaaleenheden, waarin de eigenschappen zijn gemeten. Aan het eerste deel van dit bezwaar komt men niet tegemoet, als men het tweede deel tracht te ondervangen door de metingen per eigenschap te standaardiseren, zodat ze alle variantie 1 hebben en $Y'Y$ door een correlatiematrix wordt vervangen; immers men weet nog steeds niet welke eigenschappen goed en welke slecht in de k-dimensionale deelruimte passen.

Een uitweg bestaat hierin dat men de waarnemingen in de j-de kolom van Y van

een gewicht voorziet omgekeerd evenredig met de nu nog onbekende specifieke variantie $\sigma_j^2 - q_j^2 = k_j^2$. Dit komt op hetzelfde neer als het toepassen van schaalveranderingen van die waarnemingen door ze met de (alsnog onbekende) k_j^{-1} te vermenigvuldigen. Aldus wordt Y vervangen door YK^{-1} waarin K een diagonale matrix van getallen k_j, j-de specifieke standaardafwijking, voorstelt. Door de vermenigvuldiging worden de specifieke standaardafwijkingen (en varianties) voor alle j gelijk aan 1.

De waarde van K wordt nu bepaald door de voorwaarde dat de benadering van de nieuwe covariantiematrix $K^{-1}Y'YK^{-1}$ na aftrek van de nieuwe specifieke varianties, d.i. de benadering van $K^{-1}Y'YK^{-1} - I$ door die met behulp van de bijbehorende PQ' d.w.z. door $QP'PQ' = QQ'$ waarin Q weer orthogonale eigenvectoren bevat van $K^{-1}Y'YK^{-1} - I$ met kwadraten gelijk aan de bijbehorende k (grootste) eigenwaarden, geen discrepantie ten aanzien van de varianties en de benaderingen daarvan vertoont. Anders gezegd: de diagonale elementen van $K^{-1}Y'YK^{-1} - I$ en van de bijbehorende benadering QQ' moeten exact aan elkaar gelijk zijn.

Dit criterium voor de bepaling van K dat op zichzelf reeds zeer acceptabel lijkt omdat de varianties in $K^{-1}Y'YK^{-1}$ en de benaderingen daarvan in $QQ' + I$ nu aan elkaar gelijk worden, wordt ook gevonden door toepassing van het z.g. normale factoranalyse-model, dat zegt dat de p normaal verdeelde eigenschappen van n onafhankelijke individuen een covariantiematrix hebben met de structuur $QQ' + K^2$, gevolgd door toepassing van de methode der grootste aannemelijkheid.

Op de wijze waarop de wegingsfactoren K^{-1} uit de vergelijking

$$\text{diag}\,(K^{-1}\,Y'YK^{-1} - I) = \text{diag}\,(QQ')$$

worden opgelost en de daarbij optredende moeilijkheden wordt hier niet ingegaan. Volstaan wordt met de opmerking dat dit probleem handelbaar is geworden (voor een computer). Het aangename feit doet zich nu ook voor dat men, ongeacht de verschillende schalen waarin de metingen kunnen worden uitgedrukt, steeds tot dezelfde vectoren Q komt, een z.g. fundamentele factorenbasis met dezelfde specificiteiten (en dus ook communaliteiten).

Na de vaststelling van K past men vervolgens op $K^{-1}Y$ in plaats van Y, de gehele voorafgaande hoofdcomponentenbeschouwing toe. En nu is er wél sprake van factoranalyse, zeker als men zich vergewist van de juiste keuze van het aantal k.

7. Slotwoord

Factoranalyse kan goeddeels worden beschouwd als hoofdcomponenten-analyse in een fundamentele schaal. Zij dient tot het overzichtelijk maken van metingen van

verschillende eigenschappen aan vele individuen (in dit geschrift: gezinnen). Men kan er groeperingen en andere configuraties in de individuen mee ontdekken, vaststellen welke eigenschappen nauw verwant zijn, welke eigenschappen of combinaties daarvan het meest voor verschillen tussen individuen aansprakelijk zijn, welke de meer en welke de minder zelfstandige (d.i. specifieke) eigenschappen zijn, welke relevant zijn in verband met zekere verschillen tussen individuen en tenslotte kan men op het spoor komen van nieuwe eigenschappen die het patroon in hoofdzaak lijken te beheersen.

Het is echter niet mogelijk met dit statistisch werktuig vast te stellen hoe variabelen elkaar in causale ketens of iets dergelijks beïnvloeden. Men dient ook de beruchte valkuil te vermijden die in zijn eenvoudigste vorm de waarnemer van nauwe correlatie tussen twee variabelen doet besluiten dat de een dan wel in hoofdzaak de oorzaak van de ander zal zijn.

Er is slechts sprake van een niet-onredelijke condensatie van waarnemingsmateriaal, waaruit men met gezond verstand nuttige hypothesen kan putten. Factoranalyse is een middel tot detectie van hypothesen; bewijskracht zal er zelden aan ontleend worden.

Summary

Factor analysis is introduced as a modification of principal component analysis and this in turn as a least squares approximation of a data matrix with a two-way classification (e.g. expenditures arranged according to families and different kinds of expenditure). The modification consists in a scale transformation which has to be derived from the data. The emphasis is on graphical representation of the data in view of the discovery of particular structures or relations in such a matrix.

Literatuur

Corsten, L. C. A., Over Factoranalyse, Statistica Neerlandica 22 (1968) 91–96.
Gabriel, K. R., The canonical decomposition of a matrix, Biometrika (in press).
Harman, H. H., Modern factor analysis. University of Chicago Press, 1967.
Lawley, D. N. and A. E. Maxwell, Factor analysis as a statistical method. Butterworth, London, 1963
Überla, K., Faktorenanalyse. Springer, Berlin, 1968.

Cluster-analyse*

E. J. Bijnen

1. Doel en plaats van cluster-analyse

Cluster-analyse is een van de analysetechnieken die zich in een toenemende belangstelling mag verheugen. Deze ontwikkeling doet zich in een aantal wetenschappen zoals in de psychologie, psychiatrie, biologie, sociologie, de medische wetenschappen, economie en archeologie gelijktijdig voor. Het is vooral in de eerste drie wetenschappen dat het begin van het gebruik van cluster-analyse heeft plaatsgehad en waar de vorderingen op dit gebied ook het grootst genoemd kunnen worden.

Bij de bewerking van onderzoeksmateriaal wordt cluster-analyse wel eens toegepast bij het vormen van 'homogene' groepen of clusters onderzoeksvariabelen, zoals houdingen, gedragingen, meningen, enz. Soms wil men nagaan welke variabelen sterk onderling samenhangen om tot een reductie in het aantal variabelen te komen, om daardoor een grotere greep op het verzamelde materiaal te krijgen. Men gaat er dan van uit dat deze samenvoegingen geschieden zonder aanmerkelijk verlies van relevante informatie. In een ander geval, waarop Johnson (1967, blz. 241) gewezen heeft, wil men in een vrij groot aantal vrij willekeurig verzamelde gegevens achteraf nagaan of er wellicht enige structuur in te onderkennen valt. Vanwege het ontbreken van een goede onderzoeksopzet, is dit een minder aantrekkelijk toepassingsgebied te noemen.

De clustertechnieken vinden hun voornaamste toepassingsgebied echter bij het vormen van groepen onderzoekseenheden zoals personen, bedrijven, verenigingen, enz. op basis van het geheel van verzamelde (en vermoedelijk relevante!) gegevens per onderzoekseenheid. Het analyseren van de gegevens van een onderzoekseenheid als geheel, dus als een patroon, kan zinvol zijn omdat patronen informatie kunnen verschaffen die de afzonderlijke variabelen te boven gaat. Meehl (1950) toonde aan dat het mogelijk is dat twee dichotome testvariabelen, die elk een fifty-fifty antwoordverdeling hebben op de dichotome criteriumvariabele, toch bruikbaar kunnen zijn

* Dit artikel is een bewerking van gedeelten van het proefschrift van de auteur: 'Cluster-analyse, overzicht en evaluatie van technieken', 1969, 150 blz.

bij het voorspellen van de criteriumvariabele, als men gebruik maakt van het antwoordpatroon op de beide testvariabelen (paradox van Meehl). Deze mogelijkheid doet zich voor indien de correlatie tussen beide testvariabelen bij de te onderscheiden criteriumgroepen sterk van elkaar verschillen.

Het doel van een cluster-analyse is vergelijkbare eenheden te groeperen en te onderscheiden van andere. Een cluster-analyse tracht dus groepen onderzoekseenheden of -variabelen te vormen, zodanig, dat de eenheden (variabelen) in een cluster 'veel' op elkaar gelijken maar 'niet veel' gelijkenis vertonen met eenheden (variabelen) buiten dat cluster, of zoals Ward en Hook (1963, blz. 69) zeggen:

'... desire to group large numbers of persons, jobs, or objects into smaller numbers of mutually exclusive classes in which the members have similar characteristics'.

Elke onderzoekseenheid of variabele is in een cluster-analyse dus slechts in één cluster vertegenwoordigd. Iedere clustertechniek geeft een eigen omschrijving van het 'veel' en 'niet veel' op elkaar gelijken. Men zal dan ook in elk geval na moeten gaan welke definitie de meest optimale lijkt, gedachtig de uitspraak van McQuitty (1966a, blz. 3): 'Whether or not types are found to exist depends in part on how they are defined'.

Clustertechnieken kunnen ook beschouwd worden als methoden voor het ontdekken van typen. Hempel en Oppenheim (1936) hanteerden reeds het volgende empirische typebegrip:

'die Objekte lassen sich in gewisse Gruppen zusammenfassen, derart, dass die Objekte einer Gruppe miteinander eine relativ grosse Ähnlichkeit aufweisen, während die Objekte verschiedener Gruppen sich hinsichtlich jener Eigenschaften relativ stark unterschieden'.

Deze omschrijving van het begrip type komt geheel overeen met de doelstelling van een cluster-analyse zoals wij die hierboven hebben gegeven. Cluster-analyse is dus één van de methoden tot het opstellen van typen, die, gebruik makende van empirische onderzoeksresultaten met behulp van wiskundige of statistische technieken homogene groepen onderscheiden, op basis van alle variabelen die in het onderzoek zijn betrokken (Capecchi, 1967, blz. 118–119). Clustertechnieken kunnen dus door middel van het zoeken naar typen, o.a. een bijdrage leveren tot de begrips- en theorievorming.

Een techniek die vaak in één adem met cluster-analyse genoemd wordt is factor-analyse. De doelstelling van factor-analyse is echter een enigszins andere dan die van een cluster-analyse. Het is een techniek die tracht een gegeven variabelen- of onderzoekseenhedenruimte weer te geven in een ruimte met een kleiner aantal dimensies zonder al te veel verlies aan informatie. Het is dan vereist dat de gegevens op interval niveau gemeten zijn. Bovendien beperkt men zich tot lineaire samenhangen. Het gebruik van factor-analyse stuit dus op vrij grote beperkingen.

Om het verschil tussen clusters en factoren aan te geven, stelt Cattell (1944, blz. 183)

'... that clusters are essentially representations at the descriptive level, and as such are little better than straight statements of the correlation coefficients, whereas factors are statements at the interpretive level. If the interpretations are correct the factors have more permanent value and far wider utility'.

Volgen wij echter de opvatting van Stouthard (1965, blz. 91) dat het doel of motief bij de toepassing van data-modellen (waarvan factor-analyse er één is) de materiaal-reductie is, dan moeten wij dit door Cattell gemaakte onderscheid als niet reëel beschouwen. Zowel factor-analyse als clusteranalyse zijn methoden die materiaal-reductie tot doel hebben, zij het niet op gelijke wijze. In termen van factor-analyse kan men stellen dat bij een cluster-analyse die variabelen of onderzoekseenheden bij elkaar geplaatst worden die een zelfde ladingenpatroon laten zien over de factoren die getrokken zijn.

De ontwikkeling van de cluster-analyse staat overigens niet geheel los van de factor-analyse. In de begintijd van de factor-analyse zijn er, vanwege de grote rekentechnische moeilijkheden, al vroeg methoden ontwikkeld om het aantal variabelen te reduceren, om zodoende factor-analyse mogelijk te maken. Daarnaast zijn er, al of niet als reactie op factor-analyse, methoden ontwikkeld die leiden tot samenvoeging van onderzoekseenheden of variabelen op basis van geheel andere voor-onderstellingen.

Bij het vormen van clusters kan men uitgaan van de groep als geheel en deze opdelen in homogene subgroepen, of men kan uitgaan van de afzonderlijke onderzoekseenheden of variabelen om door samenvoegingen tot clusters te komen. Onze voorkeur gaat uit naar de laatste benaderingswijze. Bij de eerstgenoemde methoden kunnen eenheden die een geringe gelijkenis vertonen met alle overige, maar ook weer niet zo sterk verschillen dat ze meteen als afzonderlijke 'groepen' worden afgesplitst, een belangrijke invloed op de te maken opdelingen hebben. Bovendien ligt het zeer voor de hand dat men, als clusters van variabelen of onderzoekseenheden gevormd moeten worden die onderling veel op elkaar gelijken, begint met de variabelen of eenheden die het sterkst op elkaar gelijken. Men heeft dan overigens geen garantie dat de 'beste' opdeling in een bepaald aantal clusters wordt verkregen.

De meer recente (na ± 1960 ontwikkelde) technieken van cluster-analyse hebben meestal betrekking op onderzoekseenheden. Sommige van de methoden kunnen echter zonder bezwaar ook voor variabelen worden gebruikt (zoals o.a. de te beschrijven syndroom-analyse van McQuitty). Bij een cluster-analyse van variabelen zal het bepalen van clusters meestal niet het uiteindelijke doel zijn. Doorgaans zal men een groot aantal scores willen vervangen door een geringer aantal. Men staat dan voor het probleem van de berekening van deze scores, omdat een eenvoudige somscore niet altijd zinvol is. Hiervoor moeten dan aparte technieken worden ontwikkeld, terwijl deze problemen bij o.a. factor-analyse reeds zijn opgelost. Bovendien zal bij variabelen een dimensie-analyse vaak aantrekkelijker zijn. Dit wil overigens

niet zeggen dat het zoeken naar clusters van variabelen niet zinvol kan zijn. Hilgendorf e.a. (1967) hebben aangetoond dat het gebruik van een techniek van McQuitty in combinatie met path-analyse een bijdrage kan leveren aan het opstellen van causale modellen.

Het toepassingsgebied van nagenoeg alle clustertechnieken is zeer ruim. Dit heeft tot gevolg dat in concrete gevallen waarin meer specifieke eisen gesteld moeten worden, de methoden niet geheel zullen voldoen. Door het opstellen van datamodellen kunnen wellicht nauwkeuriger methoden worden ontworpen voor toepassing op een beperkter gebied. Aan de methoden van McQuitty ligt reeds een data-model ten grondslag, maar dit is naar onze mening niet bruikbaar. Wij achten één of twee van zijn methoden wel bruikbaar, maar dan vanuit een geheel ander gezichtspunt.

In dit artikel zullen wij ons beperken tot twee methoden van McQuitty. Naar onze mening zijn ook de methoden die ontwikkeld zijn door Ward (1963), Johnson (1967) en Bridges (1966) zeer bruikbaar te noemen.

2. Coëfficiënten ter bepaling van de mate van gelijkheid van onderzoekseenheden

Als men zich ten doel stelt onderzoekseenheden die veel op elkaar lijken te groeperen, dan dient men over een coëfficiënt te beschikken die de mate van overeenkomst of van verschil tussen twee onderzoekseenheden aan kan geven. Men dient er zich van bewust te zijn (Cronbach en Gleeser, 1953, blz. 457 e.v.) dat de beschouwde gelijkheid geen algemene kwaliteit is, maar slechts betrekking heeft op die dimensies die in het onderzoek zijn betrokken, en dat berekening van een gelijkheidsindex reductie in de configuratie van de scores en dus verlies aan informatie betekent, zodat men rekening moet houden met de mogelijke gevolgen van de zogenaamde globale benaderingen.

Een van de meest eenvoudige coëfficiënten is de agreement score van McQuitty (1956). Deze is gedefinieerd als het aantal variabelen waarop beide respondenten een zelfde antwoord gegeven hebben. Heeft respondent A bij drie gebruikte variabelen respectievelijk de antwoordcategorieën 3, 2 en 2 aangegeven en respondent B respectievelijk 2, 2 en 3, dan is de agreement score tussen A en B één vanwege het gelijke antwoord op variabele twee. Het gebruik van deze coëfficiënt bij variabelen met een wisselend aantal antwoordcategorieën kan minder gewenst zijn, omdat de betekenis of het gewicht van een overeenkomst op een variabele dan mede afhankelijk is van dat aantal categorieën. Daarenboven geldt dat geen rekening wordt gehouden met de verdeling van de antwoorden over de antwoordcategorieën. Doorgaans lijkt dit laatste overigens van minder belang te zijn.

Als voor de gebruikte variabelen het 'transitory postulate' (Stephenson, 1953, blz. 48) opgaat, dat inhoudt dat uit fenotypische verschillen tot gelijkluidende

verschillen in de genotypische grootheid geconcludeerd moet kunnen worden (Stouthard, 1965, blz. 32), mogen we sommen, gemiddelden e.d. van de scores over de variabelen berekenen. In een dergelijk geval kan de product moment correlatiecoëfficiënt r als gelijkheidsmaatstaf gebruikt worden. Deze coëfficiënt standaardiseert echter per onderzoekseenheid naar gemiddelde en variantie, hetgeen ongewenst kan zijn. In plaats van r, die een maat voor gelijkheid is, zou dan ook D, een maat voor ongelijkheid, gebruikt kunnen worden. D is de afstand tussen twee punten in de Euclidische ruimte, waarbij de m gebruikte variabelen de assen vormen. De formule is $D_{ef} = \sqrt{\sum_j (X_{je} - X_{jf})^2}$, waarbij X_{je} de score van respondent e op variabele j aangeeft en X_{jf} de score van respondent f op die zelfde variabele. Bij het gebruik van D dient men er op bedacht te zijn dat de variabele met de grootste variantie ook de grootste invloed op deze coëfficiënt heeft. Door de variabelen vooraf te standaardiseren kan dit ondervangen worden.

Naast de hierboven aangegeven coëfficiënten zijn in de literatuur nog een groot aantal andere coëfficiënten voorgesteld. (Voor een overzicht zie: Bijnen, 1969, blz. 11–27 en Boon van Ostade, 1971). Het is niet mogelijk een coëfficiënt aan te geven die in het algemeen het meest aan te bevelen is. Bovendien is uit vergelijkende studies (Mosel en Roberts, 1954, Helmstadter, 1957, Muldoon en Ray, 1958) gebleken dat de coëfficiënten nogal afwijkende resultaten kunnen geven, terwijl de verschillen weer van geval tot geval anders kunnen zijn. Voor gegevens op interval niveau lijkt de D-coëfficiënt vaak een goede maat. Als het meten op een lager niveau heeft plaats gehad en men de relatieve gelijkheid van de onderzoekseenheden wil bepalen, is de agreement score van McQuitty wellicht de minst slechte keuze. Associatie-coëfficiënten voor nominale gegevens zullen doorgaans niet geschikt zijn voor het bovengestelde doel. Zij geven immers de mate van associatie aan, die zowel in de richting van overeenkomst als van geen overeenkomst kan gaan.

3. Twee methoden tot het opstellen van typen volgens McQuitty

3.1. *Inleiding*

Vanuit de klinische psychologie heeft McQuitty enige methoden ontwikkeld voor het analyseren van de antwoorden op een reeks vragen of een aantal kenmerken of eigenschappen van een aantal onderzoekseenheden. Het ging hem daarbij speciaal om typen psychische ziekten van elkaar te onderscheiden, hoewel hij zich bij de ontwikkeling van zijn methoden niet tot dit probleemgebied beperkt heeft. Wij zullen in het vervolg dan ook de woorden, personen en onderzoekseenheden als synoniemen van elkaar beschouwen. De antwoorden die door een subject gegeven worden op een groep situaties, zoals vragen van een test, wordt een antwoordpatroon genoemd.

Onder een type wordt verstaan een categorie binnen een bepaald classificatie-systeem, zodanig dat de leden van een type een combinatie van kenmerken hebben die uniek is voor hen, en waarbij deze kenmerken op een bijzondere wijze met elkaar samenhangen. (McQuitty, 1967, blz. 22).

Bij de verwerking van sociologisch en psychologisch onderzoeksmateriaal wordt er vaak van uitgegaan, dat de betekenis van een bepaald antwoord of kenmerk onafhankelijk is van de context waarin het is opgenomen. Meestal worden immers de items in een onderzoek afzonderlijk geanalyseerd. McQuitty gaat er nu van uit dat als er typen respondenten zijn, deze typen weerspiegeld zullen worden in de antwoorden op de tests of in andere gedragingen, niet in een 1–1 relatie tot antwoorden op een bepaalde vraag, maar in een 1–1 relatie tot een bepaald antwoordpatroon. Bij elk type behoort dus een uniek antwoordpatroon en elk antwoordpatroon van een categorie is een indicatie van het onderliggende type, waarbij natuurlijk wel verondersteld wordt dat de gestelde vragen relevant zijn. Hoewel verondersteld wordt dat de vragen een wisselende validiteit hebben, met andere woorden, dat het antwoord bij verschillende mensen door verschillende interne structuren (typen) bepaald wordt, wordt verondersteld dat de antwoordpatronen een constante validiteit hebben. Het is dan ook van belang vragen te gebruiken die een wisselende betekenis hebben naar gelang de combinatie met andere antwoorden waarin zij voorkomen. Het kan hierbij voorkomen dat een bepaald antwoord op een vraag als typisch voorkomt bij verschillende typen. Deze opvatting houdt in dat factor-analyse voor het gestelde doel ongeschikt is. Factor-analyse veronderstelt immers dat de variabelen voor iedereen hetzelfde betekenen, en daardoor kan een residuele correlatiematrix berekend worden.

Het gaat er nu om op basis van antwoordpatronen de onderliggende oorzaak, het type, te ontdekken. Er wordt van uitgegaan dat de typen zich kunnen manifesteren in vele soorten samenhangen, zodat het model niet beperkt mag zijn tot lineaire samenhangen zoals bij factor-analyse het geval is. Er moeten methoden ontwikkeld worden waarvan de analyse-uitkomsten zoveel mogelijk afhangen van de samenstelling van de data.

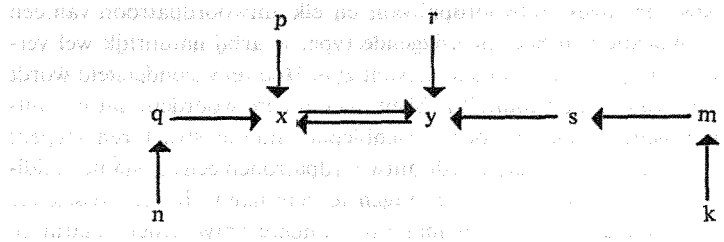
3.2. *Elementary linkage analysis*

Een type zou, zoals bij elementary linkage analysis (McQuitty, 1957, blz. 209), gedefinieerd kunnen worden als een zodanige groep onderzoekseenheden (personen) dat elk meer gemeen heeft met één ander lid van dat type dan met enig lid van een ander type, met andere woorden, iedereen is geclassificeerd met die ander waarmee hij de grootste overeenkomst heeft (McQuitty, 1967, blz. 23). Het samenstel van variabelen die alle leden van een type met elkaar gemeen hebben, zou als een prototype beschouwd kunnen worden. In plaats van agreement scores zou men volgens McQuitty hier ook correlatiecoëfficiënten of andere associatiecoëfficiënten kunnen gebruiken. De definitie van type moet dan in zoverre worden aangepast dat niet meer

wordt gesproken over kenmerken met elkaar gemeen hebben, maar over een relatieve gelijkenis of overeenkomst met elkaar bezitten.

De vorming van het eerste type zal dan ook uitgaan van het paar, dat het meest met elkaar gemeenschappelijk heeft, bijvoorbeeld x – y. Zij vormen een wederkerig paar, dat wil zeggen x heeft het meest met y gemeenschappelijk en y met x. Bij elk van deze twee worden degenen gezocht die met hen de grootste gelijkheidsindex hebben, bijvoorbeeld p, q, r en s. Vervolgens wordt voor de aldus toegevoegden nagegaan welke van de overige met deze hun hoogste index hebben, bijvoorbeeld n, m, enz. Schematisch weergegeven:

Schema 1:



Als er op deze wijze geen toevoegingen meer plaatsvinden, dan worden de geclassificeerden uit de matrix verwijderd en wordt de analyse herhaald met de overgeblevenen.

Een nadeel van deze methode is dat personen op grond van overeenkomsten op geheel andere variabelen tot het cluster kunnen worden toegevoegd dan anderen. Zo is het mogelijk dat de hoge overeenkomst van k met m op geheel andere variabelen betrekking heeft dan de overeenkomst van n met q.

Gezien de opvatting van de betekenis van de afzonderlijke kenmerken, is het hier niet mogelijk om een residuele matrix te berekenen zoals bij factor-analyse het gebruik is, waar de verklaarde variantie van de te verklaren variantie afgetrokken kan worden. Wel is het volgens McQuitty mogelijk om, als de elementary linkage analysis berust op correlatiecoëfficiënten, de gehechtheid (relevantie) van het type voor de personen aan te geven. McQuitty neemt hiertoe de ladingen op de eerste factor van de submatrix die de leden van het cluster bevat. Hij stelt immers dat de antwoorden binnen een type dezelfde betekenis hebben, zodat binnen de typen aan correlatiecoëfficiënten betekenis kan worden toegekend.

Een belangrijke beperking van de hierboven geschetste methode is wel, dat het aantal clusters uitsluitend afhankelijk is van het aantal wederkerige paren, waarvan elk cluster er één en slechts één van bevat (afgezien van ties). Immers, als er geen wederkerige paren zijn, zullen er bij de nog niet toegevoegden er altijd zijn die met leden van het eerste cluster het meeste gemeen hebben en dus worden opgenomen. Het feit dat er nog wederkerige paren zijn in de niet toegevoegden, betekent dat de toevoegingen voor het cluster op een bepaald moment gestopt moeten worden.

De elementary linkage analysis kan ook tot een hiërarchische methode worden uitgebreid, dat wil zeggen tot een methode die de gevonden typen op haar beurt tracht samen te voegen tot typen van een hogere orde (McQuitty, 1964, blz. 446). Voor elk cluster moet dan een patroon gekozen worden dat representatief geacht kan worden voor dat cluster. Deze selectie kan plaatsvinden op theoretische, praktische of statistische gronden. Op de matrix met agreement scores of andersoortige gelijkheidscoëfficiënten, kan voor deze gekozen referentie-personen weer de elementary linkage analysis worden toegepast. Dat het kiezen van deze referentiepersonen een vrij arbitraire zaak kan zijn behoeft verder geen betoog.

3.3 *Hierarchical syndrome analysis*

Een andere methode voor het opsporen van typen is de replacement versie van de hiërarchische syndroom-analyse (McQuitty, 1960b). Bij deze methode wordt er naar gestreefd dat elke persoon in een categorie geplaatst wordt zodanig dat hij meer lijkt op elk ander in die categorie dan op enig ander in enige andere categorie. Nemen wij als voorbeeld een gedeelte van de door McQuitty (1962, blz. 518 e.v.) gegeven toepassing. Hij heeft daarin een aantal bedrijven geclassificeerd op basis van 32 variabelen die aspecten van hun relatie met de vakbonden weergeven. De bedrijven waren onder andere twee bouwbedrijven (A en B), twee vervoersbedrijven (C en D), een meelfabriek (E) en een metaalfabriek (F). Tussen deze zes bedrijven ziet de matrix met de agreement scores er als volgt uit:

tabel 1: Agreement scores tussen zes bedrijven op basis van 32 variabelen.

	A	B	C	D	E	F
A		29	16	16	14	6
B	29		17	17	13	6
C	16	17		26	10	8
D	16	17	26		10	12
E	14	13	10	10		21
F	6	6	8	12	21	

Bij de hiërarchische syndroom-analyse wordt in de matrix met agreement scores in elke kolom de hoogste score onderstreept. Aldus vinden we het bedrijf waarmee die in de top van de tabel de grootste overeenkomst heeft. Van al deze onderstreepte scores wordt de hoogste genomen, hetgeen het paar aangeeft dat het meest gemeenschappelijk heeft en dus als uitgangspunt voor de analyse genomen zal worden. In tabel 1 blijkt dit paar AB te zijn. Zij worden in een categorie geplaatst, waarna de agreement scores van de categorie met de overige onderzoekseenheden bepaald moeten worden, dus in ons geval het aantal variabelen dat respectievelijk A, B en C, A, B en D, enz. met elkaar gemeen hebben. Volgens McQuitty (1960a, blz. 59 e.v.) kan men deze agreement scores schatten met behulp van de classificatie-assumptie.

Deze assumptie houdt in dat de elementen van een categorie (dus bijv. A, B en C) zoveel met elkaar gemeenschappelijk hebben als het paar dat uit hen gevormd kan worden dat het minst met elkaar gemeenschappelijk heeft. De op deze wijze geschatte agreement score is de bovengrens van de feitelijke agreement score, zodat zij meestal te hoog zal uitvallen. Maar als de eenheden tot een bepaald type behoren, zullen ze gekenmerkt worden door een groot aantal gemeenschappelijke eigenschappen, zodat het niet zeer onwaarschijnlijk is dat de feitelijke agreement score dicht bij de score van het paar met de laagste score ligt. Naarmate het aantal gemeenschappelijke variabelen tussen de paren een groter deel vormt van het totaal aantal, zal de kans op een goede benadering van de feitelijke score met behulp van de classificatie-assumptie groter worden. Bij samenvoegingen van meer eenheden zal echter het aantal gemeenschappelijke variabelen kleiner worden en dus de mogelijkheid van sterke afafwijkingen groter. Met andere woorden, de opgestelde redenering kan wel juist zijn, maar het is volstrekt niet zeker dat dit ook steeds het geval zal zijn. Als nu de agreement scores van de categorie met de overige onderzoekseenheden bepaald zijn, worden vervolgens de regels en de kolommen van de leden van de eerste categorie vervangen door die van de categorie. In deze matrix wordt opnieuw de hoogste score bepaald, enz.

Opgemerkt kan worden dat in tegenstelling tot elementary linkage analysis het aantal wederkerige paren niet bepalend is voor het aantal clusters. Een cluster in de syndroom-analyse kan meer dan één wederkerig paar in zich verenigen. Het is wel zo, dat elk wederkerig paar in de eerste samenvoegingen als de kern van een afzonderlijke categorie fungeert. Verder in de analyse kunnen deze echter worden samengevoegd.

Voor de gegevens uit tabel 1 verloopt dan de analyse (met gebruikmaking van de classificatie-assumptie) zoals weergegeven in tabel 2.

tabel 2: Replacement versie van de hiërarchische syndroom-analyse op basis van tabel 1

	A	B	C	D	E	F	AB	CD	EF	AB CD	ABC DEF
A		29	16	16	14	6					
B	29		17	17	13	6					
C	16	17		26	10	8	16				
D	16	17	26		10	12	16				
E	14	13	10	10		21	13	10			
F	6	6	8	12	21		6	8			
AB			16	16	13	6		16	6		
CD					10	8	16		8		
EF							6	8		6	
ABCD									6		
ABCDEF											

De schattingen van de agreement scores met behulp van de classificatie-assumptie kunnen zeer eenvoudig verricht worden. Moet bijvoorbeeld in ons voorbeeld de score voor ABC bepaald worden, dan valt van de paren AB, AC en BC paar AB zonder meer af, omdat op basis daarvan de categorie is gevormd. Er hoeft dus alleen nagegaan te worden of AC of BC het kleinste is, met andere woorden welke in de kolommen A en B op regel C het kleinste is. Bij samentrekking van een enkel individu met een categorie en van categorieën met elkaar, kan op gelijke wijze gehandeld worden.

Als in deze procedure twee personen i en j eenmaal bij elkaar gevoegd zijn, dan blijven zij ook in de hele verdere analyse bij elkaar. i en j kunnen dus niet afzonderlijk met k geclassificeerd worden, ook al heeft k meer met i of j gemeenschappelijk dan met enig ander. Er wordt geen rekening gehouden met de mogelijkheid dat de samenvoeging van i en j op grond van toeval, dus op grond van een reden onafhankelijk van de typologie, is geschied. Het is dus mogelijk dat een persoon niet met degene waarmee hij het meest gemeen heeft wordt samengevoegd. Aan de definitie van type is dan niet voldaan.

In de self-checking versie kunnen alle individuen geclassificeerd worden met die waarmee ze de grootste overeenkomst vertonen. De uit te voeren bewerkingen verschillen in zoverre van die van de replacement versie dat na de groepering alleen de kolommen (waarin de hoogste agreement score voorkomt) worden geschrapt, maar niet de regels. Hierdoor wordt verkregen dat de reeds gegroepeerde personen afzonderlijk blijven meelopen in de rest van de analyse.

Bij het ontbreken van elkaar overlappende typen geven de self-checking en de replacement versie, hetzelfde resultaat. Toegepast op ons voorbeeld krijgen we:

Tabel 3: Self-checking versie op basis van tabel 1

	A	B	C	D	E	F	AB	CD	EF	BCD	ABCD
A		29	16	16	14	6		16	6		
B	29		17	17	13	6		17	6		
C	16	17		26	10	8	16		8		
D	16	17	26		10	12	16		10		
E	14	13	10	10		21	13	10			10
F	6	6	8	12	21		6	8			6
AB			16	16	13	6		16	6		
CD	16	17			10	8	16		8		
EF	6	6	8	10			6	8			6
BCD											
ABCD					10	6			6		

De self-checking versie blijkt zeer arbeids-intensief te zijn. Bij materiaal waarin de onderzoekseenheden niet zeer duidelijk representanten zijn van typen, zullen de af-

zonderlijke eenheden vaak in meer dan één type voorkomen, hetgeen interpretatiemoeilijkheden met zich mee kan brengen.

De beide versies van de hiërarchische syndroom-analyse zijn volgens McQuitty (1960b, blz. 295) toepasbaar op alle soorten gelijkheidsindices. Deze methode maakt echter gebruik van de classificatie-assumptie hetgeen voor agreement scores te verdedigen valt, maar voor associatiecoëfficiënten en dergelijke een nadere argumentatie verdient, die, naar het ons toeschijnt, niet te geven valt. Er is immers een groot verschil in de aard van de agreement score en de associatiecoëfficiënten. Een agreement score heeft betrekking op een aantal kenmerken die de onderzoekseenheden gemeenschappelijk hebben. Associatiecoëfficiënten geven echter de mate van samengaan of overeenkomst tussen twee grootheden aan, hetgeen voor een combinatie van eenheden als een geheel met een andere eenheid niet aan te geven lijkt. Bij correlatiecoëfficiënten voor variabelen zou men in dit verband aan een multipele correlatie kunnen denken. Bij deze is het echter zo, dat van de groep variabelen eerst een lineaire combinatie gemaakt wordt voordat de correlatie berekend wordt. De variabelen in de combinatie worden hierbij als afzonderlijke eenheden gezien en niet als een geheel.

Toepassing van de methode op associatiecoëfficiënten zal dezelfde resultaten geven als toepassing op agreement scores, indien de indices die bij de samenvoegingen zijn gebruikt dezelfde rangvolgorde hebben als de agreement scores zouden hebben indien die op hetzelfde materiaal berekend zouden zijn. Hoewel McQuitty op een dergelijke (door hem te ruim gestelde) beperking wijst (1960b, blz. 295) blijkt in het vervolg van zijn betoog (blz. 299) niet dat hij de gevolgen van deze beperking zodanig acht dat een controle nodig is. Later (1966b) stelt hij voor, bij samenvoeging van categorieën een gemiddelde te nemen, eventueel een gewogen gemiddelde naar het aantal onderzoekseenheden in een categorie. Dit heeft echter als nadeel, dat bij een gewogen gemiddelde de bepaling van de gelijkheidsindex van een relatief sterk afwijkend individu met een categorie, de index in feite bepaald wordt door die categorie, hetgeen tot gevolg heeft dat de afwijkingen in de indices snel geringer worden bij een gelijktijdig optrekken van het niveau van de indices. Bij het gebruik van een ongewogen gemiddelde zal een dergelijk verschijnsel zich ook voordoen, maar wellicht in minder sterke mate.

Een moeilijk punt in de analyse is de bepaling van het einde van de analyse, dus de plaats waarop de categorieën niet meer samengevoegd worden. McQuitty stelt voor het aantal kenmerken (antwoorden) te gebruiken waarop de classificatie berust. Hij noemt dit de classificatie-capaciteit. Zij zal zijn de agreement score waarop tot samenvoeging is besloten, vermenigvuldigd met het aantal personen die samengevoegd worden. Samenvoegingen die een verlaging van de classificatie-capaciteit inhouden zullen achterwege blijven. Het is in dit verband mogelijk, dat een samenvoeging op basis van de hoogste agreement score een verlaging van de classificatie-capaciteit betekent, maar dat de daarop volgende wel een hogere classificatie-capa-

citeit oplevert. Zo'n tweede samenvoeging kan dan, gezien de definitie van een type, niet verwezenlijkt worden.

In het door ons gebruikte voorbeeld verloopt de classificatie-capaciteit voor de replacement versie als volgt:

Categorie	*Classificatie-capaciteit*
AB	$2 \times 29 = 58$
CD	$2 \times 26 = 52$
EF	$2 \times 21 = 42$
ABCD	$4 \times 16 = 64$
ABCDEF	$6 \times 6 = 36$

De samenvoeging van de categorieën ABCD en EF doet een daling in de classificatie-capaciteit ontstaan. Deze samentrekking zal dan achterwege worden gelaten, zodat het aantal te onderscheiden typen op twee wordt gesteld.

De classificatie-capaciteit wordt berekend met behulp van geschatte agreement scores. Deze scores kunnen bij grotere aantallen variabelen en onderzoekseenheden vrij sterk overtrokken zijn, zodat de berekening van deze capaciteit weinig zin lijkt te hebben. Als in plaats van agreement scores met andere gelijkheidsindices is gewerkt, is daardoor ook de berekening van de capaciteit zijn zin ontnomen.

3.4. *Kritiek*

Met de veronderstelling van McQuitty, dat een type bepalend is voor het antwoordpatroon op een reeks vragen en dergelijke, kunnen we in zijn algemeenheid vrede hebben. De absolute interpretatie die McQuitty er aan geeft (een patroon van precies dezelfde antwoorden of eigenschappen moet bij elk lid van het type voorkomen) komt ons echter als minder zinvol voor. Het is wellicht beter, zoals in de latent class analysis gebeurt, aan te nemen dat de variabelen in de verschillende clusters van elkaar afwijkende kansverdelingen over de antwoordcategorieën bezitten. Deze onderstelling zal het model echter aanmerkelijk gecompliceerder maken, zodat nadere beperkende assumpties vereist zullen zijn om tot een oplossing te komen.

McQuitty stelt dat bij elk type een uniek antwoordpatroon behoort en elk antwoordpatroon een indicatie is van het onderliggende type. Uit het eerstgenoemde volgt echter niet zonder meer het tweede. De definitie van een type berust bij hem op twee gedachten. Enerzijds wordt een type aangeduid als samenstel van onderzoekseenheden die onderling hoog correleren of sterk met elkaar overeenkomen, anderzijds wordt een type als een oorzakelijkheidsfactor beschouwd, die dwingend bepaalde kenmerken voor de onderzoekseenheden ten gevolge heeft. Deze twee facetten worden door McQuitty zonder meer aan elkaar verbonden, hetgeen zeker niet noodzakelijk hoeft te zijn. Vanuit de oorzakelijkheidsgedachte is het bijvoorbeeld volslagen willekeurig dat de personen die het meest met elkaar gemeen hebben

als beginpunt van de analyse genomen worden, en dus worden beschouwd als zijnde de best mogelijke representanten van het type.

Blijft evenwel dat de replacement versie van de hierarchical syndrome analysis en in mindere mate de elementary linkage analysis, waardevolle methoden lijken tot het formeren van tamelijk homogene submatrices in een matrix met gelijkheidscoëfficiënten. Dat de volgorde van samenvoeging met personen of categorieën bepaald wordt door de hoogste coëfficiënt in de matrix, moet als een goede procedure beschouwd worden. Zo wordt immers geheel afhankelijk van het materiaal de volgorde van de samenvoegingen bepaald. Bij gelijke indices van een persoon met een ander individu en met een categorie lijkt het, ofschoon het vrij arbitrair is, aan te bevelen de samenvoeging met de categorie te nemen, omdat de coëfficiënt met de categorie een groter aantal hogere coëfficiënten vertegenwoordigt.

De gehanteerde classificatie-assumptie betekent in dit verband de laagste coëfficiënt die in de submatrix van het cluster voorkomt. Het is enigszins willekeurig dit als criterium voor mogelijke toevoegingen te gebruiken, maar is naar het ons toeschijnt te verkiezen boven het gemiddelde van de coëfficiënten in een submatrix. De kans is dan minder groot dat relatief slecht gelijkende individuen aan de categorie worden toegevoegd. Op deze wijze worden dan submatrices gevormd waarin de cellen waarden bevatten gelijk aan of groter dan de laagste coëfficiënt waarop de samenvoeging berust.

Het lijkt minder gewenst om de classificatie-capaciteit te nemen als criterium voor het juiste aantal clusters. Het kan wel nuttig zijn deze als uiterste grens te laten fungeren. De zo gevormde submatrices kunnen daarna op hun samenstelling gecontroleerd worden om na te gaan of er misschien toch afwijkende variabelen zijn toegevoegd. De analyseprocedure verschaft immers geen beeld van het totaal van coëfficiënten van een bepaalde groep personen met een andere groep, zodat een eventuele lage coëfficiënt niet als een uitschieter tegenover de leden van een cluster fungeert, maar als de laagste van een groter aantal lage coëfficiënten.

4. Twee toepassingen van de syndroom-analyse van McQuitty

4.1 *Toepassing op basis van verkiezingsuitslagen*

Voor een toepassing van de replacement versie van de syndroom-analyse van McQuitty hebben we als basismateriaal genomen de uitslagen van de Tweede Kamer-verkiezingen in 1963 (C.B.S., 1963) voor een 25-tal gemeenten (zie tabel 4). De gemeenten zijn a-select gekozen, zodanig dat zij over het gehele land verspreid liggen en dat de hoofdonderscheidingen van de C.B.S.-indeling naar urbanisatiegraad (platteland, verstedelijkt platteland en steden) in gelijke mate voorkomen.

Bij de berekening van D, r of produkt-sommen (d.i. $\Sigma X_{je}X_{jf}$) moeten wij er op bedacht zijn, dat de schommelingen in de percentages stemmen op de elf kleinere

partijen tamelijk klein zijn, zodat hun invloed op de berekende coëfficiënt gering is. We moeten er dus rekening mee houden, dat de verschillen in de coëfficiënten hoofdzakelijk door de vijf grote partijen worden veroorzaakt, hetgeen een belangrijke beperking van het materiaal inhoudt. Dat wij het toch als voorbeeld genomen hebben komt, omdat op dit materiaal toepassing van factor-analyse mogelijk is, zodat vergelijkingen tussen de twee technieken mogelijk is.

Tekent men de puntenwolk tussen bijvoorbeeld de gemeenten Budel en Gronsveld, dan valt op dat het punt voor de K.V.P. ver van de overige punten verwijderd is. Dit heeft een groot effect op r, zodat we bij het gebruik van deze coëfficiënt ons van deze beperking bewust dienen te zijn. Ondanks deze tekortkoming blijft r^2 de betekenis houden van percentage verklaarde variantie.

Aan de resultaten van deze toepassing moet geen al te grote betekenis worden toegekend; zij fungeert louter als illustratie. Zou men met een bepaald doel groepen van gemeenten willen vormen, dan zou men o.a. bij de keuze van de gemeenten rekening moeten houden met de bestaande kennis over het te onderzoeken probleem.

Het verloop van de analyse op basis van de matrix met correlatiecoëfficiënten (tabel 5) is in schema 2 weergegeven. Als we het geheel van samenvoegingen bezien, dan lijkt het gewenst tot vier clusters te besluiten, namelijk:

A: 2–3–5–9–12–15–17
B: 4–6–8–10–19
C: 11–13–14–16–21–24
D: 1–18–20–22–25

Nummer 7 is niet aan cluster A toegevoegd, omdat deze de laagst voorkomende coëfficiënt sterk doet dalen.

Omdat vele auteurs hun methode vergelijken met of afzetten tegen factor-analyse, hebben wij de verkiezingsgegevens ook gefactoranalyseerd. Door ons werd de hoofdassenmethode gebruikt (Householder) en daarna werd geroteerd volgens het varimax principe. Het verloop van de eigenwaarden van de factoren voor de factor-analyse op basis van correlatiecoëfficiënten is als volgt:

factor	*eigenwaarde*
1	14,71
2	6,18
3	2,36
4	0,71
5	0,64

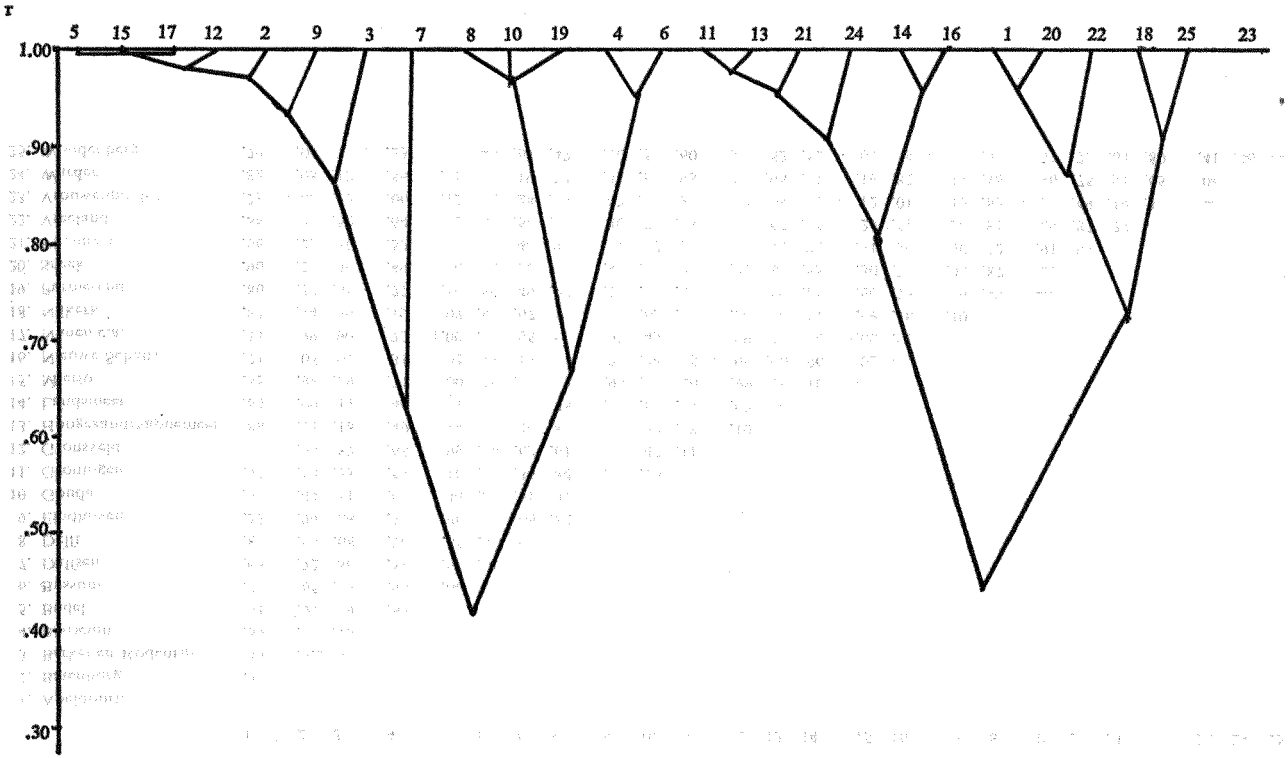
Tabel 4. Uitslagen van de Tweede Kamer-verkiezingen in 1963 in percentages (C.B.S., 1963)

	KVP	*PvdA*	*VVD*	*AR*	*CHU*	*CPN*	*SGP*	*PSP*	*LSP*	*CNVP*	*Grol*	*PEA*	*NDP*	*GPV*	*LU/VV*	*BP*
1. Apeldoorn	14,7	32,8	12,3	10,0	18,3	1,1	2,4	2,4	0,4	—	—	0,7	0,2	0,4	0,2	4,1
2. Batenburg	74,4	4,2	1,0	1,0	10,4	—	0,3	—	—	—	—	0,6	—	—	—	8,1
3. Berkel en Rodenrijs	44,2	5,9	8,2	11,6	17,9	0,1	1,2	0,7	0,3	0,0	—	0,2	0,1	8,6	0,0	1,0
4. Blaricum	31,2	14,7	29,0	3,9	3,7	1,2	0,7	7,0	2,5	—	0,4	0,5	0,0	0,2	0,4	4,6
5. Budel	86,4	7,1	1,1	0,4	0,6	0,3	0,1	0,2	0,0	—	—	0,5	0,1	0,0	0,2	3,0
6. Bussum	27,5	21,1	22,8	8,8	7,5	1,7	1,3	3,6	1,0	—	0,1	1,1	0,3	0,5	0,4	2,3
7. Dalfsen	30,5	11,4	4,3	6,3	33,1	0,1	0,9	0,6	0,1	—	—	0,1	0,0	5,7	0,0	6,9
8. Delft	28,7	32,4	10,4	9,2	7,4	3,0	2,7	3,5	0,4	0,1	—	0,5	0,2	0,4	0,0	1,1
9. Eindhoven	60,5	20,7	7,2	3,4	2,4	0,8	0,2	1,8	0,3	—	—	0,8	0,3	0,2	0,3	1,1
10. Gouda	21,5	37,8	11,3	9,8	7,4	2,0	3,9	3,5	0,3	0,2	—	0,4	0,2	0,7	0,2	0,8
11. Groningen	9,4	38,6	14,8	10,9	8,5	5,0	0,5	5,4	0,4	—	—	0,6	0,3	4,5	0,1	1,0
12. Gronsveld	83,4	3,6	1,7	—	0,3	0,2	—	0,2	—	—	—	0,2	—	—	—	10,4
13. Hoogezand/Sappemeer	9,3	47,0	13,6	6,8	5,1	7,7	0,3	6,3	0,3	—	—	0,4	0,3	1,1	0,1	1,7
14. Landsmeer	7,9	38,7	9,2	9,6	7,6	15,8	0,9	6,2	0,2	—	—	0,2	0,2	0,2	0,4	2,9
15. Mierlo	80,8	10,7	2,1	0,6	0,6	0,7	0,0	0,5	0,4	—	—	0,7	0,1	0,1	0,2	2,5
16, Nieuwe Schans	2,8	41,7	12,1	4,0	6,2	25,3	1,1	4,9	0,2	—	—	0,9	—	0,2	0,1	0,5
17. Nunen c.a.	78,9	8,2	4,6	1,2	2,7	0,1	—	0,9	0,1	—	—	0,4	0,3	0,1	0,4	2,1
18. Nijkerk	7,2	25,4	3,4	20,3	29,3	0,4	8,6	0,4	0,1	—	—	0,2	0,0	0,8	0,0	3,9
19. Purmerend	25,7	35,6	14,3	4,9	4,1	6,8	0,2	4,8	0,3	—	—	0,7	0,3	0,2	0,4	1,7
20. Sneek	13,6	33,1	8,7	19,4	16,4	2,7	0,6	3,3	0,2	—	—	0,3	0,1	0,4	0,1	1,1
21. Vlissingen	14,5	44,6	11,1	9,5	11,5	1,7	2,0	2,2	0,5	0,1	—	0,7	0,3	0,3	0,2	0,8
22. Vreeland	11,4	20,2	19,6	17,2	16,8	0,5	6,4	2,7	1,2	—	—	0,9	0,3	—	0,9	1,9
23. Vrouwenpolder	0,6	5,9	4,1	50,8	18,9	0,2	6,9	0,7	0,4	—	—	—	0,2	10,4	—	0,9
24. Warder	8,0	48,9	22,0	1,9	1,0	1,6	1,0	7,0	—	—	—	0,3	—	—	0,3	8,0
25. Woudenberg	3,0	22,9	11,8	9,6	34,2	0,4	12,6	1,2	0,3	0,1	—	0,3	0,2	0,1	0,1	3,2

Tabel 5: Correlatiematrix tussen 25 gemeenten op basis van verkiezingsgegevens

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1. Apeldoorn	—																								
2. Batenburg	.33	—																							
3. Berkel en Rodenrijs	.49	.92	—																						
4. Blaricum	.59	.67	.70	—																					
5. Budel	.31	.99	.89	.69	—																				
6. Bussum	.79	.65	.73	.95	.66	—																			
7. Dalfsen	.65	.72	.86	.51	.63	.61	—																		
8. Delft	.88	.63	.68	.76	.65	.90	.62	—																	
9. Eindhoven	.53	.94	.88	.79	.96	.81	.66	.82	—																
10. Gouda	.93	.44	.51	.67	.46	.84	.51	.97	.68	—															
11. Groningen	.92	.13	.28	.54	.16	.73	.36	.84	.41	.93	—														
12. Gronsveld	.27	.99	.87	.68	.99	.64	.62	.61	.95	.42	.11	—													
13. Hoogezand/Sappemeer	.88	.11	.18	.49	.14	.67	.26	.83	.40	.92	.98	.10	—												
14. Landsmeer	.84	.09	.17	.40	.11	.60	.28	.78	.35	.87	.94	.07	.96	—											
15. Mierlo	.35	.98	.89	.71	1.00	.69	.63	.69	.98	.50	.20	.99	.19	.16	—										
16. Nieuwe Schans	.71	—.05	.01	.31	—.02	.48	.13	.64	.21	.74	.85	—.06	.90	.96	.03	—									
17. Nunen c.a.	.34	.99	.90	.72	1.00	.70	.65	.67	.97	.49	.19	.99	.17	.14	1.00	.01	—								
18. Nijkerk	.83	.14	.39	.19	.07	.45	.67	.59	.22	.65	.67	.04	.58	.61	.09	.46	.10	—							
19. Purmerend	.86	.53	.55	.77	.56	.89	.49	.97	.76	.97	.89	.52	.89	.85	.60	.75	.59	.49	—						
20. Sneek	.96	.27	.46	.49	.26	.73	.58	.86	.48	.91	.91	.22	.85	.85	.30	.70	.30	.87	.81	—					
21. Vlissingen	.96	.25	.36	.53	.27	.74	.46	.90	.51	.97	.97	.22	.96	.92	.31	.80	.30	.72	.91	.93	—				
22. Vreeland	.88	.25	.50	.64	.22	.80	.58	.73	.40	.77	.78	.19	.67	.65	.25	.51	.26	.81	.69	.87	.77	—			
23. Vrouwenpolder	.28	—.08	.22	—.06	—.12	.12	.24	.13	—.07	.17	.23	—.13	.08	.15	—.12	.01	—.10	.63	—.00	.49	.19	.54	—		
24. Warder	.82	.08	.12	.56	.11	.69	.19	.77	.36	.87	.93	.09	.96	.87	.16	.82	.14	.46	.86	.75	.91	.65	—.04	—	
25. Woudenberg	.79	.04	.28	.23	—.05	.43	.64	.47	.10	.55	.60	—.07	.52	.53	—.03	.44	—.01	.91	.41	.72	.64	.80	.41	.46	—

Schema 2: Opbouw van de clusters volgens de syndroom-analyse



Besloten werd drie factoren (die tezamen 93% van de variantie voor hun rekening nemen) te roteren met als resultaat:

Tabel 6: Geroteerde factorladingen (|ladingen| < 0,25 zijn weggelaten)

Gemeente	*I*	*II*	*III*
1. Apeldoorn	—.80	.28	—.51
2. Batenburg		.99	
3. Berkel en Rodenrijs		.90	—.39
4. Blaricum	—.47	.72	
5. Budel		.99	
6. Bussum	—.62	.66	
7. Dalfsen		.65	—.63
8. Delft	—.75	.61	
9. Eindhoven	—.31	.94	
10. Gouda	—.85	.41	—.26
11. Groningen	—.94		—.28
12. Gronsveld		.99	
13. Hoogezand/Sappemeer	—.98		
14. Landsmeer	—.94		
15. Mierlo		.99	
16. Nieuwe Schans	—.92		
17. Nunen		.99	
18. Nijkerk	—.45		—.86
19. Purmerend	—.85	.51	
20. Sneek	—.76		—.58
21. Vlissingen	—.91		—.32
22. Vreeland	—.60		—.67
23. Vrouwenpolder			—.81
24. Warder	—.97		
25. Woudenberg	—.44		—.80

De eerste factor blijkt overwegend de gemeenten te vertegenwoordigen die niet confessioneel (PvdA, VVD, CPN, PSP) gestemd hebben, terwijl de tweede factor de KVP stemmende en de derde factor de protestants-christelijk stemmende gemeenten representeert.

Ook op produkt-sommen van de ruwe scores werd factor-analyse toegepast. Eerst werden alle produkt-sommen door 10.000 gedeeld om de matrix geschikt te maken voor computer-verwerking. Voor de uitkomsten heeft deze handelwijze geen gevolgen. Het verloop van de eigenwaarden is:

factor	*eigenwaarde*
1	5.31
2	1.97
3	0.48
4	0.13
5	0.11

Ook hier werden de eerste drie factoren geroteerd. De resultaten zijn:

Tabel 7: Geroteerde factorladingen (|ladingen| < 0,25 zijn weggelaten).

Gemeente	*I*	*II*	*III*
1. Apeldoorn		.35	
2. Batenburg	.74		
3. Berkel en Rodenrijs	.44		
4. Blaricum	.31	.26	
5. Budel	.86		
6. Bussum	.27	.29	
7. Dalfsen	.30		.29
8. Delft	.27	.35	
9. Eindhoven	.59		
10. Gouda		.40	
11. Groningen		.42	
12. Gronsveld	.84		
13. Hoogezand/Sappemeer		.50	
14. Landsmeer		.43	
15. Mierlo	.80		
16. Nieuwe Schans		.48	
17. Nunen	.78		
18. Nijkerk		.25	.36
19. Purmerend		.40	
20. Sneek		.35	.25
21. Vlissingen		.46	
22. Vreeland		.26	.26
23. Vrouwenpolder			.50
24. Warder		.54	
25. Woudenberg		.25	.33

De overeenkomst tussen de beide factor-analyses is groot. Het patroon van factorladingen blijkt bij beide hetzelfde te zijn, hoewel de eerste en tweede factor gewisseld hebben wat belangrijkheid betreft.

In de clusters die met behulp van de syndroom-analyse zijn onderscheiden blijkt hoofdzakelijk gestemd te zijn:

A: KVP

B: KVP + niet-confessioneel (PvdA, VVD, PSP, CPN)

C: Niet-confessioneel

D: Protestant + niet-confessioneel.

Vergelijkt men deze resultaten met de resultaten van de eerste factor-analyse, dan blijkt, dat de gemeenten van cluster A hoge ladingen hebben op factor II (⩾ .90), die van B op I en II (⩾ .41), die van C op I en III (⩾ .44), die van D op I (⩾ .91). Het verschil tussen de methoden, het zoeken naar de meest verklarende hypothetische dimensies en het zoeken naar overeenkomst op een zo groot mogelijk aantal punten, komt op dit materiaal dus ook in de resultaten tot uiting.

4.2 *Toepassing op basis van gegevens over werkvoorkeuren*

Voor een andere toepassing van de replacement versie van de syndroom-analyse werd gebruik gemaakt van de gegevens over werkvoorkeuren van 310 bouwvakkers, die verkregen waren in een onderzoek van het Instituut voor Arbeidsvraagstukken (zie: 'Werkoriëntaties en marktgedrag van bouwvakarbeiders, 1963). Deze gegevens zijn ook door Stouthard (1965, blz. 18–27) gebruikt voor een toepassing van de latent class analysis. In het betreffende onderzoek werden de respondenten gevraagd voor een aantal paren werkaspecten aan te geven, aan welk aspect men de voorkeur gaf.

Om voor de analyse volgens de replacement versie tot een hanteerbaar aantal personen te komen werden er 44 bouwvakkers gekozen, waarbij er 24 a-select uit de latente klasse I en 20 uit de latente klasse II werden genomen. De verhouding 24 : 20 is gebaseerd op de grootte van de klassen in het onderzoek, te weten 174 : 136 (Stouthard, 1965, blz. 22). Tussen deze 44 respondenten werden de agreement scores berekend, welke scores in een matrix werden geplaatst. Op deze matrix werd de replacement versie toegepast, waarbij gestreefd werd naar maximalisatie van de classificatie-capaciteit. Het resultaat was, dat alle onderzoekspersonen in één cluster werden samengevoegd.

Laat ons de agreement score-matrix bezien als deze zodanig hrrordend is dat alle personen bij elkaar geplaatst zijn die in hetzelfde cluster zijn geplaatst, op het punt van de analyse, dat alle personen in een cluster zijn samengevoegd uitgezonderd respondent nr. 34 (zie tabel 8). De clusters bestaan uit de respondenten:

Cluster	*Respondenten*
A	8–11–16–17–21–24–40–44
B	3–10–12–39
C	5–14–18–20–31–33–35–42
D	7–22–23–25–29–37
E	9–19–30–36–41–43
F	1–4–6–26–28–32–38
G	2–13–15–27

De clusters A en F bestaan geheel uit respondenten uit de latente klasse II, de clusters B, C en D geheel uit de latente klasse I en de clusters E en G bevatten zowel respondenten uit klasse I als uit klasse II. Uit de agreement scores van de personen in de clusters en tussen de clusters blijkt, dat we tot het bestaan van één cluster moeten besluiten. Dat op basis van de gebruikte variabelen de twee latente klassen niet onderscheiden kunnen worden met behulp van de syndroom-analyse is eigenlijk wel voor de hand liggend, als men de latente waarschijnlijkheden beziet voor de antwoord-alternatieven voor de beide latente klassen (Stouthard, 1965, blz. 22):

Tabel 9

Latente waarschijnlijkheden voor de antwoord-alternatieven	*latente klasse I*	*latente klasse II*
1. Werken op een bouw die lang duurt	.87	.40
2. Werken bij een grote aannemer	.87	.27
3. Werken op een bouw waar veel mensen werken	.43	.02
4. In een ploeg werken	.30	.07
5. Steeds dezelfde soort werkzaamheden	.38	.09
6. Werken voor een vast uurloon	.53	.83
7. Montage-bouw	.32	.10
8. Groot bouwwerk	.98	.22
9. Alléén werken	.18	.29
10. Zelfstandig werken	.80	.91
11. Veel toezicht door de uitvoerder	.32	.32
12. Veel contact met de uitvoerder	.48	.50
13. Veel contact met de aannemer of bedrijfsleider	.36	.42
14. Veel toezicht van de opzichter	.52	.46
15. Steeds met dezelfde mensen samenwerken	.90	.96
16. Blijvend bij één aannemer werken	.91	.88
17. Werk waarvan je nog wat kunt leren	.71	.79
18. In de eigen woonplaats werken	.83	.81
19. Met ouderen samenwerken	.72	.67

Alleen voor de variabelen 1, 2, 3 en 8 zijn de verschillen in de latente waarschijnlijkheden zodanig, dat men een redelijk verschil kan verwachten tussen de antwoorden in de beide klassen. De verschillen die deze vier variabelen ten gevolge hebben, zullen echter overtroefd worden door de overeenkomst op de variabelen die voor het onderscheid in latente klasse I en II minder relevant zijn. In deze toepassing blijkt het verschil van aanpak en resultaat van de latent class analysis en de syndroom-analyse. De latent class analysis zoekt naar verschillen tussen hypothetische klassen en wel zodanig, dat in de analyse getracht wordt de variabelen waarop de duidelijkste verschillen voorkomen de belangrijkste rol in de analyse te laten spelen (Gibson, 1955, blz. 72). De verschillen worden dus zo sterk mogelijk benadrukt. De syndroom-analyse zoekt daarentegen naar maximale overeenkomst en zal nu juist de verschillen zoveel mogelijk trachten weg te werken. Hetgeen, zoals uit onze toepassing blijkt, geheel andere resultaten ten gevolge kan hebben.

Het gesignaleerde verschil tussen de twee methoden is overigens niet het enige. Latent class analysis gaat uit van de veronderstelling van lokale independentie, terwijl McQuitty van het tegendeel uitgaat. Bovendien is het McQuitty-model deterministisch, in tegenstelling tot latent class analysis dat probabilistisch is.

Tabel 8: Agreement scores op basis van latent class analysis gegevens (Stouthard, 1965)

	8	11	16	17	21	24	40	44	3	10	12	29	5	14	18	20	31	33	35	42	7	22	23	25	29	37	9	19	30	36	41	43	1	4	6	26	28	32	38	2	13	15	27
8																																											
11	13																																										
16	12	16																																									
17	15	15	14																																								
21	13	17	18	15																																							
24	12	12	11	14	12																																						
40	14	16	17	14	16	11																																					
44	13	17	18	15	17	12	18																																				
3	10	10	9	10	10	11	11	10																																			
10	14	12	11	12	12	11	13	12	15																																		
12	9	11	10	13	11	12	10	11	16	14																																	
39	12	12	9	10	10	9	9	10	13	15	14																																
5	13	13	12	13	13	14	12	13	10	12	11	12																															
14	10	10	9	12	10	9	9	10	11	11	14	13	14																														
18	8	10	9	10	8	9	9	10	7	9	10	11	12	11																													
20	13	11	12	11	11	10	14	13	14	14	13	14	13	12	12																												
31	11	11	12	11	11	10	12	13	12	12	13	14	13	14	14	17																											
33	11	11	12	13	13	12	10	11	8	10	11	10	17	14	12	11	11																										
35	9	9	10	11	9	10	10	11	12	14	15	14	11	14	10	13	13	11																									
42	12	10	9	10	10	11	11	10	11	13	10	11	14	11	15	14	14	12	8																								
7	14	14	13	14	14	13	15	14	15	15	14	13	12	11	9	16	14	10	12	13																							
22	11	11	12	11	13	12	12	11	14	14	13	12	13	10	10	15	13	13	11	14	16																						
23	14	8	7	10	8	13	9	8	11	13	10	13	14	11	9	14	12	12	12	13	13	14																					
25	15	11	8	11	9	10	10	9	10	12	9	14	13	10	10	13	11	11	9	12	12	13	16																				
29	12	10	11	12	12	15	11	10	11	13	10	9	14	9	9	12	10	14	10	13	13	16	15	12																			
37	12	12	11	10	12	9	11	10	13	13	12	15	12	11	11	16	14	12	10	13	15	16	13	14	13																		
9	11	13	14	15	15	10	12	13	10	12	13	10	13	12	12	11	11	15	11	12	12	13	8	9	12	12																	
19	9	13	16	13	15	8	14	15	10	10	11	8	11	10	10	11	11	13	11	8	10	11	6	7	10	10	15																
30	8	12	13	12	14	9	11	12	11	13	14	11	10	11	11	10	10	12	12	11	11	12	7	6	11	11	16	14															
36	10	14	15	14	16	11	13	14	11	13	14	11	12	11	11	12	12	12	12	11	13	14	9	8	13	13	16	16	17														
41	12	12	13	14	14	11	13	12	13	15	14	11	12	11	11	14	12	14	12	13	15	16	11	10	15	15	16	14	15	17													
43	10	12	11	14	12	11	11	12	13	13	14	10	14	13	11	10	10	14	12	11	11	12	9	10	11	9	14	14	13	13	13												
1	11	15	14	13	15	14	14	15	12	14	13	12	15	10	12	13	13	13	11	14	14	15	12	11	14	12	13	13	14	16	14	14											
4	13	13	14	13	13	12	16	15	12	14	11	10	11	8	10	15	13	9	11	12	14	13	12	11	14	12	11	13	12	14	14	10	15										
6	12	12	13	12	14	13	13	12	11	11	8	7	10	7	5	10	8	10	8	9	13	12	11	8	15	11	10	12	11	13	13	9	12	14									
26	12	12	11	12	12	11	13	12	13	15	12	11	10	11	9	12	12	8	10	13	13	12	11	10	13	11	10	10	13	13	13	11	14	16	13								
28	12	14	13	12	14	13	13	14	11	15	12	13	14	9	11	12	12	12	12	13	13	14	13	12	13	11	12	12	13	15	13	13	18	14	11	13							
32	12	12	11	12	12	13	13	12	13	15	12	11	12	9	9	12	10	10	10	13	13	14	13	12	15	11	10	10	13	13	13	13	16	16	13	17	15						
38	11	13	14	13	13	12	14	15	12	12	11	10	9	8	8	11	11	7	11	8	12	9	8	7	10	8	9	13	12	12	10	12	13	15	14	14	12	14					
2	7	9	6	11	7	10	6	7	10	8	11	10	9	10	14	9	11	9	9	12	10	11	10	11	10	10	11	9	10	10	10	12	11	9	8	10	10	10	9				
13	8	10	11	12	12	11	9	10	7	7	8	5	12	11	9	6	8	14	8	9	7	10	9	8	13	7	12	14	11	13	11	13	12	10	13	11	11	11	10	12			
15	8	10	9	12	10	13	9	10	11	11	12	9	14	13	13	10	12	14	12	13	11	12	11	8	13	9	12	12	11	13	13	15	14	10	11	11	13	11	10	14	15		
27	10	8	7	10	8	13	9	8	13	11	10	9	12	11	9	12	12	10	10	13	13	14	15	12	15	11	8	8	7	9	11	11	12	12	13	13	11	13	10	14	13	15	
34	11	13	12	11	11	8	14	13	10	10	9	10	9	8	10	13	11	7	9	10	12	11	10	13	10	12	11	11	10	10	10	8	11	15	10	12	10	12	11	11	8	6	10

Summary

The purpose of cluster analysis is to group objects or variables into a number of classes, in such a way that every member of a class as far as possible resembles the other members of that class and is different from non-members. The techniques of cluster analysis are mainly used to group persons, institutions, unions, etc. on the basis of all the available relevant information. As such, cluster analysis can be seen as a method for detecting types and in this way it can make a contribution to concept formation. McQuitty has suggested a large number of techniques. Two of them are discussed. We do not accept the underlying data model, but from another point of view it seems a useful method, particularly the replacement version of the hierarchical syndrome analysis.

Literatuur

Boon van Ostade, A.H.:
Over type-onderzoek; Nederlands tijdschrift voor de psychologie en haar grensgebieden, 1971, 26, 82–101.

Bridges, C.C.:
Hierarchical cluster analysis; Psychological reports, 1966, 18, 851–854.

Bijnen, E.J.:
Cluster-analyse, overzicht en evaluatie van technieken, 1969, 150 blz.

Capecchi, V.:
Linear causal models and typologies; Quality and quantity, 1967, 1, 116–152.

Cattell, R.B.:
A note on correlation clusters and cluster search; Psychometrika, 1944, 9, 169–184.

C.B.S.:
Statistiek der verkiezingen 1963, Tweede Kamer der Staten Generaal, 15 mei 1963, 100 blz.

Cronbach, L.J. and Gleeser, G.C.:
Assessing similiarity between profiles; Psychological bulletin, 1953, 50, 456–473.

Gibson, W.A.:
An extension of Anderson's solution for the latent structure equations; Psychometrika, 1955, 20, 69–73.

Helmstadter, G.C.:
An empirical comparison of methods for estimating profile similarity; Educational and psychological measurement, 1957, 17, 71–83.

Hempel, C.G. und Oppenheim, P.:
Der Typusbegriff im Lichte der neuen Logik, 1936, 130 blz.

Hilgendorf, L., Clark, A.W. and Irving, B.L.:
The combined use of linkage and path analysis in the development of causal models; Human relations, 1967, 20, 375–386.

Johnson, S.C.:
Hierarchical clustering schemes; Psychometrika, 1967, 32, 241–254.

Meehl, P.E.:
Configural scoring; Journal of consulting psychology, 1950, 14, 165–171.
Mosel, J.N. and Roberts, J.B.:
The comparability of measures of profile similarity: an empirical study; Journal of consulting psychology, 1954, 18, 61–66.
Muldoon, J.F. and Ray, O.S.:
A comparison of pattern similarity as measured by six statistical techniques and eleven clinicians; Educational and psychological measurement, 1958, 18, 775–781.
McQuitty, L.L.:
Agreement analysis: classifying persons by predominant patterns of response; British journal of statistical psychology, 1956, 9, 5–16.
McQuitty, L.L.:
Elementary linkage analysis for isolating orthogonal and oblique types and typal relevancies; Educational and psychological measurement, 1957, 17, 207–229.
McQuitty, L.L.:
Hierarchical linkage analysis for the isolation of types; Educational and psychological measurement, 1960a, 20, 55–67.
McQuitty, L.L.:
Hierarchical syndrome analysis; Educational and psychological measurement, 1960b, 20, 293–304.
McQuitty, L.L.:
Multiple hierarchical classification of institutions and persons with reference to union-management relations and psychological well-being; Educational and psychological measurement, 1962, 22, 513–531.
McQuitty, L.L.:
Capabilities and improvements of linkage analysis as a clustering method; Educational and psychological measurement, 1964, 24, 441–457.
McQuitty, L.L.:
Multiple rank order typal analysis for the isolation of independent types; Educational and psychological measurement, 1966a, 26, 3–11.
McQuitty, L.L.:
Similarity analysis by reciprocal pairs for discrete and continuous data; Educational and psychological measurement, 1966b, 26, 825–831.
McQuitty, L.L.:
A mutual development of some typological theories and pattern-analytic methods; Educational and psychological measurement, 1967, 27, 21–46.

Stephenson, W.:
The study of behavior, 1953, 376 blz.
Stouthard, Ph.C.:
Data-modellen, 1965, 99 blz.

Ward, J.H.:
Hierarchical grouping to optimize an objective function; Journal of the american statistical association, 1963, 58, 236–244.
Ward, J.H. and Hook, M.E.:
Application of an hierarchical grouping procedure to a problem of grouping profiles; Educational and psychological measurement, 1963, 23, 69–81.
Werkoriëntaties en marktgedrag van bouwvakarbeiders; Rapport van het Instituut voor Arbeidsvraagstukken, 1963, 95 blz.

Vingeroefeningen in klassificatietechniek *
Een toepassing van taxonometrische klassificatie-methoden op enquêteresultaten

G. J. Randoe en T. H. M. van Kooten

1.0.0 Probleemstelling

Multivariate analysemethoden zijn heden ten dage onmisbare hulpmiddelen bij systematisch onderzoek van datasets van het type dat uit enquêtemateriaal wordt ontwikkeld.

Gegeven een specifiek probleemgebied dat een gegeven marktonderzoek bestrijkt, wordt een dataset ontwikkeld waarin een object (de ondervraagde, de eenheid van onderzoek, een bedrijf etc.) beschreven wordt door een aantal relevante variabelen die al dan niet na een vooronderzoek worden geoperationaliseerd in een serie operatoren zoals b.v. situatiekenmerken, gedragskenmerken en instellings- en/of dispositiekenmerken.

Men gaat er vervolgens van uit dat deze object-definitie in het kader van de probleemstelling aanspraak kan maken op volledigheid, en dat alle elementen van de definitie (de afzonderlijke operatoren) ieder voor zich terecht pretenderen een geldige maatstaf te zijn voor de variabelen die zij vertegenwoordigen.

Bij de keuze van het type multivariate analyse dat men op de dataset zal toepassen moeten nieuwe werkhypothesen worden opgesteld. Meestal vloeien deze rechtstreeks voort uit de marketing-probleemstelling[1].

De eerste belangrijke keuze is die van de prioriteit tussen een analyse van variabelen over objecten (R analyse) of een analyse van objecten over variabelen (Q analyse). Over deze laatste categorie gaat dit verslag.

De meest voorkomende typen van Q analyses in marktonderzoek zijn segmentatie-analyses. De uitgangshypothese is hier meestal, dat een criteriumvariabele (pivot) afhankelijk is van n – 1 'onafhankelijke' variabelen. Hebben wij te maken met operatoren van het ratio-interval- of rangorde-type, dan kan men gebruik maken van multiple discriminant-analyse. Zijn veel operatoren dichotoom, dan kan een analyse van opeenvolgende contrasterende groepen (Stouthard 1965) uitkomst brengen. Een variant op deze werkhypothese is het geval dat een criteriumset wordt

gedefinieerd (b.v. twee pivots) die uit meer dan een operator bestaat. In dat geval is de ook in marktonderzoek meest toegepaste analysemethode de canonische analyse.

Moeilijker wordt in het marktonderzoek de analyse wanneer de afhankelijkheidsrelatie niet langer uitgangspunt van de analyse is, doch juist tot onderwerp van de analyse wordt gemaakt. In het betrokken geval is dan geen segmentatieprobleem aan de orde, waarin a priori het criterium vastligt (de niet gebruiker van merk X t.o.v. de wel gebruiker van merk X), doch een probleem van latente structuuranalyse; n.l. overeenkomst van scoregedrag tussen objecten op een gegeven set operatoren.

Deze soort problemen kunnen zich voordoen bij de verwerking van beschrijvend marktonderzoek. Het clustercriterium wordt nu niet gedicteerd door de marketingprobleemstelling, maar moet het resultaat zijn van een intern consistentie-onderzoek. Het beschikbare instrumentarium voor multivariate analyse is nu aanzienlijk beperkter dan in het eerste geval. Wanneer de aard van de operatoren het toelaat kan men clusteranalyse toepassen op basis van distantiemaatstaven. De twee meest toegepaste maatstaven zijn:

$$d_{ij} = \sum_{k=1}^{n} \left| a_{ik} - a_{jk} \right| \qquad \text{(A)}$$

$$d_{ij} = \sqrt{\left[\sum_{k=1}^{n} (a_{ik} - a_{jk})^2 \right]} \qquad \text{(B)}$$

waarbij d_{ij} = afstand tussen object a_i en a_j; a_{ik} = i^e waarneming gedefinieerd in k operatoren: a_{jk} = idem j^e element.

De maatstaf (B) krijgt vaak de voorkeur omdat hij 'normalisatie' van d_{ij} mogelijk maakt en dus uitspraken toelaat over de statistische betrouwbaarheid (stabiliteit) van de verkregen structuur. Een groot nadeel is dat theoretisch gesproken de eis die men aan alle betrokken operatoren moet stellen luidt, dat het tenminste intervalschalen zijn. Onder bepaalde omstandigheden zal het niet veel uitmaken of men rangordeschalen of intervalschalen heeft gebruikt, doch met name nominale variabelen kunnen strikt gesproken niet aan dit type clusteranalyse worden onderworpen.

Clustermethoden, die in deze complicatie voorzien, zijn de iteratieve methode, ontworpen door Boon van Ostade (1969) en een eerder door Bonner (1964) ontwikkelde methode voor elementensets bestaande uit nominale kenmerken (attributen). Boon van Ostade kiest evenals Stouthard de methode van dichotomeren van alle operatoren, voorzover zij al niet als 'binaire' operatoren in het materiaal voorkomen. Bonner stelt voor om in een geval waarin zowel dichotome als multichotome operatoren in de definitie voorkomen elke multichotome operator in dichotome operatoren op te splitsen. Aan beide opvattingen kleven voordelen en be-

zwaren. Het voorstel van Bonner komt neer op een opvatting van een schaal als n onderling onafhankelijk tot stand gekomen scores. Er zijn nu veel extra maatregelen nodig om te voorkomen dat de met zo'n operator verkregen overeenkomstindices niet veel te hoog worden geschat. In deze maatregel is in het door Bonner ontworpen programma slechts sporadisch voorzien. Een praktisch probleem is bovendien dat het door Bonner geschreven programma niet meer op een moderne IBM 360 serie computer kan worden gedraaid.

Het voorstel van Boon van Ostade kan in bepaalde gevallen tot een hinderlijk informatieverlies leiden, n.l. daar waar de frequentieverdeling van de scores op de schaal niet eentoppig is. Bepaalde typen operatoren worden bij een dergelijke ingreep zelfs onherkenbaar verminkt. Dit zal b.v. altijd het geval zijn bij nominale schalen met meer dan 2 niveaus.

Dichotomeren heeft dan onvermijdelijk tot gevolg, dat de aanvankelijke indicatie dat twee heterogene groepen gezamenlijk een dergelijke tweetoppige frequentieverdeling van scores veroorzaken geheel verloren raakt.

Representeert de schaal de kleur van het object, dan is elke dichotomering van de operator in feite een handeling, waarbij kleuren willekeurig aan elkaar worden gekoppeld voordat indicaties over de juistheid van zo'n koppeling voorhanden zijn. Een veel gehoorde tegenwerping is, dat het in de praktijk erg 'meevalt'. In veel gevallen, waarin een techniek is toegepast, zonder dat aan de voor die techniek geldende voorwaarden is voldaan, komt naar het lijkt het toeval de onderzoeker te hulp. In de gevonden 'latente structuur' worden dan als belangrijke criteria die variabelen aangewezen, waarvan de operatoren wel aan de geldende voorwaarden voldeden.

Maar al te vaak ziet men dan, dat in de interpretatie van de resultaten voorbij gegaan wordt aan de mogelijkheid, dat het konijn, dat uit de hoge hoed komt hetzelfde is, dat men er van tevoren in had gestopt.

Dit verslag behandelt enkele resultaten van een onderzoek naar de bruikbaarheid van een voor taxonometrische doeleinden ontwikkelde clusteranalyse, die een oplossing aan de hand doet voor een weging van de meetkarakteristieken in die gevallen waarin objecten (in ons geval de ondervraagden in een enquête) door combinaties van verschillende soorten operatoren zijn beschreven.

Van de voor dit doel gekozen analyse was een programma beschikbaar in de bibliotheek van de IBM, t.w. het door J. Rubin en H. Friedman vervaardigde programma: 'A Cluster Analysis and Taxonomy System for Grouping and Classifying Data'.

2.0.0 Clusteranalyse op basis van overeenkomst-indices

De hier besproken clusteranalyse is ontwikkeld door Jerrold Rubin en Herman Friedman. De auteurs noemen het hier besproken analyseprincipe 'geometrisch' in tegenstelling tot het gangbare 'statistische' principe.

Bij de meer bekende vormen van clusteranalyse (opgezet volgens het statistische principe) wordt de dataset nadrukkelijk behandeld als een steekproef uit een populatie. Dit betekent dat er veel aandacht wordt besteed aan de geldigheid van de gevonden groepsstructuur als schatter van de verwachte groepsstructuur in de populatie, die de dataset geacht wordt te vertegenwoordigen.

Bij clusteranalyses, die zijn opgezet volgens het geometrisch principe vervalt dit voorbehoud (in wezen wordt de dataset opgevat als een op zichzelf staande entiteit).

Daarvoor in de plaats treedt een veel grotere vrijheid in de keuze van distantiemaatstaven. Dit heeft weer tot gevolg dat vrijwel alle beperkingen, die in statistische analyses van kracht zijn voor toelating van bepaalde typen variabelen in de definitie van een object, bij de geometrische analyse vervallen.

Het doel van de hier aan de orde zijnde clusteranalyse is de organisatie van een verzameling objecten in een beperkt aantal elkaar niet overlappende groepen waarvan de leden onderling grote gelijkenis vertonen. Tevens wordt gestreefd naar een 'optimale' verdeling, waarmee wordt bedoeld optimale homogeniteit binnen de groep en optimale heterogeniteit tussen de groepen onderling.

Binnen elke groep is ernaar gestreefd binnen de door de gebruiker van de analyse gestelde normen de onderlinge gelijkenis van de leden zo groot mogelijk te maken, terwijl daarnaast de 'afstanden' tussen de groepen onderling zo groot mogelijk is gemaakt.

Het criterium voor de vergelijking van waarnemingen (objecten) onderling is de overeenkomstcoëfficiënt (hierna aangeduid als 'overeenkomst-index').

De basis van de analyse wordt gevormd door een matrix, die paarsgewijs de relaties tussen alle objecten in de dataset beschrijft.

Het programma biedt een keuze uit verschillende rekenregels voor het in een getal weergeven van de gelijkenis tussen twee objecten. Alle hebben als principe gemeen, dat zij tot uitdrukking brengen de mate waarin twee waarnemingen (objecten) gedefinieerd in scores op k operatoren (codevoorschriften voor operationalisaties van variabelen) qua 'scorepatroon' met elkaar overeenkomen.

De matrix van overeenkomstindices wordt nu onderworpen aan een toetsingsprocedure. Alle overeenkomst-indices worden daarbij getoetst aan een subjectief gekozen norm voor gelijkheid, de 'brekingscoëfficiënt'. Deze coëfficiënt is een drempelwaarde. In eerste instantie worden alleen die objecten als 'op elkaar gelijkend' beschouwd, waarvan de overeenkomstindex een waarde heeft hoger dan de voor de brekingscoëfficiënt gekozen waarde.

Het computerprogramma bevat voorts hill climbing routines voor het itereren van een als optimaal beschouwde verdeling.

De voor de analyse toegelaten objectdefinities kunnen qua karakter zeer uiteenlopende operatoren (schaaltypen) omvatten.

Dit is o.a. mogelijk gemaakt door de hantering van een relatief criterium voor gelijkheid van het scorepatroon van twee objecten.

Wanneer het gehanteerde schaaltype vanuit theoretisch oogpunt gezien daartoe aanleiding geeft, worden niet alleen volstrekt gelijke scores als criterium voor gelijkheid geaccepteerd, doch kunnen tevens dicht bij elkaar liggende scores als gelijk of als gedeeltelijk gelijk worden geaccepteerd (fractional match).

In het hier gerapporteerde onderzoek werd de door de auteurs Rubin en Friedman zelf ontwikkelde maatstaf voor paarsgewijze overeenkomst gebruikt. Deze maatstaf is als volgt gedefinieerd:

$$S_{ij} = \frac{K_m}{K} \qquad (1)^2$$

Waarbij:
S_{ij} = de mate van overeenkomst tussen object i en object j.
K_m = het aantal operatoren waarop beide objecten eenzelfde waarde scoren.
K = het totaal aantal operatoren waarin de objecten i en j zijn gedefinieerd.

Voor de toepassing van de analyse is essentieel, dat alle objecten in K dezelfde operatoren zijn beschreven.

Niet essentieel is, dat op alle operatoren waarin het object is gedefinieerd voor elk object een score voorhanden is.

Er is in het programma voorzien in oplossingen voor datasets, waarin gegevens ontbreken.

Ook is niet essentieel, dat voor alle variabelen waarin het object is beschreven één uniform codevoorschrift voor de vertaling in operatoren is gehanteerd. Het is dus niet noodzakelijk om b.v. alle variabelen te dichotomeren, of als rangordeschaal in te brengen of iets dergelijks. Het hier besproken geometrische klassifikatiesysteem accepteert objectdefinities, waarin zowel nominale variabelen met twee meetniveaus (0 — 1) als met meer dan twee niveaus (hierna aangeduid als multichotoom) alsook verschillende soorten schalen met een hiërarchie in posities zoals rangordeschalen, intervalschalen enz. zijn vertegenwoordigd.

Speciale rekenvoorschriften voor vaststelling van de gelijkheid van scores op de afzonderlijke operatoren waarborgen een zekere mate van handhaving van het speciale karakter van het meetconcept, dat de operator vertegenwoordigt.

2.1.0 Beknopte beschrijving van het verloop van de analyse

Een uitvoerige beschrijving van het principe van de analyse, de begeleidende condities en de daaruit afgeleide rekensystemen zou meer ruimte vergen dan in dit bestek beschikbaar is.

De geïnteresseerde lezer zij daarvoor verwezen naar de door de IBM beschikbaar gestelde programmaliteratuur. Tevens is dit onderwerp iets uitvoeriger behandeld in de jaarlijkse publikatie 'Rondje Marktonderzoek' van de N.V. Inter/View te Amsterdam.

De organisatie van een verzameling objecten (dataset) in groepen met een zo groot mogelijke interne homogeniteit wordt benaderd in een iteratieproces. Er is in deze analyse dus sprake van een in hoofdzaak heuristisch programma.

Dit blijkt dan ook duidelijk uit de aard van de maatstaven, die in de diverse stadia van het iteratieproces worden gehanteerd voor de evaluatie van de in elk stadium bereikte indeling.

De eerste maatstaf is reeds genoemd. Het betreft de overeenkomstindex S_{ij}.

De tweede belangrijke maatstaf is een subjectieve norm. Subjectief is zij in die zin, dat het de gebruiker van het programma vrij staat de waarde ervan zèlf te bepalen. Het betreft hier de brekingscoëfficiënt S*. Deze waarde wordt vastgesteld binnen dezelfde grenzen als waarbinnen S_{ij} zich beweegt t.w. $0{,}0 \leqslant S^* \leqslant 1{,}0$.

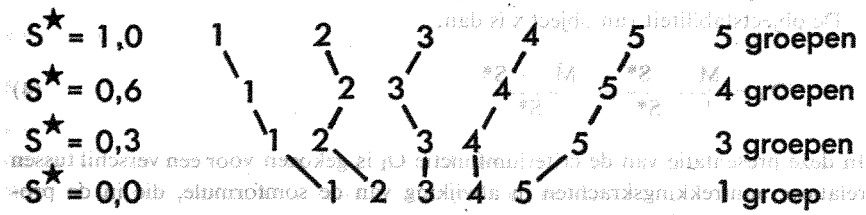
Deze drempelwaarde fungeert als toetssteen bij de ballotage van objecten voor het lidmaatschap van een bepaalde groep. Alléén die objecten worden tot de betreffende groep toegelaten, die voldoende nauwe betrekkingen onderhouden met de reeds bestaande leden van die groep. In termen van de gehanteerde maatstaven houdt dit in dat op zeldzame uitzonderingen na alle S_{ij} met deze leden groter moeten zijn dan S*.

Aannemende dat niet alle objecten van de onderzochte verzameling op een even groot aantal operatoren gelijke of als gelijk opgevatte scores vertonen, zullen de afzonderlijke S_{ij} in de basis-matrix sterk verschillende waarden hebben.

Met behulp van het criterium S* kan nu in eerste instantie structuur in de verzameling objecten (W_i) worden gebracht en tekenen zich één of meer groepen en een restgroep niet in te delen objecten af.

In onderstaande figuur is de relatie tussen de keuze van het niveau van S* en het karakter van de te verwachten groepsstructuur grafisch uitgebeeld voor eenzelfde hypothetische verzameling van 5 objecten (zie figuur 1).

Figuur 1



Een kleine afstand tussen twee of meer nummers beeldt een groep uit, een grote afstand een scheiding tussen twee groepen.

2.1.1 De Objectstabiliteit

De voornaamste maatstaf in deze analyse is de maatstaf, die de gevonden structuur evalueert naar de mate van interne homogeniteit resp. externe heterogeniteit zoals deze blijkt uit de relaties die nu binnen de groepen en tussen de groepen bestaan.

Deze maatstaf is de Objectstabiliteit ('Measure of Object Stability'). Het is een maatstaf die eerst kan worden toegepast als reeds een indeling in groepen is verkregen. Het programma genereert dus voortdurend nieuwe groepsstructuren en evalueert het resultaat achteraf aan de hand van deze maatstaf.

Aannemende, dat met behulp van toetsing van $|S_{ij}|$ aan de drempelwaarde S^* een bepaalde indeling in groepen is ontstaan, dan kunnen voor deze indeling de volgende parameters worden bepaald:

De binding van een object x met zijn groep ν_1 wordt voorgesteld door de gemiddelde som van alle S_{xj} met andere leden van de eigen groep

$$\frac{1}{\nu_1 - 1} \Sigma\ S_{xj} = M \qquad (2)$$

Hoe sterker de binding met de eigen groep, hoe vaker hoge waarden van S_{xj} zullen voorkomen; hoe groter M dus ook zal zijn.

Voor de initiële verdeling van de set in groepen $\nu_2, \ldots, \nu_p$ kan vervolgens worden bepaald de reeks waarden $\frac{1}{\nu_1} \Sigma\ t_{xj}$ voor alle groepen waarvan x geen deel uitmaakt[3]. Op dezelfde wijze als voor 2 geldt, geven deze parameters de binding van x met de 'uit'-groepen weer. Uit de gevonden waarden kan de hoogste worden gekozen. Dat is dus de voor x meest nabij gelegen 'uit'-groep. Deze hoogste waarde van

$$\frac{1}{\nu_1} \Sigma\ t_{xj} \text{ wordt aangeduid met het symbool } \bar{M} \qquad (3)$$

De objectstabiliteit is een maatstaf, die zowel S^* als M en $\bar{M}$ bevat en onderworden kan worden gebracht als een relatieve verschilwaarde van interne en externe aantrekkingskracht, gerelateerd aan de gekozen ballotagenorm S^*.

De objectstabiliteit van object x is dan:

$$O_x = \frac{M - S^*}{1 - S^*} - \frac{\bar{M} - S^*}{S^*} \qquad (4)$$

In deze presentatie van de criteriumfunctie O_i is gekozen voor een verschil tussen relatieve aantrekkingskrachten in afwijking van de somformule, die in de programmaliteratuur is gebruikt.

Het verschil is niet wezenlijk en ontstaat door tekenverwisseling in de tweede term van de formule.

Ons lijkt de hierboven gegeven formule meer correct in verband met een andere aanname in het programma, waarin de aantrekkingskracht van een object tot een lege groep geregeld wordt. Voor deze $\bar{M}$ geldt n.l. dat $\bar{M} = S^*$. Zolang het maximum aantal toegelaten groepen niet in gebruik is genomen geldt dus voor alle waarnemingen (objecten) W_i, dat $\bar{M} > S^*$.

Voor een stabiel geplaatst object (x) in een structuur, die is gebaseerd op een keuze van de brekingscoëfficiënt $S^* \approx \bar{S}_{ij}$ en waarin het aantal toegelaten groepen niet volledig is benut voor de indeling in groepen geldt als regel de relatie $S^* \leq \bar{M} < M$. Naarmate de discrepantie tussen $\bar{M}$ en M groter wordt zal O_x in waarde toenemen.

Neemt het verschil tussen $\bar{M}$ en M in grootte af, dan daalt ook O_x in waarde. Wordt $\bar{M} > M$, dan zal, afhankelijk van de waarde van S^*, O_x vroeger of later negatief worden; een kenmerk voor instabiliteit.

2.1.2 De gemiddelde en de totale objectstabiliteit

De gemiddelde en de totale objectstabiliteit zijn afgeleiden van (4) en spelen een grote rol in de beheersing van het iteratieproces.

De gemiddelde objectstabiliteit is een parameter, die een groep als geheel karakteriseert:

$$\bar{O}_{\nu p} = \frac{1}{\nu_p} \sum_{i=1}^{\nu p} O_i \tag{4a}$$

De totale objectstabiliteit is een criterium van de gehele verzameling objecten als zodanig (W_i) in een gegeven groepsverdeling t en luidt

$$E_t = \Sigma O_i \tag{4b}$$

2.1.3 Iteratiestadia

Het iteratieproces en de daarbij behorende kritische momenten kunnen nu als volgt worden beschreven:

Nadat het maximaal toegelaten aantal groepen is bepaald en het niveau van S^* is vastgelegd wordt voor alle n objecten van de verzameling O_n berekend. De dataset als geheel wordt als het ware als één groep opgevat. Voor de resterende p — 1 lege groepen geldt voor elk object dat $\bar{M} \equiv S^*$ en dus dat $\frac{S^* - M}{S^*}$ uit de somformule voor $O_i = 0$. Voor deze 'initiële' partitie wordt nu de waarde E_0 berekend welke dienst doet als toetsingsnorm voor het al of niet effectueren van een 'betere' (meer stabiele) groepsindeling als resultaat van een 'hill climbing'.

Een 'hill climbing' houdt in, dat de objecten, die in de 'één-groep-oplossing'

instabiel bleken te zijn worden overgebracht naar lege groepen, waarna wordt onderzocht of ze daar gezelschap kunnen krijgen van verwante objecten en met die objecten samen een min of meer homogene groep kunnen stichten. Hierbij worden voortdurend de waarden van de individuele, de gemiddelde en de totale objectstabiliteit berekend van alle interimoplossingen. Dit hill climbing proces wordt onderbroken op het moment dat een waarde E_1 wordt gevonden, die groter is dan E_0. Deze situatie wordt dan het eerste locale maximum genoemd. De verkregen waarde E_1 neemt nu de rol van minimumwaarde van E_0 over voor een volgende 'hill climbing'.

Voordat deze kan worden gestart worden de leden van alle nu gevormde groepen èn de leden van de residugroep volgens een bepaald systeem dooreengeschud. Het werkprincipe bij dit 'schudden' is dat elk object verhuist naar de groep waarop de waarde $\bar{M}$ van dat object in de bij E_1 verkregen groepsstructuur betrekking had. Deze hergroepering van de objecten heet 'forcing pass', een soort verbanning dus. Als alle objecten naar hun respectieve verbanningsoorden zijn overgebracht, vangt een nieuwe hill climbing aan volgens hetzelfde procédé als hierboven omschreven, totdat E_2 is bereikt. Is $E_2 > E_1$, dan is positieve voortgang geboekt op de weg naar een optimale groepsstructuur. Blijkt $E_2 \leq E_1$, dan eindigt daar de analyse en wordt de voorgaande groepsindeling (die de hoogst bereikbare E-waarde had) als 'final partition' beschouwd. Het proces wordt voortgezet tot $E_t \leq E_{t-1}$, waarna het programma alle gegevens publiceert van de structuur E_{t-1} als zijnde de optimale structuur.

De hierboven gegeven beschrijving is verre van volledig. Zo kent het programma bijvoorbeeld nog twee alternatieve procedures voor het vervolgen van een hill climbing na het bereiken van een locaal maximum of wat zich daarvoor aandient. Het zijn random 'passes' en 'assignment passes', die beide vooral dienst doen als een toets op de 'echtheid' van het locale maximum in kwestie. Het voert echter te ver om al deze details in dit bestek aan een beschouwing te onderwerpen.

2.2.0 Alternatieve overeenkomstindices voor 'geometrische' clusteranalyses

Het programma kent behalve het reeds genoemde analyseprincipe met $S_{ij} = \frac{Km}{K}$ nog enkele andere geometrische principes, zoals één met als basis de overeenkomstindex van Rogers en Tanimoto (1960) en één, die is genoemd naar de gehanteerde stabiliteitsmaatstaf, de Measure of Pairwise Strain.

De overeenkomstindex van Rogers en Tanimoto luidt:

$$S_{ij}\ (\text{Tan.}) = \frac{K_m}{2K - K_m} \qquad (5)$$

Deze vergelijkingsmethode zou volgens Rubin en Friedman (1967) het effect hebben van het contrasteren van verschillen tussen hoge en lage niveaus van S_{ij}.

De measure of Pairwise Strain is een maatstaf voor objectstabiliteit, die is gebaseerd op twee overeenkomstindices S_{ij} die alleen ex-post kunnen worden bepaald en een standaardbrekingscoëfficiënt $S^* = \frac{1}{2}$.

De functie voor de som (in dit geval het doorlopend product) van alle S_{ij} luidt

$$P = \Pi\gamma_{ij} \qquad (6)$$

waarbij $\gamma_{ij} = S_{ij}$ (Rub. en Friedm.) voor de 'eigen' groep òf
$= 1 - S_{ij}$ (,,) voor de 'uit' groep.

Elke keuze van de brekingscoëfficiënt als norm voor de partitionering ($S^{*\prime}$) leidt tot transformatie van alle γ_{ij} volgens de formule $(\gamma'_{ij})^t = \gamma_{ij}$ waarbij t zodanig is gekozen dat $(S^{*\prime})^t = S^* = \frac{1}{2}$.

De programma-auteurs delen als hun ervaring mee, dat deze clustermethode niet bevredigend werkt. Meestal werd één grote groep naast een groot aantal splintergroepen geproduceerd. Het enige voordeel van deze methode is gelegen in een veel grotere rekensnelheid.

2.3.0 De simultane verwerking van verschillende soorten variabelen in een objectdefinitie

In de programmabeschrijving in het nog te publiceren 'Rondje Marktonderzoek' zal dieper worden ingegaan op de talrijke mogelijkheden die het programma biedt om karakteristieke eigenschappen van verschillend geoperationaliseerde variabelen te verdisconteren in de berekening van S_{ij}.

In het kort kunnen de mogelijkheden als volgt worden aangeduid:

a. Toevoeging van 'dummy states' aan de schaal van de operator. De schaal wordt verlengd met een positie, die b.v. een 'match' veroorzaakt tussen twee objecten op de bewuste operator, waar de score zèlf een non-match opgeleverd zou hebben en omgekeerd. Deze maatregel voorkomt het optreden van S_{ij} ter waarde van 1,0 en 0,0, die in datasets met veel binaire kenmerken in de objectdefinitie anders zeer frequent zouden voorkomen.

Het voordeel van deze optie is dan, dat de binaire kenmerken qua invloed op de clustering worden ingeperkt en niet de neiging krijgen om over de schalen (operatoren met meer dan twee niveaus) te domineren. Bij schalen met meer dan twee niveaus kan deze optie worden gebruikt om bepaalde posities van de schaal een extra-gewicht toe te kennen bij het genereren van een groepsindeling.

b. Voor schalen waar aan een rangorde van niveaus een bepaalde hiërarchie is toegekend kan het programma naar believen scores op verschillende niveaus die echter dicht genoeg bij elkaar liggen nog opvatten als een 'gedeeltelijke' match.

c. Variabelen met een continu karakter (zoals lengte, temperatuur, etc.) kunnen als zodanig worden verwerkt. In het programma wordt dan de standaarddeviatie berekend van de aangetroffen frequentieverdeling en met behulp van de z–waarde worden de grenzen vastgesteld, waarbinnen waarden op verschillende niveaus nog als gedeeltelijke of volledige match worden aangemerkt voor de bepaling van S_{ij}.

3.0.0 Onderzoek naar het effect van toepassing van verschillende brekingscoëfficiënten

Een belangrijk onderzoeksdeel in het totale project is het experiment geweest, waarbij een tiental kleine clusteranalyses is uitgevoerd bij een variabel niveau van de brekingscoëfficiënt.

Wij hebben daarin een steekproef van 40 objecten getrokken uit het basisbestand van 496 ponskaarten, afkomstig van een onderzoek uitgevoerd door N.V. Inter/View betreffende de plaats van de huisvrouw in communicatiekanalen die relevant zijn voor informatie over nieuwe produkten. De trekking vond met behulp van een random-getallen-lijst plaats.

De analyse bestond in wezen uit 2 runs: een op 14 variabelen en een op 28 variabelen. In beide gevallen berekende de computer een gelijkheids-matrix op basis van de gekozen 40 objecten.

In de eerste stap, die van de 14-variabelen-berekening, herhaalde de computer, uitgaande van de aan het begin opgestelde matrix 5 maal een clustering. De brekingscoëfficiënten waarbij dit geschiedde waren: 0.2, 0.5, gemiddelde, 0.8 en 0.9. Dezelfde procedure werd toegepast op de 28-variabelen analyse. Hier moest uiteraard opnieuw een gelijkheidsmatrix worden berekend, omdat de gelijkheidswaarden tussen de objecten nu op een groot aantal extra variabelen berustten.

Een uitvoerig verslag van de resultaten is om redenen van beschikbare ruimte niet hier opgenomen. Het is gepubliceerd als onderdeel van de jaarlijkse uitgave van INTER/VIEW N.V. 'Rondje Marktonderzoek'.

De voornaamste conclusies uit dit experiment waren:

De verkregen klassificatie was ondanks de (te) geringe omvang van de dataset gedeeltelijk voor zinvolle interpretatie vatbaar. Uiteraard kwamen alleen de grootste groepen voor interpretatie in aanmerking. Deze interpretatie bevestigde in grote lijnen bestaande theorieën over demografische kenmerken en communicatiegedrag van huisvrouwen, waarop het onderzoekmateriaal betrekking had.

Deze eerste reeks experimenten op zeer kleine datasets heeft voorts duidelijk gemaakt dat:

a. De hantering van dichotome, multichotome en rangorde-schalen ter karakterisering van variabelen inderdaad mogelijk is zonder dat een duidelijk aanwijsbare bevoordeling aan de dag treedt voor hetzij korte hetzij lange schalen bij de keuze van de classificatiecriteria.

b. De 'Similarity-based' clusteranalyse heeft schijnbaar met alle typen multivariate analyse gemeen dat operatoren met een sterk a-symmetrische frequentieverdeling niet aan bod komen bij de selectie van groeps- of clustercriteria. De typische reactie van deze analyse op de aanwezigheid van zulke operatoren in de objectdefinitie is een algemene niveau-verhoging van de matrix S_{ij} en een vermindering van de clusterabiliteit van het materiaal.

Niet bekend is voorshands of deze sterk a-symmetrisch bezette operatoren, voorzover toegelaten in de objectdefinitie, nog belangrijke bijdragen leveren aan de afzondering van minderheidsgroepen. Dit laatste is zeker niet onaannemelijk, doch voorlopig ontbreken nadere aanwijzingen.

Men zal eerst een experiment moeten uitvoeren met een aanzienlijk grotere dataset voordat hierover uitspraken mogelijk zijn.

Een dergelijk experiment stuit echter op practische bezwaren, welke een dergelijke observatie in de weg staan. Hierop komen wij in dit verslag nog nader terug.

4.0.0 Probleemstelling voor stabiliteitsonderzoek

De toepassing van een non-hiërarchische clusteranalyse op basis van een matchcoëfficiënt als overeenkomst-index (in ons geval met optimaliseringsprocedures: hill-climbing en herstructurering) heeft voor steekproefdata enkele theoretische nadelen waarvan de practische reikwijdte diende te worden onderzocht. Waar de gehanteerde schatter voor overeenkomst tussen twee waarnemingen in deze analyse geen enkele verwantschap vertoont met gangbare, op Euclidische distantiemaatstaven gebaseerde overeenkomstindices, blijft 'normalisatie' van deze maatstaf voor de eigenlijke clusteranalyse achterwege. Als gevolg van dit gemis wordt toepassing van significantietoetsen op het gevonden resultaat een zeer gecompliceerde kwestie. Voor zover ons bekend is er nog geen oplossing voor dit probleem voorhanden.

Om toch een inzicht te krijgen in de significantie van de door dit analysetype gegenereerde clusters is daarom een redenering gevolgd die neerkomt op consistentieonderzoek binnen de beschikbare set van steekproefdata. De in een steekproef gevonden clusterstructuur is representatief voor de structuur van de bijbehorende populatie, indien het resultaat reproduceerbaar is in een nieuwe ad random getrokken steekproef van dezelfde populatie. Men kan zich een indruk vormen van deze eigenschap door een beschikbare dataset te partitioneren in een willekeurig

aantal subsets ($n_1 + \ldots + n_r$), volgens een random procedure, en de clusteranalyse achtereenvolgens te herhalen op sub-set n_1, $n_1 + n_2$, $n_1 + n_2 + \ldots + n_r$.

Het is dan eenvoudig een kwestie van vergelijken van alle achtereenvolgens bereikte clusterstructuren om vast te stellen bij welke $n_1 + \ldots + n_r$ de gevonden structuren onderling zo weinig verschillen, dat men van een stabiele oplossing mag spreken. Vooropgesteld dat bij een steekproeftrekking van de N observaties zowel als bij de partitionering van N observaties in $n_1 + \ldots + n_r$ subgroepen steeds aan de random voorwaarde is voldaan, is het aannemelijk dat de eerste stabiele oplossing tevens tenminste representatief is voor de te schatten populatiestructuur.

Weliswaar is daarmee geen betrouwbaarheidsinterval bekend voor de resultaten, doch met inachtneming van dit voorbehoud zijn de gevonden relaties dan wel als practisch bruikbaar te beschouwen.

Prof. Van Tulder signaleerde terecht, dat de definities van de overeenkomstindices in deze 'match-methoden' (Rogers en Tanimoto 1960, Rubin en Friedman 1967) op zichzelf aanleiding kunnen geven tot zeer instabiele resultaten, en wel vooral daar waar men uitkomsten uit kleine datasets tracht te reproduceren na aanzienlijke uitbreiding van de set.

Wanneer wij de overeenkomst-index $S_{ij} = \frac{K_m}{K}$ hanteren, dan worden de data voor de bewerking getransformeerd in een matrix S_{ij} ter grootte van $\frac{n(n-1)}{2}$ elementen.

Indien de waarnemingen W_1 en W_2 op 10 operatoren zijn gedefinieerd waarvan er 5 een volle match opleveren, dan is $S_{12} = \frac{5}{10} = \frac{1}{2}$.

Het is dan denkbaar dat waarneming W_2 en waarneming W_3 op de 5 overgebleven operatoren ook een volle match te zien geven, zodat uiteindelijk de volgende relatie geldt:

$$\begin{tikzcd} & W_1 & \\ & S_{1.2} = 0{,}5 / S_{1.3} = 0{,}0 \arrow[ul] \arrow[ur] \arrow[dl] \arrow[dr] & \\ W_2 & S_{2.3} = 0{,}5 \arrow[l] \arrow[r] & W_3 \end{tikzcd} \tag{7}$$

Gegeven deze eigenschap van het criterium S_{ij} kan een toevoeging van n_2 obser-

vaties aan de oorspronkelijke set n_1 ingrijpende gevolgen hebben voor de opnieuw gegenereerde clusterstructuur.

De basisgegevens voor de analyse worden getransformeerd in een matrix S_{ij} met $\frac{n(n-1)}{2}$ elementen.

Uitbreiding van de set met Δn observaties veroorzaakt een uitbreiding van de matrix S_{ij} met:

$$\frac{\Delta n\ (\Delta n - 1)}{2} + n.\Delta n \text{ elementen} \tag{8}$$

Een verdubbeling ($n_1 = n_2$) van de dataset veroorzaakt een uitbreiding van $2n^2 - n$ elementen.

Kleine afwijkingen in de datastructuur kunnen in zo'n geval aanzienlijke deformaties veroorzaken in de matrix S_{ij}.

Dit theoretisch bezwaar kan echter practisch van slechts geringe betekenis zijn, namelijk wanneer relaties van het type 7 tot de uitzonderingen behoren. Is bovendien aan random voorwaarden voldaan bij trekking van sequente steekproeven, dan is de verwachting dat een toevoeging van Δn geen disproportionele uitbreiding van dit type relaties oplevert.

Is onder die omstandigheden de Δn relatief klein, dan moet in het sequentie-experiment een fase bereikbaar zijn waarin deze oorzaak van instabiliteit verwaarloosbaar klein is geworden. De hier opgeworpen vraag is eens te meer pregnant, omdat de toepassingskosten van het hier behandelde programma progressief toenemen bij een uitbreiding van het aantal waarnemingen (objecten). De besparing op de analysekosten wordt indrukwekkender naarmate de omvang van de voor de clusteranalyse gekozen steekproef kleiner kan zijn. Hoe kleiner de n_1 waarbij de invloed van bovenstaande soort incongruenties verwaarloosbaar klein wordt bij extrapolatie van het gevonden resultaat, hoe toepasselijker de analyse wordt.

4.1.0 Keuze van de stabiliteitscriteria

Bij de vergelijking van de clusterstructuren gevonden na analyse in achtereenvolgens n_1, $n_1 + n_2$, $n_1 + n_2 + \ldots . n_r$ sequente steekproeven kan men de volgende indicaties voor stabiliteit van de gegenereerde clusterstructuur hanteren:

4.1.1 Clusteridentiteit

Indien men de analyse op n_1 observaties opvat als voortgekomen uit de analyse op $n_1 + n_2$ observaties dan is bij volstrekte stabiliteit de verwachting dat de resultaten van n_1 respectievelijk $n_1 + n_2$ observaties een identieke indeling in gelijke groepen oplevert waarbij voor n_1 geldt:

1 dat dezelfde clusters worden aangetroffen, verminderd met n_2 observaties
2 dat elke cluster in absolute omvang evenredig kleiner wordt met de verkleining van de dataset.

Een moeilijk te beantwoorden vraag is of bij een uitbreiding van de dataset een splitsing van een cluster in een of meer subclusters ook ontoelaatbaar moet worden geacht.

Onzes inziens is een dergelijke gebeurtenis voor practische toepassing wel toelaatbaar, indien men daar de analyse toepast met de wetenschap dat de beperkte omvang van de dataset geen grote detailleringen mogelijk maakt. Wel moet dan worden verlangd dat de gehanteerde stabiliteitsmaatstaf een relatief lage waarde te zien geeft, zodat het optreden van dit fenomeen bij uitbreiding van de dataset met redelijke zekerheid kan worden voorspeld.

4.1.2 Objectstabiliteit en gemiddelde groepsstabiliteit

In het programma zelf wordt gebruik gemaakt van de objectstabiliteit O_i (zie (4)). Voor de groep als geheel hanteert men de splitting function:

$$\bar{O} = \frac{1}{\nu_1} \sum_{i=1}^{\nu_1} O_i \tag{9}$$

Kleine waarden van $\bar{O}$ indiceren hoge $\bar{M}$ respectievelijk lage M waarden en zullen tevens vaak optreden bij:

$$S^* > \frac{1}{N} \Sigma S_{ij} \tag{10}$$

zoals uit substitutie in (4) kan worden afgeleid.

4.1.3 Een distantiemaatstaf voor groepsscores op afzonderlijke variabelen

Voor de analyse van steekproefdata is voorwaarde voor de stabiliteit van de gevonden groepsstructuur in een random partitie n_1 objecten ($n_1 \subset N$) dat een toevoeging van n_2 random getrokken objecten ($n_2 \subset N - n_1$) een clusterstructuur oplevert, die binnen bepaalde grenzen per overeenkomstige groep eenzelfde patroon van relaties van groepsscores versus totaalscores op K variabelen vertegenwoordigen als in de clusteranalyse op n_1 objecten.

De mate van dissociatie van de groepskarakteristiek van groep ν_1 binnen de steekproef n_1 moet dezelfde zijn als van ν_1', welke is gegenereerd binnen twee ge'match'te steekproeven $n_1 + n_2$, die voor de analyse zijn samengevoegd, wanneer ν_1' de pendant is van ν_1 uit n_1.

Aangenomen werd, dat, als bij twee achtereenvolgende vergrotingen van de dataset (steekproef) n_1 met gematchte steekproeven n_2 en n_3 deze dissociaties voor verwante groepen ν_1 (uit n_1), ν_1' (uit $n_1 + n_2$) en ν_1'' (uit $n_1 + n_2 + n_3$) een zelfde orde van grootte vertoonden over alle K operatoren, waaruit deze karakteristiek is opgebouwd, deze zelfde dissociatie geacht kon worden te bestaan in de groep N_1 uit de populatie N_p.

De mate van dissociatie d moet worden uitgedrukt in een maatstaf, die voor alle soorten operatoren op overeenkomstige wijze kan worden berekend.

Is deze maatstaf beschikbaar en vergelijkt zij b.v. de gemiddelde scores X_i over K operatoren afzonderlijk dan kan de hypothese over de representativiteit worden aanvaard, indien blijkt dat:

$$d\left(\frac{1}{v_1}\sum_{i=1}^{v_1} X_{ik}, \frac{1}{n_1 = v_1}\sum_{i=v_1+1}^{n_1-v_1} X_{ik}\right) \approx$$

$$\approx d\left(\frac{1}{v'_1}\sum_{i=1}^{v'_1} X_{ik}, \frac{1}{n_1 + n_2 - v'_1}\sum_{i=v'_1+1}^{n_1+n_2-v'_1} X_{ik}\right) \qquad (11)$$

en zo vervolgens voor alle verwante groepen en alle operatoren.

Deze maatstaf is van vitaal belang omdat men zeker in marktonderzoek in de eerste plaats behoefte heeft aan geldige uitspraken over groepskenmerken. Tevens is echter voor marktonderzoek een van de grootste problemen de keuze van een maatstaf die als distantiecriterium simultaan voor de diverse soorten operatoren kan worden gehanteerd.

Het klassieke probleem is de veelsoortigheid van de gehanteerde meetmethodes (schaaltechnieken) waardoor de operatorschaal dichotoom alsook multichotoom kan zijn, doch evenzeer een rangorde of intervalschaal, of zelfs een ratioschaal kan vertegenwoordigen.

Rubin & Friedman geven in het programma voor 'continuous' en 'rankorder'-variables gemiddelden en standaarddeviaties kennelijk met als oogmerk een distantiemaatstaf. Om begrijpelijke redenen zien zij voor de 'discrete variables' (in ons geval de nominale variabelen) van het gebruik van deze criteria af.

Hiermee echter gaat de mogelijkheid verloren om alle relevante variabelen op 'goodness of fit' te vergelijken en wordt een evaluatie van de stabiliteit van de clusterstructuur als zodanig een hoogst onoverzichtelijke en subjectieve aangelegenheid.

Naar ons oordeel zou hier de hantering van de Cramèr's associatiecoëfficiënt φ'[4] het voordeel hebben opgeleverd dat hij voor alle voorkomende typen operatoren een geldige associatie-maatstaf is. Weliswaar is hij als toets in sommige gevallen relatief inefficiënt (omdat er voor intervalschalen uiteraard een efficiëntere toets is), maar het is toch redelijk dit voordeel aan het voordeel van vergelijkbaarheid op te offeren.

Een ander bezwaar van de φ' associatie-coëfficiënt is, dat bij een zeer kleine n de waarde van φ' soms niet geheel onafhankelijk meer is van n. Dit heeft zijn oorzaak in de toepassing van de correctie van Yates bij de berekening van de waarde van χ^2 voor de dichotomieën. Bij een groter aantal df krijgt men daarentegen de beschik-

king over de likelihood ratiotoets voor de berekening van χ^2, welke toets ook een goede match geeft voor χ^2 bij kleine n zonder een correctie in de teller van de breuk waarin n voorkomt.

Het gevolg is bij de vergelijking van de behaalde waarden in φ' voor de afzonderlijke variabelen tussen kleine en grote steekproef-analyses wèl, dat een correctie voor φ' noodzakelijk is indien df = 1 en geen correctie indien df $>$ 1.

De correcties blijven echter òf gering, òf verliezen hun waarde zodat wij ook dit bezwaar niet als doorslaggevend hebben aangemerkt.

Bij de berekening van de associatiecoëfficiënt over groepsscores op operatoren met meer dan twee niveaus is overigens om redenen van technische aard wel een transformatie toegepast. Hierbij werd elke schaal tot drie niveaus teruggebracht. De invloed van deze vereenvoudiging op het niveau van φ' werd als verwaarloosbaar klein opgevat[5].

5.0.0 Experiment

Als datamateriaal voor de test deed dienst het resultaat van een enquête onder 496 huisvrouwen uitgevoerd door N.V. Inter/View, waarin vragen over communicatiegedrag t.a.v. koopbeslissingen en over de plaats van de huisvrouw in het communicatienetwerk over (nieuwe) produkten waren opgenomen.

De ontbrekende scores op de operatoren die voor de analyse in aanmerking kwamen werden geschat op grond van berekende correlaties met de bekende waarnemingen, dan wel random verdeeld over klassen in die gevallen dat een ruwe indicatie nader moest worden gepreciseerd en aangevuld.

Via een 2-traps procedure werd de dataset verdeeld in 10 random subsets van 40 objecten en een restgroep van 96 objecten.

Uit deze 10 subsets werden er 3 gekozen op grond van een zo groot mogelijke match met de totaalscores van alle operatoren en een goede onderlinge match.

Uit alle voor de definitie van het object beschikbare operatoren werden achtereenvolgens 9 resp. 12 operatoren gekozen zodanig, dat van alle toegepaste schaaltypen elk type zo veel mogelijk vertegenwoordigd was. De samenstelling van de operatoren-set was in beide sequentie-experimenten als volgt:[6]

Tabel 1

Aantal operatoren van het schaaltype:	*Experiment met:* 9	12
dichotoom (0–1)	3	4
multichotoom	2	2
rangorde	3	4
ratio	1	2

In beide gevallen werd de analyse uitgevoerd op datasets binnen 3 onderling zo goed mogelijk gematchte steekproeven n_1; n_2; n_3 van elk 40 objecten met als programma:

run 1	n_1	40 obj.
run 2	$n_1 + n_2$	80 obj.
run 3	$n_1 + n_2 + n_3$	120 obj.

De volgende optiekeuze was voor alle experimenten dezelfde t.w.:

1. $S_{ij} = \frac{K_m}{K}$ (overeenkomstindex volgens Rubin & Friedman)
2. $S^* = \frac{1}{N} \Sigma S_{ij}$ (standaardoptie)
3. geen dummyvariabelen en dummystates toegelaten ter onderdrukking van full matches op binaire kenmerken.
4. het maximum aantal groepen (25) was toegelaten.

Wel gevarieerd werd de startoptie. In een enkel geval werd er geëxperimenteerd met de optie: start met gegeven initiële partitie. Deze optie werd getest om te zien of de output van een $n_1 + n_2 + n_3$ analyse gebruikt kon worden voor het bereiken van een aanzienlijke besparing van Cpu-tijd in de analyse van een dataset $n_1 + n_2$, en had geen enkele invloed op de gedaante van de uiteindelijk gerealiseerde clusterstructuur.

5.1.0 Resultaten

Het experiment, dat door periodiek gebrek aan voldoende geldmiddelen veel tijd heeft gekost en niet geheel voltooid kon worden, leverde niettemin zeer leerzame resultaten op. Van deze resultaten zullen we speciaal de analyse gericht op onderzoek van de stabiliteit van het gevonden resultaat nader belichten.

5.1.1 Clusteridentiteit

In de in de respectieve runs gegenereerde groepen werden de objectnummers vergeleken. De diagrammen 2 en 3 geven het resultaat weer van de identificatie van de objecten in de beide sequentie-experimenten.

Naar de omvang van de gevormde groepen beoordeeld geven beide experimenten ongeveer eenzelfde ontwikkeling te zien. De omvang van verwante groepen varieert sterk bij de reductie van 80 naar 40 objecten. Diezelfde variatie wordt aanzienlijk kleiner bij de reductie van 120 naar 80 objecten. In de run met 9 operatoren verdwijnt er bij de reductie van 80 naar 40 objecten zelfs een groep ter grootte van 7,5% van alle waarnemingen. Bij de andere reducties verdwenen er hoogstens groepen ter grootte van 2,5%.

Ook de omvang van de overige mutaties indiceert, dat het experiment met 9 operatoren iets meer variaties toelaat in de groepsomvang dan dat met 12 operatoren.

Diagram 2 TRANSFORMATIE VAN CLUSTERSTRUCTUUR IN 3 RUNS VAN 120-80-40 OBJECTEN EN 9 OPERATOREN

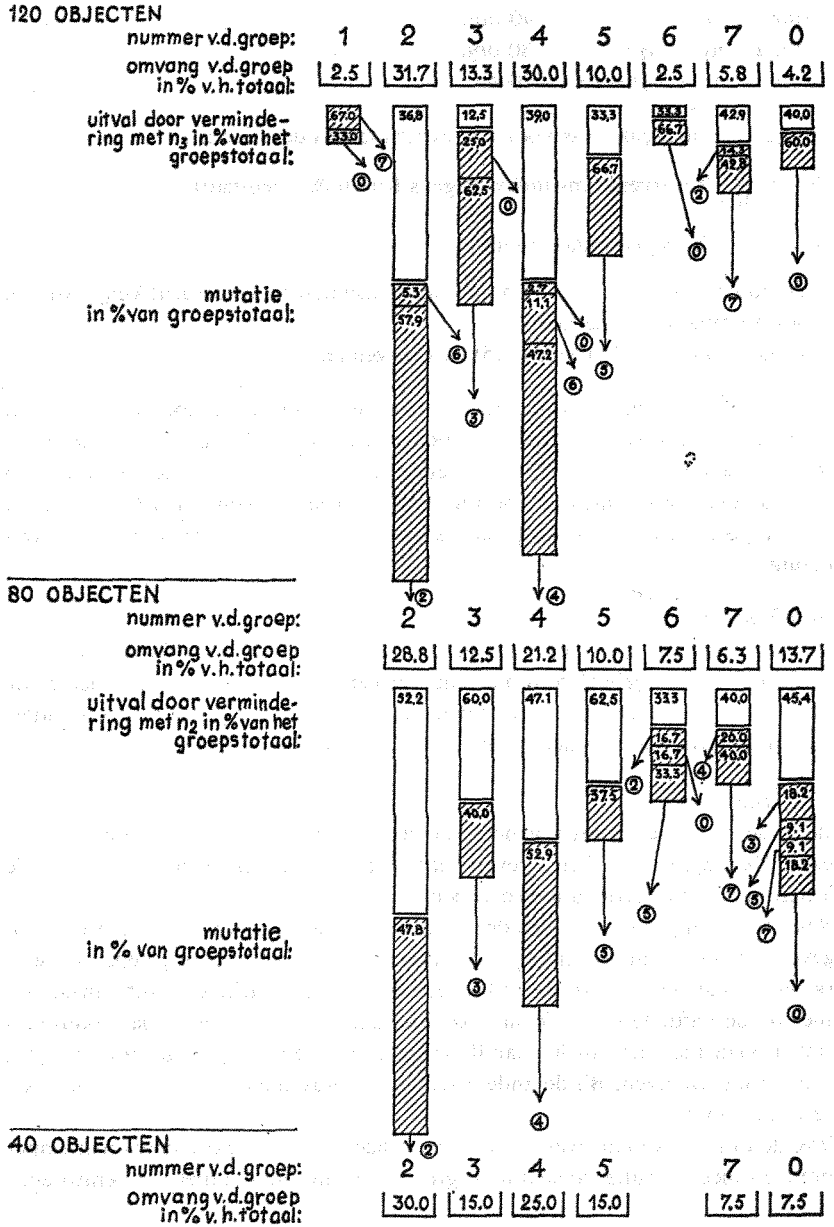
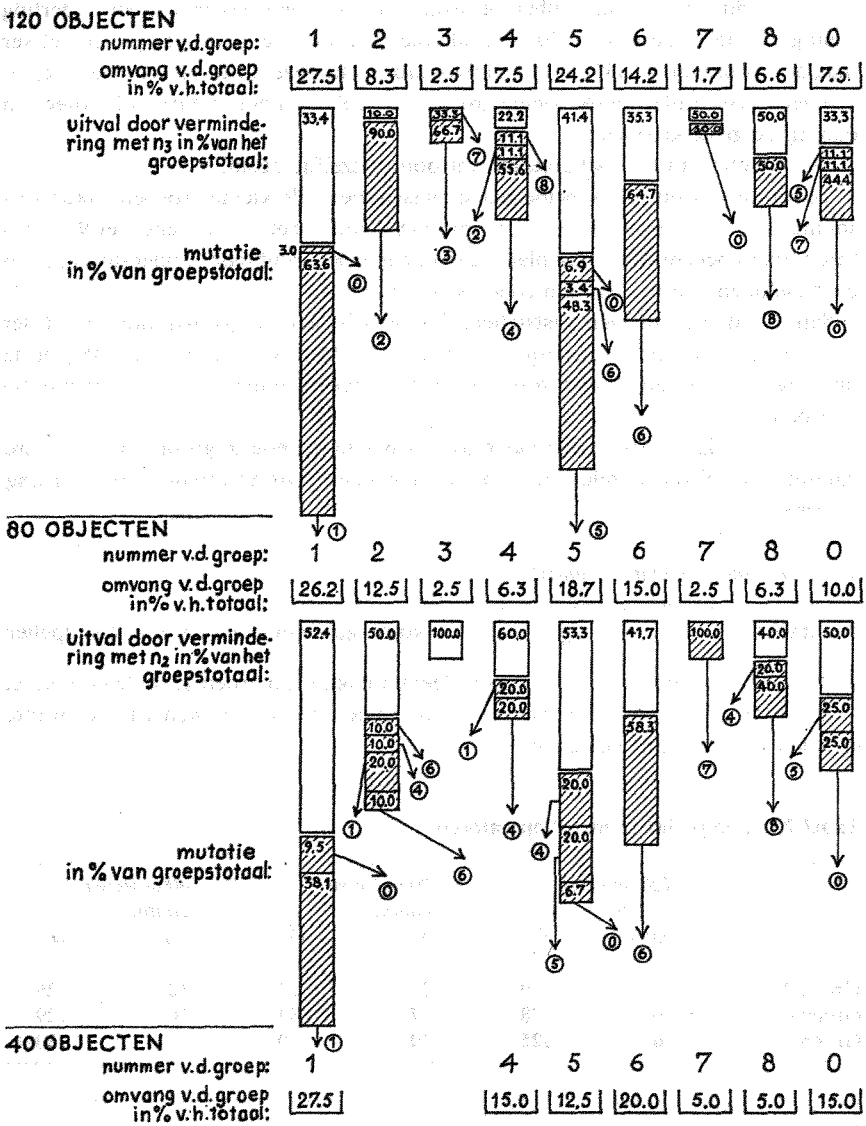


Diagram 3 TRANSFORMATIE VAN CLUSTERSTRUCTUUR IN 3 RUNS VAN 120-80-40 OBJECTEN EN 12 OPERATOREN



Voor wat betreft de fracties van de groepen, die tijdens de reductie afvallen kan een norm gesteld worden en vergeleken met de feitelijke uitval.

Ook in dit opzicht verschillen de resultaten van beide experimenten onderling weinig. Bij de reductie van 120 naar 80 objecten, waar de uitval per groep niet ver mag afwijken van $\pm$ 33%, zijn de afwijkingen van deze norm vrij aanzienlijk, alhoewel excessen alleen bij de kleine groepen (< 15% van het totaal) veel groter zijn dan bij de grote groepen.

De reductie van 80 – 40 objecten vertoont hetzelfde beeld.

Tenslotte valt op dat in beide reductiefasen diezelfde kleine groepen vaker versplinteren dan de grote. Deze versplintering vindt in het experiment met 9 operatoren fractioneel minder vaak plaats dan in het experiment met 12 operatoren, doch de verschillen zijn te klein om conclusies toe te laten.

Samenvattend kan men vaststellen, dat zelfs bij een vergroting van de dataset tot 120 objecten ook in de grote groepen in de clusterset variaties tot 50% in de omvang van de groep kunnen voorkomen (zie experiment met 9 operatoren: mutatie in groep 4).

Deze variatie is voor marktonderzoek-toepassingen nog te groot, zodat grotere datasets noodzakelijk zullen zijn om tot een goede schatting van de groepsomvang te komen.

5.1.2 Gemiddelde objectstabiliteit

De statistic $\frac{1}{\nu_p} \sum_{1}^{\nu_p} \frac{M - S^*}{1 - S^*} + \frac{S^* - \overline{M}}{S^*}$ wordt gepresenteerd als een voorspeller voor de coherentie van de objecten die deel uitmaken van groep Π in de populatie.

Wanneer wij de gevonden waarden voor beide runs uitschrijven voor verwante groepen krijgen wij de volgende overzichten:

Tabel 2. A. Experiment met 9 operatoren

	120 objecten		*80 objecten*		*40 objecten*	
	aantal obj.	$\overline{O}$	*aantal obj.*	$\overline{O}$	*aantal obj.*	$\overline{O}$
Groep 2	38	.29	23	.27	12	.26
Groep 4	36	.28	17	.33	10	.29
Groep 3	16	.25	10	.30	6	.30

B. Experiment met 12 operatoren

	120 objecten		*80 objecten*		*40 objecten*	
	aantal obj.	$\overline{O}$	*aantal obj.*	$\overline{O}$	*aantal obj.*	$\overline{O}$
Groep 1	33	.30	21	.30	11	.34
Groep 6	17	.27	12	.27	8	.28
Groep 2	10	.18	10	.15		
Groep 5	29	.26	15	.28		

Uit beide overzichten blijkt, dat een toename van N resp. n_1 geen aanleiding geeft tot een toename van $\overline{O}$.

Voor de hier gereproduceerde cijfers blijkt een tendens tot handhaving van het eenmaal bereikte niveau.

Uit de hier niet gereproduceerde cijfers voor de kleinere groepen kan men duidelijk constateren, dat indien de omvang van de groep zeer klein is, de gemiddelde objectstabiliteit sterk kan stijgen als resultaat van de geringe omvang van de groep zelf. Dit is ook zeer voor de hand liggend en vloeit direct voort uit wat reeds eerder is opgemerkt over de hoedanigheden van de matrix S_{ij} zelf. Hoe kleiner het aantal S_{ij} hoe groter de kans op symmetrie in de onderlinge relaties (a–b), (b–c), (c–a).

Zouden we al aannemen dat de stabiliteit van de gevonden groepen een neiging tot toenemen moet vertonen naarmate de dataset groter is, dan wordt in *deze* analyse dat beeld weer teniet gedaan door deze typische hoedanigheid van

$$\frac{1}{\nu_p} \sum_{1}^{\nu_p} O_i.$$

Deze speculatie sluit echter geenszins alle andere verklaringen uit, zodat er naar ons oordeel reden is voor de conclusie, dat de gemiddelde objectstabiliteit wel een bruikbaar constructie-criterium is voor de analyse zelf, doch niet voldoet als criterium ter beoordeling van de reproduceerbaarheid van het resultaat in een grotere steekproef.

5.2.0 Individuele objectstabiliteit

Het introduceren van nieuwe relaties tussen elementen binnen de bestaande dataset als gevolg van uitbreiding met nieuwe elementen behoeft ook bij steekproefmateriaal geen verontrustend verschijnsel te zijn, zolang deze relaties nog binnen de marge vallen, waarmee stochastische waarnemingen behept zijn.

De individuele stabiliteit van de leden van een groep kan ook nog als maatstaf dienen bij een andere analyse van de achtereenvolgens gevormde en onderling verwante groepen. Als men na de bepaling van de identiteit van de gevormde groepen op grond van de in de groep aangetroffen leden deze leden rangschikt naar de grootte

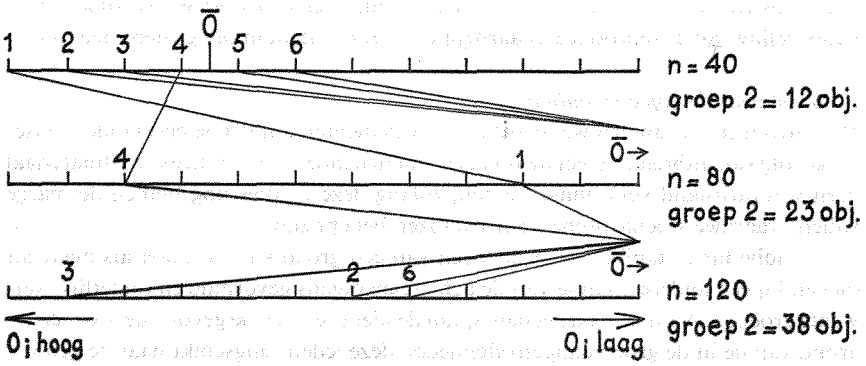
van de maatstaf O_i dan kan men veronderstellingen maken voor de rangorde van de plaats waarin men de bestaande leden in de uitgebreide groep moet aantreffen. De in het klassificatieproces nieuw gevormde groep moet, als zij geheel identiek is met de eerder geklasseerde en verwante groep, aan de oude leden een rangorde toekennen, die overeenkomt met de vorige. Onder de nieuw toegevoegde leden kunnen en zullen stellig ook leden zijn met een hoge stabiliteitswaarde, doch zij moeten òf de oude hooggeklasseerde leden en bloc verbannen naar een lagere plaats òf deze leden in hun gelederen opnemen, b.v. door interpolatie. In de nieuwe reeks van stabiliteitswaarden, gerangschikt van hoog naar laag, moeten de oude hooggeklasseerde leden dus nog steeds op een hoge plaats worden aangetroffen. Als de oude leden A1, A2, A3 zijn genoemd naar de hoogte van hun oorspronkelijke stabiliteitswaarden en de nieuwe leden B1, B2, B3 dan moet de nieuwe reeks deze hiërarchie op een of andere wijze respecteren b.v. A1, A2, B1, A3, B2, B3 .... of iets dergelijks.

In beide experimenten zijn de grootste drie verwante groepen op deze eigenschap onderzocht.

Het zou eentonig worden wanneer voor elke groep uit elk experiment uitvoerig verslag werd gedaan van de feitelijke rangorde. Hieronder volgen daarom twee voorbeelden van aangetroffen hiërarchieën, die typisch zijn voor de gehele analyse. De ene vertegenwoordigt de groep, die het beste aan de gestelde eis beantwoordt en de andere de groep die het slechtst voldeed op dit punt.

In onderstaande diagrammen is de plaatsverschuiving aangegeven van de eerste zes leden van de betrokken groep (naar rangorde van de gevonden stabiliteitswaarde in de klassificatie van 40 elementen) na herhaalde toevoeging van 40 nieuwe elementen (resp. 80 en 120 elementen). De leden zijn genummerd naar de rangorde van hun plaats in de analyse op 40 elementen en behouden dit rangnummer in de vervolganalyses in het diagram. De diagrammen tonen de lotgevallen van de eerste 6 leden.

Fig. 4

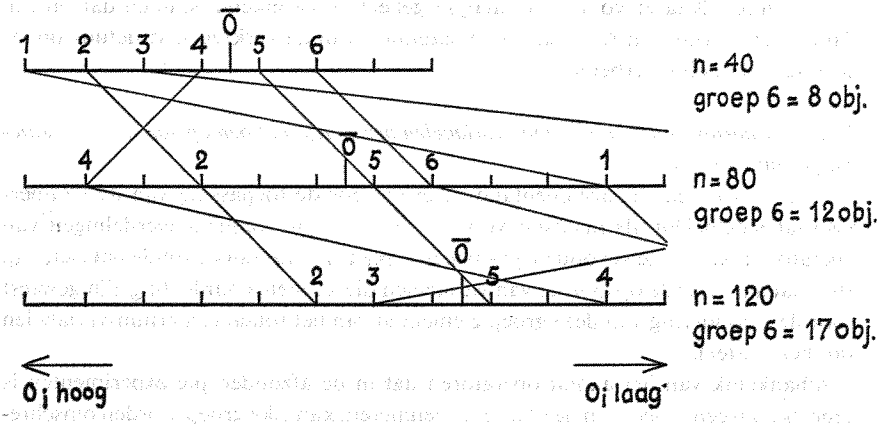


Het minst geslaagd verliep de rangschikking bij de groepen 2 (12 leden in de 40 analyse; 23 leden in de 80 analyse; 38 leden in de 120 analyse) bij het experiment met 9 operatoren (zie fig. 4). De plaats van O gemiddeld ($\bar{O}$) is ook in het diagram aangegeven.

4 van de 6 leden werden na de uitbreiding van 40 naar 80 radicaal van hun hoge plaats gedrongen. Sommige verdwenen zelfs geheel. De uitbreiding tot 120 elementen bracht 3 van de 6 leden terug op hun oorspronkelijk hoge plaatsen zij het in een radicaal andere volgorde. Voor een zo omvangrijke groep geen gunstig resultaat.

Des te merkwaardiger is daarom het tweede hier gerapporteerde resultaat over de groepenreeks 6 in het experiment met 12 operatoren. Het betrof hier een qua omvang zeer bescheiden groep. Niettemin kwamen 5 van de 6 oorspronkelijk hoogst geklasseerde leden inderdaad op de eerste 12 plaatsen van de nieuwe groep terecht (zie fig. 5) en 4 van de eerste 6 na verdere uitbreiding op de eerste 12 plaatsen van de nieuwe groep in de 120-analyse.

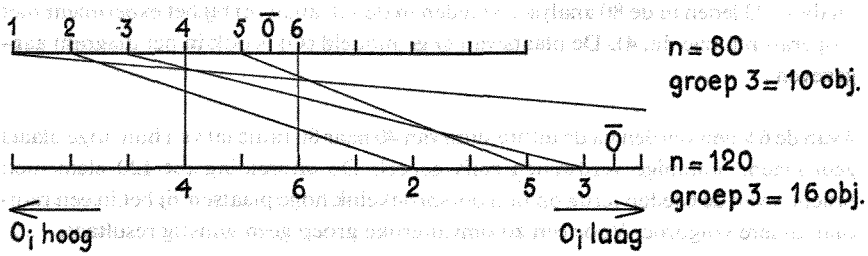
Fig. 5



De lijnen geven aan hoe de objecten met de grootste object-stabiliteiten verschuiven tussen de opeenvolgende data-bestanden.
$\bar{O}$ = gemiddelde objectstabiliteit van de groep.

Deze beschouwingswijze illustreert onzes inziens enigszins de stabiliteit van de gevormde groepen en de invloed die toevoeging van nieuwe elementen uit een ge-'match'te steekproef heeft op de nieuwe klassificatie.

Fig. 6



Wanneer wij het eerder aangenomen startpunt voor deze analyse echter vervangen door de 80-objecten analyse, dan blijkt het resultaat aanzienlijk gunstiger uit te vallen.

Men zou nu de eis moeten stellen dat de eerste 6 objecten uit de 80-analyse op de eerste 9 plaatsen van de overeenkomstige groep in de 120-analyse terecht moeten komen. De groepenreeks 3 in de analyse op 9 operatoren diene als voorbeeld van het resultaat dat nu wordt verkregen.

Alleen objectnummer 3 overschrijdt de gestelde grens en komt op positie 10 terecht. Dit is onzes inziens ook een argument voor het standpunt dat 40 objecten als omvang van een dataset voor deze analyse geheel ontoereikend is, doch dat bij een dataset ter grootte van 80 objecten de stabiliteit van de verkregen structuur op dit criterium duidelijk verbetert.

5.3.0 Opsporing van kenmerkende variabelen per groep met behulp van de associatiecoëfficiënt van Cramèr

In 4.1.3 werd reeds de argumentatie gegeven voor de toepassing van de Cramèr-associatiecoëfficiënt als maatstaf voor de mate waarin frequentieverdelingen van operatoren binnen de gevonden groepen afwijken van die van de totale dataset, c.q. als maatstaf voor de opsporing van variabelen die kennelijk aanleiding zijn geweest voor de afzondering van deze groep elementen van het totaal (criterium-variabelen van het cluster).

Afhankelijk van het aantal operatoren dat in de afzonderlijke experimenten is gebruikt om een object van de dataset te definiëren, kan elke groep worden omschreven in reeksen van 9 resp. 12 φ'-waarden. Voor stabiliteit van de gevonden clusters wordt nu als eis gesteld dat de reeksen-waarden, die voor verwante groepen werden verkregen ook een hoge mate van onderlinge overeenkomst moeten vertonen.

Noemen wij de verwante groepen uit twee opeenvolgende groeistadia in ons onderzoek v_a en v_a en berekenen wij de associatie-coëfficiënt voor alle operatoren afzonderlijk in de vorm van een vergelijking van de score v_{ak} van de groep zèlf met die van alle andere groepen binnen het bestand tezamen $(v_{ak}, (n-v_{ak}))$, dan geldt hier als criterium voor stabiliteit:

$$\varphi'(\nu_{ak}, (n - \nu_{ak})) \approx \varphi'(\nu'_{ak}, (n' - \nu'_{ak})) \quad (12)$$

Op waarden van φ', berekend uit de χ^2-waarden met df = 1 is een correctie van toepassing in verband met de afhankelijkheid van deze waarden van φ' van de grootte van de steekproef (dataset) n.

In de hieronder opgenomen overzichten zijn deze waarden van φ' dan ook afzonderlijk gegroepeerd.

5.3.1 *Resultaten*

De berekeningen zijn slechts uitgevoerd over die groepen uit de afzonderlijke clusteranalyses, waarvan de oorspronkelijke groep (gegenereerd in de dataset van 40 elementen) een omvang had die toetsing toeliet. In beide experimenten waren slechts twee van dergelijke groepen aanwezig. Een derde groep kreeg eerst in de analyse over 80 elementen een omvang van betekenis en is van daaraf ook in de berekening opgenomen.

Voor het experiment met 9 operatoren zag het overzicht van φ' waarden voor de groepenreeks 2 er als volgt uit:

Tabel 3

groep 2					
aantal objecten in de analyse			*120*	*80*	*40*
operator:					
nr.	*type*	*df*	φ'	φ'	φ'
3	nominaal	1	.24	.10	.15
4	rangorde	1	.10	.01	—
7	nominaal	1	.67	.59	.60
8	nominaal	1	.40	.40	.40
1	nominaal	2	.11	.21	.17
2	rangorde	2	.18	.25	.21
5	nominaal	2	.16	.39	.25
6	ratio	2	.15	.15	.29
9	rangorde	2	.23	.14	.28

N.B. Sommige nominale variabelen waren verwerkt als dichotome operatoren (0–1) en andere als multichotome. Bij deze laatste werden wèl meerdere niveaus onderscheiden, doch deze werden opgevat als niveaus waartussen geen andere relatie bestond dan onderlinge exclusiviteit.

Uit de vergelijking van overeenkomstige φ' waarden komt een duidelijk tekort aan stabiliteit aan het licht. Operator 7 (bezit bankrekening) heeft bij de classificatie kennelijk de functie van criteriumvariabele vervuld, doch de schommelingen in φ' waarden zijn vrij aanzienlijk.

Voor een deel mag deze toeneming van φ' bij stijging van n worden toegerekend aan de invloed van de Yates-correctie bij berekening van χ^2.

Het resultaat is dan echter weer in zoverre inconsistent dat voor operator 8 (bezit koelkast) die ook een belangrijke rol vervult bij de typering van de groep, en waarvoor de correctie ook opgaat, dit verschijnsel weer niet opgaat. Integendeel, hier liggen de φ' waarden weer keurig bij elkaar.

Voor multichotome operatoren blijkt het niveau van de φ'-waarden in de successieve datasets minder stabiel. Dit kunnen we zien in tabel 5 waar een overzicht wordt gegeven van de 4-reeks. Hier is operator 3 (abonnement op de consumentengids) een aardig voorbeeld van een 'instabiele' operator.

Tabel 4

groep 4					
aantal objecten in de analyse			*120*	*80*	*40*
operator:					
nr	*type*	*df*	*φ'*	*φ'*	*φ'*
3	nominaal	1	.23	.14	—
4	rangorde	1	.08	.05	.27
7	nominaal	1	.66	.50	.52
8	nominaal	1	.39	.32	.33
1	nominaal	2	.03	.24	.32
2	rangorde	2	.34	.42	.44
5	nominaal	2	.20	.28	.20
6	ratio	2	.15	.01	.34
9	rangorde	2	.38	.55	.71

De indruk wordt gewekt dat het overwicht dat dit type clusteranalyse geeft aan nieuw geïntroduceerde relaties zich vooral uit in grillige wijzigingen in de mate waarin variabelen op het tweede plan bijdragen tot de classificatie.

Sommige verdwijnen successievelijk, andere komen op. De belangrijkste criteriumvariabelen echter handhaven hun functie, evenals de rangorde waarin zij hun rol van criteriumvariabele vervullen.

Merkwaardig is, dat de groepenreeks 2 in het experiment met 9 operatoren vrijwel dezelfde leden omvatte als de groepenreeks 1 uit het 12-operatoren-experiment.

De toevoeging van operatoren voor: gezinsgrootte, het lezen van damesbladen en het lezen van nationale dagbladen gaf dus nauwelijks aanleiding tot hergroepering!

Bezien we de φ' waarden die in deze eerste reeks 2 bij 9 operatoren worden behaald op 12 operatoren (reeks 1), dan zien we dat operator 10 consistent de hoogste φ' scoort (zie tabel 5). Deze operator is dezelfde als 7 uit het eerste experiment (bezit bankrekening) en heeft zijn rol als criteriumvariabele dus behouden.

De operator 11 (bezit koelkast) wordt echter nu in betekenis overschaduwd door een nieuwe operator, n.l. gezinsgrootte. Tegelijk met deze nieuwe operator gaat nu een andere oude bekende een veel belangrijker rol spelen, nl. operator 8 (dezelfde als 5 uit het '9'-experiment). Het betreft hier een multichotome schaal voor gezinssamenstelling.

Het resultaat suggereert als verklaring voor dit verschijnsel dat het koelkastbezit 'spurious' is op gezinsgrootte, zodat bij weglating van deze criteriumvariabele (gezinsgrootte) het koelkastbezit als vervangende variabele wordt aangegrepen om een bijdrage te leveren voor de classificatie-'maat': bezit bankrekening. Deze immers blijft gehandhaafd als belangrijkste criteriumvariabele.

Weliswaar laat dit overzicht ook nog belangrijke schommelingen zien in verwante waarden van φ', doch deze kennelijke mogelijkheid tot een zinvolle interpretatie van de wisseling in betekenis van de afzonderlijke variabelen geeft alle hoop dat bij een grotere dataset het gedrag van de afzonderlijke variabelen vrij nauwkeurig kan worden geanalyseerd.

Voor de overige groepen die qua omvang deze analyse toelieten zijn de waarden van φ' ook berekend. Zij onderstrepen de hierboven opgeworpen suggesties, reden waarom wij ervan afzien deze resultaten ook nog uitvoerig te rapporteren.

Uit de overzichten kan men concluderen dat voor de variabelen die in de clusteranalyse duidelijk de rol van criteriumvariabele hebben gespeeld (gezien hun hoge waarde van φ') geldt, dat mutaties in de dataset geen invloed van betekenis hebben gehad op het vervullen van die rol.

Andere variabelen echter, die slechts een ondergeschikte bijdrage leverden aan de vorming van de clusters, geven soms flinke afwijkingen in het scorepatroon te zien bij mutaties in de dataset.

De verbetering in stabiliteit die deze achtereenvolgende uitbreidingen tot 120 elementen zou kunnen brengen is daar nauwelijks merkbaar.

De implicaties van deze conclusies zijn dat indien men de analyse op kleinere datasets toepast alleen die groepen betrouwbaar zijn welke $\pm 20\%$ of meer van alle waarnemingen omvatten. Over die groepen maken dan alleen uitspraken welke het scorepatroon van de criteriumvariabelen zélf betreffen én van de variabelen die volgens de Cramèr-statistic waarden van .30 en meer behalen aanspraak op geldigheid.

Waarschijnlijker is echter dat er vooral in het geval van steekproefwaarnemingen grotere datasets nodig zijn, bv. van ± 200 waarnemingen, om uit de resultaten van een taxonometrische clusteranalyse meer gedetailleerde conclusies te trekken zonder dat de geldigheid van de conclusies twijfelachtig zou kunnen worden.

Tabel 5

groep 1 score-reeks van φ' waarden bij 12 operatoren

aantal objecten in de analyse			*120*	*80*	*40*
operator: nr	*type*	*df*	*φ'*	*φ'*	*φ'*
3	ratio	1	.09	.09	.02
4	nominaal	1	.25	.22	.24
5	nominaal	1	.22	.17	.18
6	rangorde	1	.08	.16	.15
10	nominaal	1	.62	.54	.56
11	nominaal	1	.36	.37	—
1	nominaal	2	.09	.20	.17
2	rangorde	2	.31	.22	.46
7	rangorde	2	.42	.42	.52
8	nominaal	2	.41	.38	.41
9	ratio	2	.21	.19	.44
12	rangorde	2	.12	.31	.24

6.0.0 Kosten en commerciële bruikbaarheid van de analysemethode

Ten aanzien van de kosten welke aan het gebruik van dit programma zijn verbonden kunnen de volgende opmerkingen worden gemaakt.

Volgens de programmabeschrijving stijgt de cpu-tijd kwadratisch met de toename van het aantal objecten.

De cpu-tijd zou, naar de verwachting van Rubin en Friedman tevens lineair toenemen met het aantal groepen in de eindpartitie.

De door ons verrichte exercities zijn te gering in aantal en omvang, en naar hun aard onvoldoende op dit probleem afgestemd geweest dan dat hierover een afdoende antwoord kan worden gegeven.

In de resultaten welke wij hebben kunnen verzamelen blijkt deze 'groeps-invloed' vooralsnog niet aanwezig te zijn.

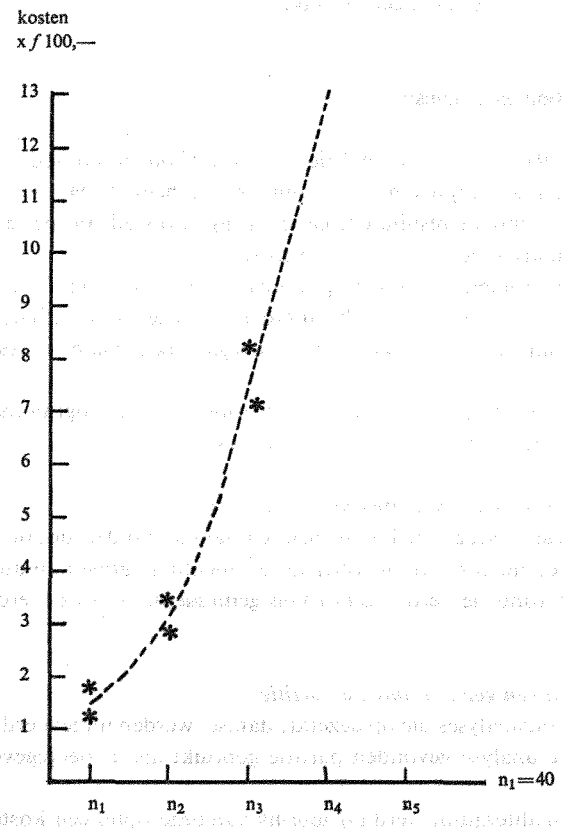
Gezien de positieve indicaties ten aanzien van de ontwikkeling tot commercieel gebruik van deze clusteranalyse dringt zich de noodzaak op een voorspelling te geven van de analysekosten bij uitbreiding van de dataset.

Daartoe hebben wij alle tot nu toe bekende verbanden in onze analyse tussen de kosten voor cpu-tijd en omvang van de dataset in beeld gebracht.

Tabel 6

aantal objecten	*aantal operatoren*	*cpu-tijd*	*kosten van de cpu-tijd (fl)*
40	9	.0127	125
80	9	.0434	297
120	9	.1510	828
40	12	.0238	182
80	12	.0557	350
120	12	.1179	732

Fig. 7



De samenhang tussen de gevonden punten in de figuur toont een sterk stijgend verloop van de proces-kosten bij uitbreiding van de dataset. Het lijkt zelfs niet uitgesloten dat hier geen sprake is van een kwadratisch verband, doch van een nog hoger exponentieel verband.

Een en ander zou een waarschuwing moeten inhouden aan het adres van de marktonderzoeker die dit geometrische clusterprincipe op grote schaal wil toepassen.

Het valt in dit verband aan te bevelen met een kleine dataset te beginnen en die, afhankelijk van de specifieke probleemstelling, daarna uit te bouwen in volgende clusterruns.

Het ligt in de bedoeling van de auteurs van dit artikel een vervolg-onderzoek te verrichten waarin aan groepenbeperking en variabelenkeuze grotere aandacht zal worden besteed dan tot nu toe het geval is geweest.

Deze werkwijze lijkt juist daarom relevant omdat in groepenbeperking en zinvolle variabelen-keuze de kostenbesparing kan zijn gelegen die deze analyse geschikt maakt voor algemeen commercieel gebruik.

7.0.0 Samenvatting en conclusie

Het door Jerrold Rubin en Herman Friedman ontwikkelde en geprogrammeerde geometrische klassificatiesysteem (programmabibliotheek IBM) is bij wijze van proef toegepast op enquêteresultaten. De ervaringen opgedaan met enkele proefnemingen kunnen als volgt worden samengevat.

Toekomstige gebruikers van dit programma doen er goed aan om zeer kostenbewust tewerk te gaan De kosten van het uitvoeren van de analyse stijgen progressief bij een toename van het aantal *waarnemingen* (objecten), dat aan analyse wordt onderworpen.

De door de auteurs opgedane praktische ervaring met het programma strekt zich uit over de volgende facetten van het klassificatiesysteem:

1. Toepassing van verschillende brekingscoëfficiënten

Deze proefnemingen leidden tot de conclusie, dat de als 'standaardoptie' aanbevolen keuze: $S^* = \bar{S}_{ij}$ meestal tot een zinvoller groepsindeling leidde dan andere opties, zelfs indien deze rond de eerder genoemde gemiddelde waarde werden gekozen ($S^* \gtrless \bar{S}_{ij}$).

2. Toepassing van een gegeven initiële partitie

Bij sommige vervolganalyses die op dezelfde dataset werden uitgevoerd, werd de bij een voorafgaande analyse gevonden partitie gebruikt als invoergegeven voor een volgende analyse.

In de programmaliteratuur werd bij gebruik van deze optie een kostenbesparing in het vooruitzicht gesteld.

In de praktijk is van een dergelijke kostenbesparing niets gebleken.

3. Stabiliteit van de door de analyse bereikte groepsstructuur

Op grond van de door Prof. van Tulder geformuleerde theoretische oorzaak voor mogelijke instabiliteit van het resultaat van de analyse werd een experiment opgezet. Dit experiment hield in een onderzoek van de stabiliteit van groepsstructuren, die achtereenvolgens werden gegenereerd in een steeds groeiende dataset.

De initiële dataset werd daartoe twee keer uitgebreid met nieuwe datasets van gelijke omvang.

De drie datasets bevatten dezelfde operatoren. De afzonderlijke sets ontstonden door random trekking uit een groter databestand en waren zodanig gekozen, dat de totaalscores per operator per set zoveel mogelijk identieke of gelijkvormige frequentieverdelingen hadden.

Drie hypothesen over kenmerken van een stabiel clusterresultaat werden achtereenvolgens getoetst op de resultaten van het groei-experiment.

Hieruit kwam als bevinding naar voren, dat geen van de datasets van 40, resp. 80 en 120 waarnemingen een voldoende omvang had om aan alle gestelde eisen van stabiliteit te voldoen.

Wèl werd de stabiliteit van de set van 120 waarnemingen groot genoeg geacht om ten aanzien van de voornaamste (= meest omvangrijke) groepen redelijk betrouwbare uitspraken te doen met betrekking tot hun meest opvallende klassificatiekenmerken.

In de analyses over 120 waarnemingen bleek het bovendien mogelijk het effect van een uitbreiding van de object-definitie (d.w.z. het toelaten van meer variabelen tot de berekening van de overeenkomstindex) nauwkeurig te analyseren.

Deze bevinding riep de vraag op of men de oorspronkelijke stabiliteitseisen niet onnodig streng had gesteld.

De hier beproefde analyse-methode schept een dilemma bij de toepassing, speciaal in die gevallen waarin men een probleemstelling vanuit een 'kritisch'-methodologisch standpunt wenst te benaderen. Gebleken is immers, dat een dataset toch tenminste 200–250 waarnemingen zal moeten omvatten, wil het resultaat zowel genuanceerde alsook betrouwbare conclusies mogelijk maken.

In dat geval dreigen echter tevens de kosten van de analyse zó hoog te worden, dat een toepassing voor commerciële doeleinden meestal niet verantwoord zal worden geacht.

De hoop is gevestigd op kostenbesparingen, die worden geëffectueerd bij drastische beperking van het voor de klassificatie toegelaten aantal groepen.

Gevreesd moet echter worden, dat zo'n beperking tot stand komt vooral ten nadele van de latent aanwezige kleine groepen.

Voor een kritische analyse, die juist veel nadruk legt op het belang van kleine groepen en hun kenmerken, zal deze eigenschap veelal een onoverkomelijk bezwaar blijken te zijn.

Het is onzes inziens van groot belang, dat er een vervolg op dit testprogramma komt. Zowel onze vooronderstelling over de kosten van de analyse als die over het effect van groepenbeperking op de kosten zijn van zodanig belang voor het bepalen van de toepassingsmogelijkheden, dat toetsing ervan dringend gewenst is.

Zolang dit niet is gebeurd kan slechts worden vastgesteld, dat de analyse praktisch bruikbare resultaten belooft, doch dat de wijze van verkrijging van die resultaten vooralsnog tamelijk kostbaar lijkt te zijn.

Summary

The application of 'statistical' techniques of cluster analysis on survey data often requires many transformations in the 'shape' of the variables used to define the objects of the survey (respondents as defined by scores on a set of variables).
Some transformations in fact distort the very nature of the variable. This seems to justify the use of geometrical clustering techniques as developed for purposes of numerical taxonomy.
This report deals with attempts to use a taxonomy system for grouping survey data. The program used has been developed for IBM by J. Rubin and H. Friedman. Several trial runs were made to test the implications of varying options in the program and to obtain indications with regard to the validity of the results.
The tantalizing conclusion of these explorations is that although the results are not unpromising even in the case of relatively small samples, the minimum sample size of 200-250 objects can only be processed at comparatively high cost. Further exploration will be needed to find ways of reducing the cost of the analysis.

Noten

* De samenstellers van dit artikel zijn grote dank verschuldigd aan de N.V. Inter-View te Amsterdam voor het beschikbaar stellen van het onderzoeksmateriaal en de financiën, alsmede aan de N.V. v/h Nederlandse Stichting voor Statistiek te Den Haag en het I.P.M. te Schiedam voor financiële bijdragen die voortzetting van het onderzoek mogelijk maakten. Tevens zijn zij veel dank verschuldigd aan prof. dr. J.J. van Tulder voor zijn waardevolle bijdrage aan de discussie en zijn suggesties voor aanvullend onderzoek.

1. Voor een handzaam overzicht van relaties tussen probleemstelling en keuze van het type analyse zij verwezen naar Jagdish N. Sheth: 'The Multivariate Revolution in Marketing Research', *Journal of Marketing* 35 : 1 (1971) pp. 13–19 en van dezelfde auteur: 'Multivariate Analysis in Marketing'; *Journal of Advertising Research* 10 : 1 (1970) pp. 29–40.
2. De in dit verslag gekozen symbolen wijken in sommige gevallen af van de door de auteurs van de programmaliteratuur gehanteerde symbolen. In de hierna volgende analyse van de resultaten moesten vele begrippen worden gebruikt (o.a. bij de beschrijving van toetsen op het resultaat) waarvoor symbolen gebruikelijk zijn, die identiek waren met die, welke

de auteurs voor aanduiding van geheel andere begrippen gebruikten. Wij bieden de lezers gaarne onze excuses aan voor het hierdoor veroorzaakte ongerief.

3. In de programmaliteratuur wordt terwille van de theoretische ontwikkeling van de stabiliteitsmaatstaf onderscheid gemaakt tussen de 'IN'groepsovereenkomst S_{ij} en de 'UIT'groepsovereenkomst t_{ij}.
4. Voor een beschrijving van de Cramèr associatiecoëfficiënt zie o.a. William L. Hays, *Statistics* (N. York 1963) pp. 604–606.
5. De reden van deze transformatie is gelegen in complicaties die ontstaan bij de berekening van φ' in subgroepen van zeer kleine steekproeven, zoals in ons experiment in bepaalde gevallen noodzakelijk werd.
   Het toelaten van alle voor de clusteranalyse onderscheiden niveaus in de berekening zou immers hebben geleid tot grote aantallen cellen met hooguit één waarneming, en een te frequente verwaarlozing van deze '1'-cellen ($\ln 1 = 0$) voor de berekening van de associatiecoëfficiënt. Is de kans op cellen ter grootte van één waarneming zeer klein, dan vervalt deze noodzaak tot transformatie en kan bijgevolg scherper getoetst worden.
6. Appendix II bevat een overzicht van de variabelen waarvan de operatoren in de clusteranalyse zijn opgenomen.

Appendix I

Literatuur

1. *A.H. Boon van Ostade*
   'Een klassificatiemethode voor psychologische data'.
   Dissertatie Nijmegen 1969.
2. *R.E. Bonner*
   'On Some Clustering Criteria'.
   IBM Journal of Research and Development. Vol 8. Jan. 1964, p. 22 t/m 32.
3. *H.P. Friedman and J. Rubin*
   'On Some Invariant Criteria for Grouping Data'.
   Journal of the American Statistical Association: 62, 1967. p. 1159 t/m 1178.
4. *D. Rogers and T. Tanimoto*
   'A Computer Program for Classifying Plants'.
   SCIENCE 132. p. 1115 t/m 1118. 1960.
5. *J.H.G. Segers en Ph.C. Stouthard*
   'Analyse door middel van opeenvolgende contrasterende groepen'.
   Tijdschrift Sociale Wetenschappen: 6:3 (1963) p. 221 t/m 241.

Appendix II

Overzicht van de in de analyse betrokken variabelen

Voor het stabiliteitsonderzoek werden de clusteranalyses op hetzelfde databestand uitgevoerd op basis van resp. 9 en 12 operatoren voor de berekening van de paarsgewijze match.

De variabelen, die door deze operatoren vertegenwoordigd werden waren de volgende:

nummer van de operator in de analyse met ... operatoren 9	12	*aantal niveaus van de operator*	*naam van de variabele in het onderzoek waaruit de gegevens afkomstig zijn*
1	1	9	Urbanisatiegraad van de woonplaats van de ondervraagde huisvrouw
2	2	8	Leeftijd van de ondervraagde huisvrouw
–	3	5	Abonnement op één of meer damesbladen
–	4	2	Abonnement op opinieblad
3	5	2	Abonnement op Consumentengids
4	6	4	Abonnement op één of meer nationale dagbladen
–	7	9	Gezinsgrootte
5	8	4	Gezinssamenstelling (aanwezigheid van en geslacht van kinderen)
6	9	5	Schaal Opinieleiderschap. Opgebouwd uit antwoorden op 4 vragen over opinieleiderschap over 4 interessegebieden*
7	10	2	Bezit bankrekening
8	11	2	Bezit koelkast
9	12	9	Gezinsinkomen

* Als de huisvrouw zichzelf beschreef als belangrijk informant op één interessegebied, dan kreeg zij een score 'één uit vier'; gaf zij aan belangrijk informant te zijn op twee gebieden, dan kreeg zij een score 'twee uit vier' enz.

Bibliografie multivariate analyse *

ANDERSON, T.W. *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley & Sons, 1958.

ATTNEAVE, F. Dimensions of Similarity. *American Journal of Psychology*, Vol. 63, pp. 516–556.

BANKS, S. Why People Buy Particular Brands, in R. Ferber and H. Wales (eds.), *Motivation and Behavior*. Homewood, Ill.: Richard D. Irwin, 1968.

BARBAN, A.M. AND W.F. GRUNBAUM. A Factor Analytic Study of Negro and White Responses to Advertising Stimuli. *Journal of Applied Psychology*, Vol. 49, pp. 274-279.

BENSON, P.H. Individual Exposure to Advertising and Changes in Brands Bought. *Journal of Advertising Research*, Vol. 7, No. 4, December 1967, pp. 27-32.

BIRDWELL, A.E. in L.G. Smith (ed.). *Reflections on Progress in Marketing*. Chicago: American Marketing Association, 1964.

BOX, G.E.P. A General Distribution Theory for a Class of Likelihood Criteria. *Biometrika*, Vol. 36, pp. 317-346.

BRODY, R.P. AND S.M. CUNNINGHAM. Personality Variables and the Consumer Decision Process. *Journal of Marketing Research*, Vol. 5, February 1968, pp. 50-57.

BURT, C. *The Factors of Mind*, New York: Macmillan and Company, 1941.

BURT, C. Factorial Analysis of Qualitative Data, *British Journal of Psychology*, Vol. 3, pp. 166-185.

BURT, C. Correlations between Persons. *British Journal of Psychology*, Vol. 38, pp. 59-96.

CAMPBELL, B.M. *The Existence and Determinants of Evoked Set in Brand Choice Behavior*, Unpublished Ph. D. Dissertation, Columbia University, 1968.

CATTELL, R.B. The Three Basic Factor Analytic Research Designs— Their Interrelationships and Derivatives. *Psychological Bulletin*, Vol. 49, pp. 499-520.

CATTELL, R.B. (ed.), *Handbook of Multivariate Experimental Psychology*. Chicago: Rand McNally & Co., 1966

CATTELL, R.B. (ed.), Guest Editorial: Multivariate Behavioral Research and the Integrative Challenge. *Multivariate Behavioral Research*, Vol. 1, January 1966, pp. 4-23.

CLARINGBOLD, P.J. Multivariate Quantal Analysis. *Journal of the Royal Statistical Society*, Series B, Vol. 20, pp. 398-405.

CLAYCAMP, H.J. Characteristics of Owners of Thrift Deposits in Commercial Banks and Savings and Loan Associations. *Journal of Marketing Research*, Vol. 2, May 1965, pp. 163-170.

* Literatuuroverzicht, behorende bij het artikel 'Multivariate Analysis in Marketing', van Jagdish N. Sheth, Journal of Advertising Research, Vol. 10, Nr. 1, February 1970. Overgenomen met toestemming van © Advertising Research Foundation, Inc.'

CLEVENGER, T., G.A. LAZIER AND M.L. CLARK. Measurement of Corporate Images by the Semantic Differential. *Journal of Marketing Research*, Vol. 2, February 1965, pp. 80-82.

COLLINS, G. On Methods: Factor Analysis. *Journal of Advertising Research*, Vol. 1, No. 3, September 1961, pp. 28-32.

COOLEY, W.W. AND P.R. LOHNES. *Multivariate Procedures for the Behavioral Sciences*. New York: John Wiley & Sons, 1962.

COOMBS, C.H. *A Theory of Data*. New York: John Wiley & Sons, 1964.

CRONBACH, L.J. AND G.C. GLESER. Assessing Similarity Between Profiles. *Psychological Bulletin*, Vol. 52, 1955, pp. 281-302.

DAY, G.S. *Buyer Attitudes and Brand Choice Behavior*. Unpublished Ph.D. Dissertation, Columbia University, 1967.

DUBOIS, P.H. *Multivariate Correlation Analysis*. New York: Harper and Brothers, 1957.

DUDEK, F.J. Relations Among Television Rating Devices. *Journal of Advertising Research*, Vol. 4, No. 3, September 1964, pp. 24-28.

DUDECK, F.J. AND E. THOMAS. Scaling Preferences for Television Shows. *Journal of Applied Psychology*, Vol. 48, 1964, pp. 257-260.

EASTON, A. Corporate Style Vs. Corporate Image, *Journal of Marketing Research*, Vol. 3, May 1966, pp. 168-174.

EHRENBERG, A.S.C. On Methods: The Factor Analytic Search for Program Types. *Journal of Advertising Research*. Vol. 8, No. 1, March 1968, pp. 55-70.

EKEBALD, F.A. AND S.F. STASCH. Criteria in Factor Analysis. *Journal of Advertising Research*, Vol. 7, No. 3, September 1967, pp. 48-57.

EVANS, F.B. Psychological and Objective Factors in the Prediction of Brand Choice: Ford versus Chevrolet. *Journal of Business*, Vol. 31, October 1959, pp. 340-369.

EZEKIEL, M. *Methods of Correlation Analysis*, New York: John Wiley & Sons, 1941.

FARLEY, J.U. Dimensions of Supermarket Choice Patterns. *Journal of Marketing Research*, Vol. 5, May 1968, pp. 206-208.

FARLEY, J.U. Why Does "Brand Loyalty" Vary Over Products? *Journal of Marketing Research*, Vol. 1, November 1964, pp. 9-14.

FISHER, R.A. *Statistical Methods for Research Workers*, New York: Haftner Publishing Co., 1958.

FISHER, R.A. The Use of Measurements in Taxonomic Problems. *Annals of Eugenics*, Vol. 7, 1936, pp. 179-184.

FRANK, R.E. AND P.E. GREEN. Numerical Taxonomy in Marketing Analysis: A Review Article. *Journal of Marketing Research*. Vol. 5, February 1968, pp. 83-93.

FRANK, R.E., W.F. MASSY, AND D.G. MORRISON. In Smith, L.G. (ed.). *Reflections on Progress in Marketing*. Chicago: American Marketing Association, 1965, pp. 312-323.

FRANK, R.E. AND W.F. MASSY. Innovation and Brand Choice: The Folger's Invasion. In Greyser (ed.) *Toward Scientific Marketing*. Chicago: American Marketing Association, 1964.

FRANK, R.E., W.F. MASSY, AND D.C. MORRISON. Bias in Multiple Discriminant Analysis. *Journal of Marketing Research*, Vol. 2, August 1965, pp. 250-258.

FRANK, R.E., A.A. KUEHNN, AND W.F. MASSY (eds.). *Quantitative Techniques in Marketing Analysis*. Homewood, Ill.: Richard D. Irwin, 1962.

GALES, K. Discriminant Functions of Socio-Economic Class. *Applied Statistics*, Vol. 6, June 1957, pp. 123-132.

GATTY, R. Multivariate Analysis for Marketing Research: An Evaluation. *Applied Statistics*, Vol. 15, November 1960, pp. 157-172.

GIRSCHICK, M.A. Principal Components. *Journal of the American Statistical Association*, Vol. 31, 1936, pp. 519ff.

GREEN, B.F. Latent Structure Analysis and Its Relation to Factor Analysis. *Journal of the American Statistical Association*, Vol. 47, March 1952, pp. 71-76.

GREEN, P.E., M.H. HALBERT AND P.S. ROBINSON. Canonical Analysis: An Exposition and Illustrative Example. *Journal of Marketing Research*, Vol. 3, February 1960, pp. 32-39.

GREEN, P.E., R.E. FRANK AND P.S. ROBINSON. Cluster Analysis in Test Market Selection. *Management Science*, Vol. 13, April 1967, pp. 387-400.

GRUBB, E.L. AND G. HUPP. Perception of Self, Generalized Stereotypes, and Brand Selection. *Journal of Marketing Research*, Vol. 5, February 1968, pp. 58-63.

GUTTMAN, L. The Principal Components of Scale Analysis. In S.A. Stouffer, et al. (eds.). *Measurement and Prediction*. Princeton, N.J.: Princeton University Press, 1950.

HARMAN, H. *Modern Factor Analysis*, 2nd Edition, Chicago: The University of Chicago Press, 1967.

HARPER, R. Factor Analysis as a Technique for Examining Complex Data on Foodstuffs. *Applied Statistics*, Vol. 5, March 1956, pp. 32-48.

HARVEY, J. What Makes a Best Seller. In R. Ferber and H. Wales (eds.). *Motivation and Behavior*, Homewood, Ill.: Richard D. Irwin, 1968.

HORST, P. *Factor Analysis of Data Matrices*, New York: Holt, Rinehart and Winston, 1965.

HORST, P. Relations Among m Sets of Measures. *Psychometrika*, Vol. 26, 1961, pp. 129-149.

HOTELLING, H. Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, Vol. 24, 1961, pp. 129-149.

HOTELLING, H. The Generalization of Student's Ratio. *Annals of Mathematical Statistics*, Vol. 2, 1931, pp. 360-378.

HOTELLING, H. Relations Between Two Sets of Variables. *Biometrika*, Vol. 28, 1936, pp. 321-377.

HOWARD, J.A. AND J.N. SHETH. *A Theory of Buyer Behavior*. New York: John Wiley & Sons, 1969.

ITO, R. Differential Attitudes of New Car Buyers. *Journal of Advertising Research*, Vol. 7, No. 1, March 1967, pp. 38-42.

JOHNSON, P.O. The Quantification of Qualitative Data in Discriminant Analysis. *Journal of the American Statistical Association*, Vol. 45, March 1950, pp. 65-70.

JOWETT, G.H. Factor Analysis, *Applied Statistics*, Vol. 7, 1958, pp. 114-125.

JOYCE, T. AND C. CHANNON. Classifying Market Survey Respondents. *Applied Statistics*, Vol. 15, November 1960, pp. 191-215.

KENDALL, M.G. *A Course in Multivariate Analysis*, London: Charles Griffin & Company, 1957.

KERNAN, J.B. Choice Criteria. Decision Behavior, and Personality. *Journal of Marketing Research*, Vol. 5, May 1968, pp. 155-165.

KING, C.W. The Innovator in the Fashion Adoption Process. In L.G. Smith (ed.), *Reflections on Progress in Marketing*. Chicago: American Marketing Association, 1964.

KING, W.R. Early Prediction of New Product Success. *Journal of Advertising Research*, Vol. 6, No. 2, June 1966, pp. 8-13.

KING, W.R. Marketing Expansion —A Statistical Analysis. *Management Science*, Vol. 9, July 1963, pp. 563-573.

KING, W.R. Performance Evaluation in Marketing Systems. *Management Science*, Vol. 10, July 1964, pp. 659-666.

KING, W.R. On Methods: Structural Analysis and Descriptive Discriminant Functions. *Journal of Advertising Research*, Vol. 7, No. 2, June 1967, pp. 39-43.

KIRSCH, A.D. AND S. BANKS. Program Types Defined by Factor Analysis. *Journal of Advertising Research*, Vol. 2, No. 3, September 1962, pp. 29-32.

LAWLEY, D.N. AND A.E. MAXWELL. *Factor Analysis as a Statistical Method*, London: Butterworth & Co., 1963.

LAZARSFELD, P.F. Latent Structure Analysis. In Koch (ed.). *Psychology: A Study of a Science*. New York: McGraw-Hill, 1959.

LAZARSFELD, P.F. The Logical and Mathematical Foundation of Latent Structure Analysis. In S. Stouffer, *et al.* (eds.). *Measurement and Prediction*. Princeton: Princeton University Press, 1950.

LUNN, J.A. New Developments in Consumer Behavior. In D. Pym (ed.). *Social Science and Modern Management*. London: Penguin Books, 1967.

MASSY, W.F. Principal Components Regression in Exploratory Statistical Research. *Journal of the American Statistical Association*. Vol. 60, March 1965, pp. 234-256.

MASSY, W.F. Television Ownership in 1950: Results of a Factor Analytic Study. In R.E. Frank, A.A. Kuehn and W.F. Massy (eds.). *Quantitative Techniques in Marketing Analysis*. Homewood, Ill.: pp. 440-460.

MASSY, W.F. Applying Factor Analyses to a Specific Marketing Problem. In S.A. Greysor (ed.), *Toward Scientific Marketing*. Chicago: American Marketing Association, 1964, pp. 291-307.

MASSY, W.F. On Methods: Discriminant Analysis of Audience Characteristics. *Journal of Advertising Research*, Vol. 5, No. 1, March 1965, pp. 39-45.

MCDONALD, R.P.A. General Approach to Nonlinear Factor Analysis. *Psychometrika*, Vol. 27, 1962, pp. 397-415.

MINDAK, W.A. Fitting the Semantic Differential to the Marketing Problem, *Journal of Marketing*. Vol. 25, April 1961, pp. 28-33.

MORRISON, D.F. *Multivariate Statistical Methods*, New York: McGraw-Hill, 1967.

MORRISON, D.G. Measurement Problems in Cluster Analysis. *Management Science*, Vol. 13, August 1967, pp. B-775—B-780.

MUKHERJEE, B.N. A Factor Analysis of Some Qualitative Attributes of Coffee. *Journal of Advertising Research*, Vol. 5, No. 1, March 1965, pp. 35-38.

MYERS, J.G. AND F. NICOSIA. On the Study of Consumer Typologies. *Journal of Marketing Research*, Vol. 5, May 1968, pp. 182-193.

NUNNALLY, J. C. The Analysis of Profile Data, *Psychological Bulletin*, Vol. 59, 1962, pp. 311-319.

OSGOOD, C.E., G.J. SUCI AND P.H. TANENBAUM. *The Measurement of Meaning*, Urbana: University of Illinois Press, 1957.

OSGOOD, C.E. AND G.J. SUCI. A Measure of Relation Determined by Both Mean Differences and Profile Information. *Psychological Bulletin*, Vol. 49, 1952, pp. 251-252.

PERRY, M. Discriminant Analysis of Relations Between Consumer's Attitude Behavior and Intention (forthcoming in *Journal of Advertising Research*).

PESSEMIER, E.A., P.C. BURGER, AND D.J. TIGERT. Can New Product Buyers Be Identified. *Journal of Marketing Research*, Vol. 4, November 1967, p.. 349-354.

PILGRIM, F.J. AND J.M. KAMEN. Pattern of Food Consumption and Preferences through Factor Analysis. *Journal of Marketing*, Vol. 24, 1959, pp. 68-72.

RAMOND, C.K. Factor Analysis: When to Use It. In A Shuchman (ed.). *Scientific Decision Making in Business*, New York: Holt, Rinehart, and Winston. Inc., 1963.

ROBERTSON, T.S. AND J.N. KENNEDY. Prediction of Consumer Innovators: Application of Multiple Discriminant Analysis. *Journal of Marketing Research*, Vol. 5, February 1968, pp. 64-69.

ROSS, J. Mean Performance and the Factor Analysis of Learning Data. *Psychometrika*, Vol. 29, 1964, pp. 67-73.

ROTHMAN, J. Formulation of an Index of Prosperity to Buy. *Journal of Marketing Research*, Vol. 1, May 1964.

RULON, P.J. Distinctions Between Discriminant Analysis and Regression Analysis and a

Geometric Interpretation of the Discriminant Function. *Harvard Educational Review*, Vol. 21, 1951, pp. 80-90.

SETHI, S.P. *An Investigation into the Mediating Effects of Socio-Psychological Variables Between Advertising Stimulus and Brand Loyalty*. Unpublished Ph.D. Dissertation, Columbia University, 1958.

SHETH, J.N. A Review of Buyer Behavior, *Management Science*, Vol. 13, August 1967, pp. B-718—B-756.

SHETH, J.N. *A Behavioral and Quantitative Investigation of Brand Loyalty*. Unpublished Ph.D. Dissertation, University of Pittsburgh, 1966.

SHETH, J.N. A Factor Analytic Model of Brand Loyalty, *Journal of Marketing Research* (Vol. 5, November 1968, pp. 395-408).

SHETH, J.N. Factor Analysis of Classificatory Data. August 1967.

SHETH, J.N. Estimating Parameters Using Factor Analysis. January 1968.

SHETH, J.N. Dimensions of Brand Preference for Durable Appliances. July, 1968.

SHETH, J.N. Multivariate Analysis in Marketing. *Journal of Advertising Research*, Vol. 10, No. 1, February 1970, pp. 29-39.

SHETH, J.N. AND W. RING. Correlates of General Attitudes. July 1968.

SOKAL, R.R. Numerical Taxonomy, *Scientific American*, Vol. 215, December 1966, pp. 106-116.

SOKAL, R.R. AND P.H.A. SNEATH. *Principles of Numerical Taxonomy*, San Francisco: Freeman & Co., 1963.

SPEARMAN, C. General Intelligence Objectively Determined and Measured. *American Journal of Psychology*, Vol. 15, 1904, pp. 201-293.

STEFFLRE, V. Market Structure Studies. Unpublished manuscript, July 1965.

STEPHENSON, W. *The Study of Behavior*. Chicago: University of Chicago Press, 1953.

STEPHENSON, W. Public Images of Public Utilities. *Journal of Advertising Research*, Vol. 3, No. 4, December 1963, pp. 34-39.

STOETZEL, J.A. Factor Analysis of the Liquor Preferences of French Consumers. *Journal of Advertising Research*, Vol. 1, No. 4, December 1960, pp. 7-11.

SWANSON, C.E. The Frequency Structure of Television and Magazines. *Journal of Advertising Research*, Vol. 7, No. 2, June 1967, pp. 8-14.

TORGERSON, W.S. *Theory and Methods of Scaling*, New York: John Wiley & Sons, 1958.

TRAVERS, R.M.W. The Use of a Discriminant Function in the Treatment of Psychological Group Differences. *Psychometrika*, Vol. 4, 1939, pp. 25ff.

TRYON, R.C. *Cluster Analysis*, Ann Arbor, Mich.: Edwards Brothers, 1933.

TRYON, R.C. General Dimensions of Individual Differences: Cluster Analysis versus Factor Analysis. *Educational and Psychological Measurement*, Vol. 18, 1958, pp. 477-495.

TRYON, R.C. AND D.E. BAILEY. The BC TRY Computer System of Cluster and Factor Analysis. *Multivariate Behavioral Research*, Vol. 1.

TUCKER, L.R. Determination of Parameters of a Functional Relation by Factor Analysis. *Psychometrika*, Vol. 23, 1958, pp. 19-23.

TWEDT, D.W. A Multiple Factor Analysis of Advertising Readership. *Journal of Applied Psychology*, Vol. 36, June 1952, pp. 207-215.

VINCENT, N.L. A Note on Stoetzel's Factor Analysis of Liquor Preferences. *Journal of Advertising Research*, Vol. 2, No. 1, March 1962, pp. 24-27.

WELLS, W.D. Backward Segmentation. In J. Arndt (ed.). *Insights into Consumer Behavior*, New York: Allyn and Bacon, 1968.

Literatuurbesprekingen

P.E. Green, and D.S. Tull, Research for Marketing Decisions, Englewood Cliffs, N.J., Second edition 1970, 644 blz.

Dit werk behoort tot de nog beperkte kategorie boeken, die op wetenschappelijk niveau een systematisch overzicht van marktonderzoektechnieken geven. Het heeft een grote belangstelling ontmoet, hetgeen blijkt uit de tweede druk die in 1970 op de eerste uitgave uit 1966 volgde.

De aanvullingen in de tweede druk betreffen onder meer een nieuw hoofdstuk over meerdimensionale schaaltechnieken en een aanmerkelijke uitbreiding van de beschouwingen over multivariate analyse. Ook zijn in tal van andere hoofdstukken wijzigingen doorgevoerd. Aan het einde is een hoofdstuk toegevoegd over toepassingen van de Bayesiaanse besluitvorming in het marktbeleid.

Het boek behandelt marktonderzoek als het systematisch en objektief zoeken en analyseren van informatie die van belang is voor de identifikatie en de oplossing van 'marketing'-vraagstukken. Om die reden wordt eerst een hoofdstuk gewijd aan 'Management Decisions', waarbij de besluitvorming onder onzekerheid een belangrijke plaats inneemt. Met name de Bayesiaanse besluitvorming wordt uitvoerig behandeld; de grote voorkeur voor deze benadering treft men op talrijke plaatsen in dit boek aan. De behandelde marktonderzoektechnieken zijn gegroepeerd rond 'information from respondents', 'information from experiments' en 'information from simulation'.

Vrijwel het grootste gedeelte van het boek is gewijd aan 'information from respondents'. Onderwerpen die onder meer aan de orde komen zijn de opbouw van een vragenlijst, het formuleren van vragen, het trekken van de steekproef. Aandacht wordt besteed aan het meten van attitudes – schaaltechnieken – en aan het motivatieonderzoek. Multivariate statistische technieken om uit deze 'information from respondents' samenhangen af te leiden, worden uitvoerig besproken, te weten regressie-analyse, canonische korrelatie, diskriminantanalyse, komponent-analyse, faktor- en clusteranalyse.

'Information from experiments' is vooral gericht op proefopzetten in het markt-

onderzoek en op de daarbij gebruikte statistische technieken, zoals variantieanalyse en faktoriële schema's.

In het onderdeel 'Information from simulation' komen onder meer aan de orde: Monte Carlo-technieken, het nabootsen van ondernemers- en konsumentengedrag en het heuristisch programmeren. Hierbij wordt vooral de betekenis van simulatie voor het marktonderzoek met voorbeelden toegelicht. Het voorlaatste hoofdstuk is gewijd aan voorspellingstechnieken, waarbij onder meer het voorspellen met behulp van trendextrapolatie, 'exponential smoothing' en regressieanalyse aan de orde komen. Het boek wordt afgesloten met een hoofdstuk waarin voorbeelden worden gegeven over de toepassing van Bayesiaanse besluittechnieken.

Een aantal kritische opmerkingen, die overigens in de schaduw staan van onze grote waardering voor dit boek zijn de volgende. In de beschouwing over de Bayesiaanse besluitvorming in hoofdstuk 2, p. 58 ware meer aandacht te besteden aan de verdeling van $P(z_i|s_j)$ voor s_j: toestand van de natuur en z_i: resultaat van het onderzoek. Hoewel een behandeling van het verschil tussen funktionele en kausale samenhang voor marktonderzoekers zeer belangrijk is, lijkt de hier gegeven beschouwing over kausaliteit niet vrij van gemeenplaatsen, bv. p. 77: 'Plato, Aristotle, St. Thomas Aquinas, Kant, Hume, Newton, Locke, Mill and Freud are but a few of those who have contributed to this inquiry'.

Het komt ons voor dat de klassieke procedure van steekproeftrekking wel weinig aandacht krijgt in vergelijking met de Bayesiaanse benadering: systematische, gestratificeerde, cluster- en getrapte steekproeven worden erg kort samengevat.

Het hoofdstuk over voorspellingstechnieken is bescheiden. De hier ontwikkelde tegenstelling tussen 'correlation techniques' en 'econometric techniques' is wazig.

Ondanks deze kritische opmerkingen achten wij dit boek een rijk, wetenschappelijk verantwoord en zeer aktueel overzicht van de thans beschikbare onderzoektechnieken in het marktonderzoek. Een aantal onderwerpen, zoals de multidimensionale schaaltechnieken en de Bayesiaanse besluitvorming zijn duidelijk stimulerende aanzetten tot vernieuwing in het marktonderzoek.

Mede door het groot aantal voorbeelden, literatuurverwijzingen en opgaven is het werk van Green en Tull een zeer geschikt leerboek, dat als naslagwerk kan dienen; het is een studieboek, geen boek met pasklare recepten.

Hoewel de statistische technieken eenvoudig worden behandeld lijkt ons een elementaire kennis van de wiskundige statistiek wenselijk voor een vruchtbare bestudering van dit boek.

M. T. G. Meulenberg

W. W. Cooley and P. R. Lohnes, Multivariate Data Analysis, New York, 1971, 364 blz.

In 1962 publiceerden Cooley & Lohnes hun eerste boek getiteld: 'Multivariate Procedures for the Behavorial Sciences'.

Thans wordt hetzelfde onderwerp door deze auteurs opnieuw, maar in een uitgebreide en in sommige opzichten verbeterde vorm behandeld.

De stijl is die van een helder en systematisch geschreven studie- en handboek. Terminologie en notatie zijn met zorg gekozen.

Het boek is gericht op het practische uitvoering van multivariate analyse, niet op de statistische theorie. De verschillende methoden worden in de inleiding gegroepeerd naar twee gezichtspunten: (a) men heeft één of meer verzamelingen variabelen, en (b) men heeft één of meer populaties. Zo ontstaan de vier groepen

Q1: Hoofdassen-analyse en factor-analyse;
Q2: Multivariate variantie-analyse en discriminant-analyse;
Q3: Multiple, canonische, en multiple-partiële correlatie;
Q4: Multivariate covariantie-analyse.

De volgorde van de hoofdstukken wijkt van deze indeling inzoverre af, dat begonnen wordt met de multiple correlatie, omdat deze op eenvoudige wijze de introductie mogelijk maakt van enkele veel-voorkomende reken-methodes uit de lineaire algebra. Daarna worden achtereenvolgens de onderwerpen uit Q1, Q2, Q3, Q4 besproken.

Het boek begint met een algemeen overzicht, een korte inleiding in het rekenen met matrices en vectoren, en een illustratie van het gebruik van de computer programmeertaal FORTRAN (de auteurs bevelen de lezer aan een korte FORTRAN-cursus te volgen of een inleidend FORTRAN-leerboek door te nemen ten einde multivariate conputer-programma's te kunnen lezen).

In hoofdstuk II worden de multivariate normale verdeling besproken, en de covariantie en correlatie-matrix.

Elk der volgende hoofdstukken bespreekt een multivariate analyse methode. Begonnen wordt met een uiteenzetting van het doel en de mathematische structuur van de analyse. Waar nodig wordt daarbij ingegaan op bepaalde rekenmethodes (zoals de oplossing van de algemene eigenwaarden vergelijking bij de berekening van canonische correlaties).

De analyse methode wordt daarna eerst geïllustreerd aan een vereenvoudigd kunstmttig voorbeeld, en vervolgens aan een voorbeeld dat aan een reël onderzoek is ontleend. Tenslotte zijn aan elk hoofdstuk Fortran-computer-programma's in extenso toegevoegd (hierover straks nog een opmerking).

Aan de methodelogie van multivariate analyse wordt niet uitvoerig aandacht besteed. In dit opzicht draagt de geboden stof sterk het karakter van een kookboek. Ook zijn de 'oefeningen' die aan elk hoofdstuk zijn toegevoegd, meer gericht op de

rekentechniek dan op de keuze en de evaluatie van een methode. Toch kunnen deze opgaven de lezer stimuleren om de uitkomsten van een analyse kritisch te bezien en zich af te vragen in hoeverre de uitkomsten zijn vraagstelling beantwoorden.

Keuze van methode en interpretatie der uitkomsten zijn uiteraard de belangrijkste fasen van een onderzoekproject. Voorwaarde voor juiste keuze en juiste interpretatie is dat men de implicaties van een methode onderkent. Men moet de analyse kunnen interpreteren in termen van empirische variabelen en empirische relaties. Een nauwkeurige kennis van het analyse-procedure is daarvoor noodzakelijk, en het boek van Cooley en Lohnes biedt deze basis in duidelijke en overzichtelijke vorm.

In totaal bestaan ca 60 pag. uit teksten van computerprogramma's. Het is ondoenlijk de juistheid van de programma's na te gaan, maar men mag aannemen dat zij zorgvuldig getest zijn. Sommige zijn opgezet als 'chanin'-programma's d.w.z. dat men achtereenvolgens verschillende programma's moet uitvoeren, om uit de begin-data het uiteindelijke gewenste resultaat te krijgen, en dat kan wel eens omslachtig zijn. Overigens ben ik niet overtuigd van het nut van volledige vermelding van deze programma's. De meeste rekencentra bezitten zulke programma's wel, of kunnen ze opvragen, en om ze uit te voeren zal men toch enige assitentie van een programmeur nodig hebben.

Liever zou ik hebben gezien dat de auteurs deze pagina's hadden besteed aan methodologische beschouwingen, of aan bijzonderheden, die zich in de praktijk van een onderzoek kunnen voordoen, maar die in handboeken soms moeilijk te vinden zijn.

Zo vond ik dat wel erg weinig aandacht is besteed aan de vraag of en hoe men comunaliteiten moet schatten bij factoranalyse. Geen woord over scheve rotatie van factoren. Bij de bespreking van canonische correlaties wijzen de auteurs er niet op dat de berekende canotische correlatie vooral in kleine steekproeven een ernstige overschatting kan zijn van de ware canonische correlatie. Storend vond ik dat slechts terloops verband wordt gelegd tussen discriminant-analyse en multivariate-variantie-analyse, en dat op pag. 13 een onderscheid daartussen wordt gemaakt dat nergens op slaat of zeer slecht is uitgedrukt. In hoofdstuk over factoriële discriminant-analyse (beter: factoriële multivariate variantie-analyse) is de gelegenheid gemist om een vergelijking te maken met univariate vaiantie-analyse met zg. 'repeated measures'. In dit hoofdstuk wordt bovendien op zeer onduidelijke wijze voor de significantie-toetsing verwezen naar een eerder hoofdstuk. Hetzelfde bezwaar geldt tegen een verwijzing (op pag. 227) naar Bartlett's approximatie-toets voor Wilks' Lambda (op p. 175). Dezelfde onduidelijkheid zat trouwens ook in het vroegere boek uit 1962.

Samenvattend: Een vooral op de berekeningsmethoden- en in dat opzicht duidelijk en systematisch geschreven – studieboek, dat zeker geschikt is als inleidingen naslagwerk voor wie naslagwerk voor wie althans de moeite wil nemen zich series

in deze materie te verdiepen. Kennis van de voornaamste begrippen uit de elementaire statistiek is wel vereist, – en men moet niet bang zijn voor formules. De besproken voorbeelden zijn voldoende illustratief voor de toepassingsmogelijkheden.

E. E. Ch. I. Roskam

Uit de marktonderzoekwereld

NEDERLAND

Examen Marktonderzoek 1971
Het eerstvolgende examen Marktonderzoek zal worden afgenomen op zaterdag 27 november 1971 (deel I en II) en op zaterdag 29 januari 1972 (deel III en IV). Het examen is uitsluitend schriftelijk.
De examencommissie is als volgt samengesteld: prof. dr. G. J. Aeyelts Averink, voorzitter; prof. dr. J. van Tulder, vice-voorzitter; R. Swier, lid en G. Mulder, secretaris. Gecommitteerde namens het Ministerie van Economische Zaken is prof. dr. ir. M. Th. G. Meulenberg.
De aanmelding voor dit examen sluit op 1 november 1971.
Examenprospectussen en inschrijfformulieren, alsmede alle andere gegevens inzake het examen zijn verkrijgbaar bij het secretariaat van de examencommissie Marktonderzoek, Frederik Hendriklaan 110 te Leiderdorp.

BELGIË

Bronnen voor marktonderzoek
De afdeling Marktverkenning van de Nederlandsche Kamer van Koophandel voor België en Luxemburg te Brussel heeft een lijst samengesteld, getiteld:

BRONNEN VOOR MARKTONDERZOEK IN BELGIË.

De lijst is verkrijgbaar à *f* 7,50 voor leden van de Nederlandsche Kamer van Koophandel voor België en Luxemburg en *f* 15,— voor niet-leden, bij het bijkantoor te 's-Gravenhage, Nassauplein 24.

Statistische marketinggids
Bij vele ondernemingen bestaat bij export een groot en veelal constant gebrek aan feitelijke gegevens over de markt. Men weet te weinig af van het land waarheen men

exporteert en van de markt of deelmarkt die men wil bewerken. Aan algemene voorlichting op dit terrein bestaat dan ook grote behoefte.
Het is om deze reden dat de Nederlandsche Kamer van Koophandel voor België en Luxemburg een Statistische Marketinggids voor België heeft samengesteld. Het rapport bevat circa 170 tabellen, verdeeld over 157 pagina's, met gedetailleerde informatie over bevolking, inkomen, industrie, handel en publiciteit. De teksten van de tabellen zijn eveneens vermeld in het Frans en in het Engels.
De kosten van het rapport bedragen *f* 35,— voor leden van de Kamer en *f* 50,— voor niet-leden.
Besteladressen: Koningsstraat 93, 1000 Brussel, en Nassauplein 24, 's-Gravenhage.

E.S.O.M.A.R.

'Translating advanced advertising theories into research reality'

There will be five main papers presented at the above Seminar. (Madrid, 3–6 November 1971).
They are 'A suggested explanation of the mechanisms of advertising' by Marcel Marc; 'Practical frameworks for advertising planning and research' by Mrs. Mary Tuck; 'A simulation model for the development and control of products, their packaging and advertising' by W.A.K. Frost; 'Advertising models in the practical application for decisions on advertising content and creative treatment and on the use of advertising media' by Arthur Juchems and the seminar will end with the paper by Dr. Timothy Joyce in which he will comment on all of papers previously presented.
All papers and discussions will be in English.
Many people have already registered for the Seminar and there are only a limited number of places left. If you would like to participate please send in your registration as soon as possible to: E.S.O.M.A.R. Central Secretariat, Raadhuisstraat 15, Amsterdam.

Uit de buitenlandse tijdschriften

Journal of Marketing Research
Vol. 8, Nr. 1, February 1971

Market segmentation: a strategic management tool
Richard M. Johnson
Repetition in Media models: a laboratory technique
Michael L. Ray and Alan G. Sawyer
Trade area boundaries: some issues in theory and methodology
Louis P. Bucklin
Modeling store choice behavior
David A. Aaker and J. Morgan Jones
A system construct for evaluating retail market locations
L.A. White and J.B. Ellis
Television show preference: application of a choice model
Donald R. Lehmann
The application of attitude immunization techniques in marketing
Stewart W. Bither, Ira J. Dolich and Elaine B. Nell
Evolutionary Model building
Glen L. Urban and Richard Karash
A multivariate analysis of personality and product use
David L. Sparks and W.T. Tucker
Multidimensional scaling and individual differences
Paul E. Green and Vithala R. Rao
Models for responses of customers with a varying rate
Tarow Indow
Stability of appliance brand awareness
George S. Day and Robert W. Pratt, Jr.

Vol. 8, Nr. 2, May 1971

Market measurement and planning with a simultaneous-equation model
Randall L. Schultz
Market measures from salesmen: a multidimensional scaling approach
Ronald E. Turner
World-of-mouth communication in the industrial adoption process
John A. Martilla
Identifying the convenience-oriented consumer
W. Thomas Anderson, Jr.
Identification of determinant attributes: a comparison of methods
Mark I. Alpert
Multivariate analysis of differences in buyer decision time
Joseph W. Newman and Richard Staelin
An industrial pricing decision using Bayesian multivariate analysis
Donald G. Frederick
Television program types
Ronald E. Frank, James C. Becknell and James D. Clokey
Empirical studies on individual response to exposure patterns
Leonard M. Lodish
What big agency men in Europe think of copy testing methods
Harper W. Boyd, Jr. and Michael L. Ray
A computer simulation model for forecasting catalog sales
George C. Michael

Vol. 8, Nr. 3, August 1971

Stages of consumer dicision making
Terrence O'Brien
Testing cognitive models through computer-controlled experiments
G. David Hughes and Jose L. Guerrero
Personality variables and cigarette brand choice
Joseph N. Fry
Measurement of husband-wife influence in consumer purchase decisions
Harry L. Davis
Generalized change agents and innovativeness
John O. Summers
Clusters of consumer interests and opionion leaders' spheres of influence
David B. Montgomery and Alvin J. Silk

Five dimensions of the industrial adoption process
Urban B. Ozanne and Gilbert A. Churchill, Jr.
An experimental study of industrial buyer behavior
Richard N. Cardozo and James W. Cagley
The use of GERT in the planning and control of marketing research
A. Coskun Samli and Carl Bellas
Using cluster analysis to improve marketing experiments
Georges S. Day and Roger M. Heeler
Comparative cluster for world markets
S. Prakash Sethi
Conjoint measurements for quantifying judgmental data
Paul E. Green and Vithala R. Rao

Journal of the Market Research Society
Vol. 13, Nr. 2, April 1971

Cluster analysis – a new tool in electricity
J.G. Boggis and I. Held
Survey data in Courts of Law
Peter Menneer
Preference analysis: evaluation of promotional activity in consumer markets
S.T. Buck and T.St.G. Jephcott
Consumer attitudes and brand usage: comments
Mary Tuck
Consumer attitudes and brand usage – two answers
M. Bird and A.S.C. Ehrenberg
1984 and the market research counter-revolution
C. Greenhalgh

Vol. 13, Nr. 3, July 1971

Publicity on social issues: the need for research
Tom Corlett and Michael Barnes
Market segmentation-the realities of buyer behaviour
Martin Collins
Experimentation as a tool for the retailer
Yoram Wind, Susan P. Douglas and Aaron Ascoli
The world's simplest test of significance
Jack Field

Journal of Advertising Research
Vol. 11, nr. 1, February 1971

Brain wave measures of media involvement
Herbert E. Krugman
On advertising wear out
Valentine Appel
Frequency effects revisited
Michael L. Ray, Alan G. Sawyer and Edward C. Strong
Assessing TV pretest audiences
John J. Wheatley
Attitudes and media exposure
Charles W. King and John O. Summers
How program environment affects TV commercials
John R. Kennedy
Perceptual mapping to reposition brands
Henri Assael

Vol. 11, nr. 2, April 1971

Psychographics for market segmentation
Ruth Ziff
Reactions to blacks in newspaper ads
B. Stuart Tolley and John J. Goett
TV commercial recall correlates
Kevin J. Clancy and David M. Kweskin
A new way to find opinion leaders
James S. Fenton and Thomas R. Leggett
Dagmar revisited – eight years later
Donald C. Marschner
Responses to parts of TV commercials
Joseph R. Murphy

Medewerkers

E.J. Bijnen (30) studeerde sociologie aan de Katholieke Hogeschool te Tilburg. Hij promoveerde in 1969 aan dezelfde hogeschool. Sinds 1963 is hij als wetenschappelijk medewerker verbonden aan de Katholieke Hogeschool voor het vak Statistiek en methoden van sociologisch onderzoek.

L.C.A. Corsten studeerde aan de Landbouwhogeschool en aan de University of North Carolina. Hij was achtereenvolgens medewerker bij de afdeling wiskunde van de Landbouwhogeschool, bij het Instituut voor Rassenonderzoek voor Landbouwgewassen, en bij het Natuurkundig Laboratorium der N.V. Philips. Sinds 1963 is hij hoogleraar in de Wiskundige statistiek aan de Landbouwhogeschool. Hij vervulde onlangs een gasthoogleraarschap aan de Hebrew University te Jerusalem.

Tim van Kooten (26) studeerde economische sociologie aan de Katholieke Hogeschool Tilburg. Na zijn studie werkte hij korte tijd bij het reclamebureau Van Maanen, Ogilvy & Mather N.V. te Amsterdam, waarna hij in juni 1971 in dienst kwam bij N.V. v/h Nederlandse Stichting voor Statistiek, als hoofd van de afdeling industrieel marktonderzoek.

Gerrit Jan Randoe (41) studeerde economische sociologie aan de Universiteit van Amsterdam. Daarna was hij verbonden aan reclamebureau Van Maanen, Ogilvy & Mather N.V., eerst als marktonderzoeker, later ook als marketing-adviseur. Sinds april 1971 is hij adjunkt-directeur van de N.V. v/h Nederlandse Stichting voor Statistiek in Den Haag alwaar hij o.m. de supervisie heeft over projecten van fundamenteel onderzoek.

E.E.Ch.I. Roskam studeerde van 1950–1958 psychologie te Leiden. In 1966 werd hij medewerker van de vakgroep mathematische psychologie aan het Psychologisch laboratorium van de Katholieke Universiteit te Nijmegen. Hij promoveerde in 1968 op een proefschrift over nonmetrische schaalmethoden. Sinds het begin van dit jaar is hij gewoon hoogleraar – op persoonlijke titel – in de Mathematische psychologie te Nijmegen.