Patrick Van Kenhove *

Clusteranalyse en factoriële correspondentie-analyse in marktsegmentatieonderzoek: een praktische illustratie

In deze bijdrage wordt, aan de hand van een voorbeeld uit de praktijk, geïllustreerd hoe twee belangrijke multivariate analysetechnieken - clusteranalyse en factoriële correspondentieanalyse - aangewend kunnen worden in marktsegmentatieonderzoek. Alle te nemen stappen worden eerst kort theoretisch besproken en vervolgens toegepast. Veel aandacht wordt besteed aan de mogelijke valkuilen.

Inleiding

In deze tekst wordt eerst kort ingegegaan op de omschrijving van de technieken clusteranalyse en factoriële correspondentieanalyse (FCA). Vervolgens wordt aan de hand van een concreet voorbeeld aangetoond hoe beide technieken op zich en hun combinatie nuttig kunnen zijn voor marktsegmentering. De verschillende stappen in een clusteranalyse en een correspondentieanalyse worden doorlopen. Op dat moment wordt ook kort ingegaan op een aantal kenmerken en op de werking van beide technieken. Veel aandacht wordt besteed aan de interpretatiemogelijkheden van de output.

* Universiteit Gent (Faculteit Economische en Toegepaste Economische Wetenschappen en De Vlerick School voor Management). De auteur wenst drie anonieme referees te danken voor hun zeer waardevolle opmerkingen.

Economisch en Sociaal Tijdschrift, 1996/2, blz. 281-309

1. Clusteranalyse en factoriële correspondentieanalyse: afbakening, omschrijving en doel

De doelstelling van clusteranalyse kan als volgt omschreven worden: gegeven een steekproef van n individuen of objecten, elk gemeten op p variabelen, ontwikkel een classificatiesysteem om die n individuen of objecten te groeperen in g klassen, met g kleiner dan n (Everitt, 1974, blz. 1).

Het doel van clusteranalyse is gevallen (individuen, produkten, merken, stimulusmateriaal enz.) in groepen te sorteren, zodanig dat een hoge mate van gelijkenis bestaat tussen gevallen in dezelfde groep en een lage mate van overeenkomst tussen gevallen die behoren tot verschillende groepen (Wishart, 1987, blz. IV).

Clusteranalyse is een exploratieve methode om classificatieproblemen op te lossen. De techniek wordt gebruikt om een mogelijk aanwezige, doch niet gekende structuur in de gegevens bloot te leggen. Het is dus een hypothesegenererend instrument en niet hypothesetoetsend. De techniek is ontstaan in de biologie, maar wordt vandaag gebruikt in alle wetenschappelijke disciplines waar classificatie aan de orde is.

Clusteranalyse wordt in het marktonderzoek hoofdzakelijk aangewend voor datareductie (Punj en Stewart, 1983, blz. 134). Een groot aantal individuen of produkten of stimuli wordt samengevoegd tot een beperkt aantal intern homogene en extern heterogene groepen, die clusters of segmenten genoemd worden. De toepassingen in de marketing zijn vooral marktsegmentering en marktafbakening. Bij marktsegmentering wordt de techniek gebruikt om individuen te groeperen met gelijke behoeftenpatronen, gelijke karakteristieken of gelijke reacties op stimuli uit de marketing mix. Bij marktafbakening wordt clusteranalyse gebruikt om produkten of merken te groeperen die tot eenzelfde markt behoren en met elkaar concurreren.

Clusteranalyse is een verzamelnaam voor een bijna ontelbaar aantal technieken en procedures, wat de uiteindelijke keuze van de - voor het specifieke probleem - relevante procedure niet vergemakkelijkt.

Factoriële correspondentieanalyse (FCA) is een vorm van "multidimensional scaling" (MDS). MDS is een verzameling van mathematische

technieken om de onderliggende verborgen structuur in een datamatrix te visualiseren.

MDS is een verzamelnaam voor diverse technieken die een aantal gemeenschappelijke kenmerken bezitten. Tussen een of meer sets van objecten wordt een associatie berekend. Deze associatie geeft aan hoe similair of hoe dissimilair de twee vergeleken sets van objecten zijn. De output is één (of meer) ruimtelijke voorstelling(en) met daarin een puntenconfiguratie of een vectorconfiguratie of beide. Elk punt of elke vector komt overeen met een van de objecten. In de ruimtelijke voorstelling (map) worden de punten op een zodanige manier gevisualiseerd dat de objecten, naarmate ze meer similair (of minder dissimilair) in de datamatrix zijn, dichter bij elkaar liggen. De map vormt een al dan niet geslaagde visuele representatie van de verborgen structuur in de data (Kruskal en Wish, 1978, blz. 6).

Factoriële correspondentieanalyse is een vorm van MDS. Slechts sinds de tweede helft van de jaren tachtig kent de techniek een zekere populariteit in het marktonderzoek. Ze was in Frankrijk o.l.v. Benzécri reeds langer populair onder de naam "analyse des correspondances". In de Angelsaksische literatuur worden soortgelijke methoden gebruikt onder verschillende namen: dual scaling, optimal scaling, bi-plots, principalecomponentenanalyse voor kwalitatieve data enz. In essentie komen al deze methoden neer op de mathematische ontbinding van een matrix in zijn basisstructuur (matrixdecompositie), die de onderliggende dimensies voor rij- en kolomkenmerken bevat, alsook de gewichten die men aan deze dimensies kan toekennen (Hoffman en Franke, 1986; Carroll, Green en Schaffer, 1986, 1987; Carroll en Green, 1988; Benzécri, 1969). Het gebruik van verschillende benamingen had tot gevolg dat de methode in het verleden een tamelijk obscuur karakter had.[1]

Het doel van FCA bestaat erin de m rijen en n kolommen uit een contingentietabel visueel weer te geven als een set van m rij- en n kolompunten, in een ruimte met zo weinig mogelijk (liefst twee) dimensies. Een contingentietabel is in feite niets meer dan een eenvoudige kruistabel.

1 Met dank aan een anonieme referee voor deze aanvullingen.

2. Clusteranalyse en FCA als technieken voor marktsegmentatie: een uitgewerkt voorbeeld

Het uitvoeren van een clusteranalyse en een correspondentieanalyse vergt het doorlopen van een aantal stappen (De Pelsmacker en Van Kenhove, 1994). De relevante stappen worden, zoals eerder gesteld, geïllustreerd aan de hand van een concreet voorbeeld. Elke stap heeft een eigen problematiek. Om die op te lossen moet in elke stap een keuze worden gemaakt uit een waaier van procedures. Hier worden enkel die procedures behandeld die het meest gepast zijn voor de problematiek van het voorbeeld. Telkens zal wel summier verwezen worden naar de andere mogelijke procedures. Voor een uitvoerige behandeling en vergelijking van al deze procedures wordt naar de gespecialiseerde literatuur verwezen (Wishart, 1987; Everitt, 1974).

3. Probleemstelling

Het betreft hier het probleem van een Belgische fabrikant van modieuze dameskleding, die zijn produkten verkoopt via de betere modeboetieks (SLM, 1988). Vanuit uitgebreid onderzoek weet de opdrachtgever wie zijn finale consumente is en wat haar behoeften en haar voorkeurkleding zijn. Via het hier uitgevoerde onderzoek wenst de opdrachtgever te weten in welke wijken (winkelcentra) van Brussel hij distributiepunten moet zoeken om zijn collectie aan te bieden. Het hier uitgevoerde onderzoek zal dus een antwoord moeten bieden op de volgende vragen:

- waar koopt de modegevoelige consumente in Brussel haar kleding?
- wat is het beeld en het profiel van elk winkelcentrum in Brussel bij die modegevoelige consumente?

Uit vroeger uitgevoerd onderzoek (ITCB, 1988) is bekend dat het direct oplossen van beide vragen een te gesimplifieerd beeld van de werkelijkheid zou geven. Verschillende modegevoelige consumenten kunnen de verschillende winkelcentra om verschillende redenen op een andere manier gaan percipiëren en prefereren. Het wordt daarom wenselijk geacht de modegevoelige consumenten te segmenteren met als criterium de relatie die zij ontwikkelen met een of meer winkelcentra. Er zal dus gezocht worden naar een waaier van variabelen die ons in staat stellen zo goed mogelijk een beeld en een verklaring te geven voor de relatie van de consumente met haar winkelcentra.

4. Selectie, meetniveau en vergelijkbaarheid van attributen

A. Selectie van de juiste variabelen (attributen)

Het uitgangspunt van clusteranalyse is de classificatie van n objecten (individuen) op basis van hun antwoorden op p variabelen (attributen). De selectie van de p attributen is cruciaal en zeer delicaat (Milligan, 1980, blz. 325).

Een belangrijke fout is het niet opnemen in de studie van alle relevante attributen voor het onderzoek. Dat kan het gevolg zijn van een onvolledig uitgevoerde exploratieve fase van de gegevensverzameling. Aldus kunnen belangrijke dimensies ontbreken en zal de clusteroplossing wellicht statistisch goed maar inhoudelijk suboptimaal zijn. Het is immers onmogelijk om een dergelijke fout te herstellen. Een voorbeeld hiervan is een recent uitgevoerde clusteranalyse in de koekjesmarkt (SLM, 1992). Een heel belangrijke verklarende dimensie bij de vraag waarom bepaalde mensen bepaalde koekjes kopen, is de gebruikssituatie. Met andere woorden: dient het koekje om thuis bij de tv op te eten, om met de kinderen mee te geven naar school, of wordt het gepresenteerd aan mogelijke visite, enz. Indien de onderzoekers waren vergeten deze dimensie op te nemen in de vragenlijsten, zou de clusteranalyse wellicht technisch goede maar inhoudelijk zeer suboptimale clusters opgeleverd hebben.

De onderzoeker dient er dus voor te zorgen dat het meetinstrument inhoudvalide is. Dat kan door een minutieus uitgevoerd vooronderzoek via "desk research" en kwalitatieve technieken.

Toegepast op het voorbeeld van dit artikel wordt in eerste instantie exploratief kwalitatief onderzoek uitgevoerd om de relevante gegevenssets te verzamelen. Om de relevante attributen te verzamelen, dienen we eerst de literatuur en de bestaande marktonderzoeken over het onderwerp na te pluizen. Vervolgens wordt met een aantal vrouwen uit de doelgroep een individueel gesprek gevoerd via een diepte-interview via projectieve technieken. Er is geopteerd voor deze techniek omwille van de diepgang die via dergelijke ondervraging wordt bereikt. Er is niet geopteerd voor groepsdiscussies wegens het persoonlijke karakter van het onderwerp en de mogelijkheid van sociaal wenselijke antwoorden (De Pelsmacker en Van Kenhove, 1994). In totaal worden 18 consu-

menten ondervraagd. Deze consumenten worden geselecteerd aan de hand van een contactblad met de volgende criteria:
- de consumente is een vrouw tussen 18 en 60 jaar;
- zij is een trouwe bezoekster (minstens 1 keer per maand) van een Brussels winkelcentrum;
- zij koopt kleding in dit winkelcentrum aan.

Er wordt gestreefd naar representativiteit wat de winkelcentra betreft. Elke keer wordt aan de respondenten gevraagd om de relevante lijst van winkelcentra op te stellen en aan te vullen.

Op basis van de resultaten uit de exploratieve fase (de hierboven vermelde "desk research" en de diepte-interviews) worden vervolgens de lijsten met de relevante karakteristieken opgesteld. Die lijsten bevatten tien winkelcentra en drie lijsten (respectievelijk 23, 18 en 23 uitspraken uit de diepte-interviews). Een (gedeeltelijk) overzicht van de lijsten is opgenomen in tabel 1 (zie verder).

B. Meetniveau van de attributen

Hoe worden de variabelen gemeten? De meeste softwarepakketten voor clusteranalyse maken een analyse mogelijk op continue en binaire variabelen.

Een binair attribuut is een attribuut dat ofwel aanwezig ofwel afwezig is voor een individu. Het kan slechts twee waarden hebben: 0 of 1. Gegevens bekomen via de "pick any"-methode zijn hiervan een voorbeeld. Hierbij dient een respondent uit een lijst van bijv. 24 attributen er 5 te kiezen die het best zijn/haar mening weergeven.

Een continue variabele, daarentegen, kan veel meer waarden aannemen dan een binaire. De respondent geeft een score op een schaal met meer nuances dan 0 en 1, de afstanden tussen de scores zijn in principe gelijk. Voorbeelden hiervan zijn gegevens verkregen via een semantische differentiaal, een Likert- of een stapelschaal. Uit een lijst van bijv. 24 attributen dient de respondent nu voor alle 24 een score te geven van bijv. 1 tot 7.[2]

2 Een referee merkt terecht op dat Likert-schalen, semantische differentialen en stapelschalen in feite geen echte continue variabelen zijn. Het zijn pseudo-continue variabelen. Respondenten gebruiken de schaal niet op dezelfde wijze. Sommigen scoren eerder aan de extreme zijde van de schaal, terwijl anderen dat niet doen. Als gevolg daarvan heeft de schaal een verschillende betekenis voor verschillende respondenten.

Beide methoden hebben voor- en nadelen. De "pick any"-methode gaat snel. De respondent hoeft niet alle variabelen te scoren. De schaalmethoden (semantische differentiaal of likertschaal of stapelschaal) hebben als belangrijkste nadeel dat bij een groot aantal variabelen de vragenlijst zeer lang wordt, waardoor de respondent kan afhaken. Het grote voordeel van deze methode is echter wel dat de verkregen gegevens aan statistisch meer en krachtiger procedures onderworpen kunnen worden dan de "pick any"-data. De keuze tussen beide zal dus in hoofzaak afhangen van de lengte van de vragenlijst en het soort statistische analyses dat de onderzoeker wil uitvoeren (Van Kenhove, 1989, blz. 290).

In het voorbeeld is de "pick any"-methode gebruikt. Hoe is dat concreet gebeurd? Een vragenlijst wordt via persoonlijk interview afgenomen van respondenten die aan de eerder gestelde criteria (zie bij de exploratieve fase) beantwoorden. De criteria worden als volgt geoperationaliseerd:
- Opdat de ondervraagde persoon zeker tot de doelgroep zou behoren, moet zij merkkleding kopen die tot hetzelfde concurrentieel veld behoort als dat van de opdrachtgever. Met concurrentieel veld wordt hier bedoeld die merken die verkrijgbaar zijn in de distributiepunten waar ook het merk van de opdrachtgever te koop is. Een lijst van 41 merken wordt daartoe opgesteld.
- Wanneer de aangesproken persoon voldoet aan het eerder gestelde leeftijdscriterium en tijdens een bepaalde tijdsperiode een van de merken heeft gekocht, wordt zij uitgenodigd voor een persoonlijk interview.

De respondenten worden proportioneel gerekruteerd in elk van de tien winkelcentra. De effectieve steekproef bestaat uit 302 respondenten.

Aan de hand van de vragenlijst worden eerst een aantal gedrags- en identificatievragen gesteld. Vervolgens worden de drie lijsten van kenmerken doorlopen voor het meest geprefereerde winkelcentrum. De gevolgde methode is de "pick any"-gegevensverzameling. Telkens wordt gevraagd om uit elk van de lijsten met attributen vijf woordjes of uitspraken te selecteren die het meest van toepassing zijn op het meest geprefereerde winkelcentrum. De "pick any"-methode is gekozen omwille van de snelheid. De enquêtes mochten niet te lang duren. Uit het vooronderzoek was immers gebleken dat de respondenten uit de doelgroep niet langer dan vijf tot tien minuten tijd wilden vrijmaken.

Op basis van deze ondervraging wordt nu een datamatrix opgesteld, bestaande uit 302 rijen (de respondenten) en 64 attributen (de bovengenoemde karakteristieken). Deze datamatrix bevat enkel de getallen 1 en 0. Een 1 staat voor selectie van een bepaald attribuut door de respondente, 0 staat voor niet-selectie.

De doelstelling bestaat erin consumentensegmenten te vinden die intern zo homogeen mogelijk en extern zo heterogeen mogelijk zijn qua relatie tot hun geprefereerde winkelcentra. Daartoe zal een clusteranalyse worden uitgevoerd op deze datamatrix, met als doel groepen van respondenten te vinden met een zo gelijklopend mogelijk antwoord op de 64 attributen. Een cluster is dus een groep respondenten met gelijksoortige verlangens wat winkelcentra betreft.

C. Vergelijkbaarheid van de attributen

Attributen zijn vergelijkbaar wanneer ze in dezelfde meeteenheid gesteld zijn. In het voorbeeld is aan deze voorwaarde voldaan. Op elke vraag is enkel een 1 of een 0 een geldig antwoord. Attributen zijn echter niet altijd vergelijkbaar. Een eenvoudig voorbeeld kan dit verduidelijken. Van drie respondenten meet men twee attributen, nl. de leeftijd (gemeten in jaren) en het inkomen (gemeten in franken).

Respondent	Leeftijd	Inkomen
1	34	900.000
2	26	950.000
3	36	1.600.000

De verschillen tussen de respondenten zijn:
Respondent 1 versus 2: (34 - 26) + (950.000 - 900.000) = 50.008
Respondent 1 versus 3: (36 - 34) + (1.600.000 - 900.000) = 700.002
Respondent 2 versus 3: (36 - 26) + (1.600.000 - 950.000) = 650.010

Het is duidelijk dat deze vergelijkingen zinloos zijn. De variabelen leeftijd en inkomen hebben een dusdanig verschillend gewicht in de analyse dat de variabele leeftijd geen enkele rol speelt. Men zal de variabelen vergelijkbaar maken door ze eenzelfde gewicht te geven. Dat is mogelijk door ze te standaardiseren: elke variabele krijgt een gemiddelde 0 en een standaardafwijking 1, zodat de waarden van twee in totaal verschillende meeteenheden gemeten variabelen perfect vergelijkbaar worden.

Bij de vergelijkbaarheid van de attributen kan zich een tweede probleem voordoen. Dikwijls is er een onevenwichtige verdeling van de attributen over onderliggende dimensies (constructen) gemeten via die attributen (Punj en Stewart, 1983, blz. 144). Veronderstel twee belangrijke dimensies in de data en tien attributen die relevant zijn voor het probleem. Wanneer bijvoorbeeld de eerste 9 attributen dimensie 1 meten en slechts 1 attribuut (het tiende) de tweede dimensie, dan ontstaat een duidelijk onevenwicht. In het geval dat de eerste 9 attributen ongeveer perfecte substituten zijn voor elkaar, krijgt in werkelijkheid 1 attribuut gewicht 9 en het andere attribuut gewicht 1.

Bij continue variabelen kan dit probleem worden opgelost via principale-componentenanalyse. De oorspronkelijke variabelen worden gereduceerd tot een kleinere set van factoren of onderliggende dimensies. De factorscore (de score van elke respondent op elke factor) wordt vervolgens gebruikt als input voor clusteranalyse. Bij binaire variabelen kan de analyse van de intercorrelatie (gemeten via bijv. tetrachorische correlatiecoëfficiënten; zie Long, 1983) een oplossing bieden. Bij een groot aantal binair gescoorde attributen zal het moeilijk zijn om op basis van visuele inspectie van de intercorrelaties te concluderen of er sprake is van een evenwichtige spreiding van de attributen op de dimensies. Een mogelijke oplossing die door een referee werd gesuggereerd is eerst correspondentieanalyse (FCA, zie verder) op de data toe te passen en de respondenten vervolgens op de scores op de FCA-dimensies te clusteren. In het voorbeeld wees de correlatie tussen de 64 attributen op een vrij evenwichtige spreiding van de attributen op mogelijk onderliggende dimensies. Er waren slechts enkele attributen met een intercorrelatie (in absolute termen) hoger dan 0,7 ($0<r<1$).

5. Keuze van een clusteralgoritme

Eerder is reeds vermeld dat clusteranalyse een verzamelnaam is voor een zeer groot aantal clusteralgoritmes. De vraag naar een passend clusteralgoritme dringt zich dus op. De resultaten kunnen immers zeer sterk verschillen naar gelang van het gekozen algoritme. De keuze van het juiste clusteralgoritme is dus van zeer groot belang. Het is wellicht

de belangrijkste beslissing die de onderzoeker dient te nemen in de hele clusteranalyse. Uit onderzoeken (o.a. Milligan, 1980; Wishart, 1987) komen nochtans een aantal constanten naar voren die toepasbaar zijn op marktsegmentatie.

Alle hiërarchische samenvoegende methoden van clusteranalyse leiden meestal tot niet-optimale resultaten. Hiërarchische samenvoegende methoden vertrekken van n individuen die elk één cluster voorstellen. Vervolgens worden de twee clusters met de hoogste similariteit of kleinste dissimilariteit of afstand samengevoegd. De procedure kan meestal voortgezet worden tot alle individuen in 1 cluster zijn opgenomen. De resultaten worden meestal visueel voorgesteld in een dendrogram, dat de opeenvolgende samenvoegingen visueel voorstelt (Everitt, 1974, blz. 8).

Opeenvolgende samenvoegingen van individuen zijn definitief. Dit betekent dat een individu, eenmaal toegewezen aan een cluster, niet meer van cluster kan veranderen. De zwakte hiervan is dat foutieve klasseringen in het begin van het proces onherstelbaar zijn. Vroegere foutieve klasseringen kunnen bijvoorbeeld het gevolg zijn van "outliers", dat zijn individuen die een zeer apart antwoordpatroon hebben en in feite tot geen enkele cluster behoren. Hoe groter de dataset en hoe vroeger in het proces een misclassificatie optreedt, des te groter de invloed op de volledige structuur zal zijn. Onderzoek toont aan dat hiërarchische clusterprocedures afwijkingen gaan vertonen bij hoge niveaus van bezetting: dit is wanneer meer dan 90% van alle individuen opgenomen zijn in de analyse (Punj en Stewart, 1983, blz. 143). Wegens de mogelijke aanwezigheid van "outliers" is het toepassen van alleen een hiërarchische clusteranalyse geen goede praktijk. Deze methoden zijn echter zeer populair en veelgebruikt. Een aantal belangrijke statistische pakketten neemt enkel die methoden op. De methoden lijden echter aan zodanige gebreken dat ze met de grootste voorzichtigheid aangewend moeten worden.

Het minst gevoelig voor deze gebreken is de methode van Ward. Die gaat ervan uit dat bij elke stap in de analyse nauwkeurig kan worden gemeten wat het informatieverlies is dat het groeperen van clusters teweegbrengt. Dat kan gebeuren door de som te berekenen van de gekwadrateerde afwijkingen van elk individu in een cluster tot het gemiddelde van die cluster. Alle mogelijke combinaties van twee clusters worden in elke fase nagegaan en de fusie van die twee clusters die de

minste toename in de som van de gekwadrateerde afwijkingen meebrengt, wordt doorgevoerd.

De methode wordt ook minimumvariantiemethode genoemd, omdat zij zal proberen clusters te genereren met een zo klein mogelijke variantie binnen elke cluster. De methode maakt gebruik van de "error sum of squares"-index en is zowel bruikbaar voor binaire als voor continue metingen. Voor een uitvoerige bespreking van de wiskundige onderbouwing wordt verwezen naar Wishart (1987, blz. 91).

De "error sum of squares"-index heeft als voordeel dat binnen de bekomen clusters de binnenvariantie geminimaliseerd wordt. Dit leidt tot sferische, zeer dichte, goed afgebakende clusters. Dat komt overeen met de doelstelling van clusteranalyse voor marktsegmentatie, die erin bestaat de n respondenten te groeperen in k clusters (k<n) die intern zo homogeen en extern zo heterogeen mogelijk zijn t.o.v. alle in de clusteranalyse opgenomen karakteristieken (de attributensets).

Behalve hiërarchische methoden wordt ook veel gebruik gemaakt van zogenaamde "K-means"-partitiemethoden. "K-means"-partitiemethoden starten met de n te clusteren individuen in k (k<n) a priori bepaalde clusters te verdelen. De partities kunnen al dan niet toevalsmatig gebeuren. De algoritme tracht vervolgens de individuen te realloceren naar die cluster waartoe de afstand tot het centrum minimaal is. De procedure stopt wanneer dat voor alle individuen gerealiseerd is. Het hele proces verloopt dus iteratief. Cruciaal hierbij is de startpositie of de manier waarop respondenten aan een startcluster worden toegewezen. In de praktijk gebeurt dat dikwijls toevalsmatig. Onderzoek toont echter aan dat alle "K-means"-partitiemethoden slechte eindoplossingen genereren wanneer voor een "at random" startpositie wordt gekozen. Dit staat in schril contrast tot de gangbare praktijk (Anderberg, 1973, blz. 160; Milligan, 1980, blz. 334).

Onderzoek toont aan dat hiërarchische methoden, meer in het bijzonder de methode van Ward, excellente startposities kunnen genereren. Uit dit onderzoek blijkt de superioriteit van de "K-means"-methoden boven de hiërarchische methoden wanneer gekozen wordt voor een startconfiguratie die gebaseerd is op de resultaten van een hiërarchische methode (Milligan, 1980, blz. 339; Wishart, 1987, blz. 145).

Dit leidt meteen tot de opbouw van een goede procedure die uit twee stappen bestaat. Eerst wordt een hiërarchische clusteranalyse uitgevoerd via de methode van Ward. Er wordt gestopt op een te groot aantal clusters. De resultaten hiervan dienen als input voor een "K-means"-methode. In het voorbeeld is bij de toepassing van de "K-means"-methode niet verder gewerkt met de methode van Ward. Deze methode leidt tot het bepalen van afstanden (of gelijkenissen) tussen individuen waarbij een 0-0-overeenkomst evenveel betekenis heeft als een 1-1-overeenkomst. Daar de gegevens verzameld zijn via de "pick any"-methode, betekent een 0 alleen maar dat de respondent het desbetreffende attribuut niet geselecteerd heeft. Het volgende citaat maakt dit duidelijk: "For example, suppose the data units are animals and the variables are 'has feathers', and 'has webbed feet'. Dogs and cats (...) would fall into cell D because there is no way they could have such attributes. It would be misleading to allow these 0-0 matches to contribute to the measure of association between dogs and cats" (Anderberg, 1973, blz. 88). Voor een overzicht van associatiematen voor binaire variabelen wordt verwezen naar Kaufman en Rousseeuw (1989).[3]

Deze methoden en associatiemaatstaven vinden we o.a. terug bij Wishart; ze zijn opgenomen in het op pc beschikbare programmapakket Clustan (Wishart, 1987, blz. 145).

Toegepast op het voorbeeld geeft dit de volgende bewerkingen. Eerst wordt een hiërarchische clusteranalyse via de methode van Ward toegepast tot uiteindelijk 10 clusters overblijven. Dit aantal is subjectief bepaald maar toch voldoende groot als startconfiguratie voor de "K-means"-methode. Op basis van het uitgevoerde kwalitatieve onderzoek verwachten we slechts 4 clusters. De 10 clusters uit de hiërarchische clustering dienen vervolgens als vertrekpunt voor een eerste reallocatie via de "K-means"-methode. Een optimum wordt bereikt. Vervolgens worden die twee clusters samengevoegd waarvan de centra de kleinste afstand tot elkaar hebben. Binnen de 9 overblijvende clusters worden de individuen opnieuw gerealloceerd tot een optimum bereikt wordt. Opnieuw worden die twee clusters samengevoegd waarvan de centra de kleinste afstand tot elkaar hebben. Dezelfde procedure wordt herhaald voor 8, 7, 6, 5, 4, 3 en tenslotte 2 clusters.

3 Met dank aan een anonieme referee.

Naast hiërarchische en partitiemethoden bestaan er nog talrijke andere methoden voor clusteranalyse. Voor een bespreking hiervan wordt verwezen naar de literatuur (o.a. Everitt, 1974: Wishart, 1987). Volledigheidshalve wordt kort iets gezegd over "clumping"-methoden wegens hun groeiende populariteit. "Clumping"-methoden maken overlappende clusters mogelijk. Ze verlaten de hypothese van wederzijds exclusieve en collectief exhaustieve clusters. Overlappende clusteranalyse vertrekt van het feit dat bijv. merken kunnen concurreren in meer dan een competitieve set, of dat een individu tot meer segmenten kan behoren. Clusters kunnen aldus overlappen. Bekende procedures zijn "adclus" en "mapclus". Een beschrijving hiervan zou ons opnieuw te ver leiden. Daarvoor wordt verwezen naar de gespecialiseerde literatuur (Arabie en Carroll, 1980; Arabie, Carroll, DeSarbo en Wind, 1981).

6. Bepaling van het optimale aantal clusters en evaluatie van de clusterstructuur

In het voorbeeld is een clusteranalyse uitgevoerd via de combinatiemethode van Wishart. Een oplossing is berekend voor 10, 9, 8, ..., tot 2 clusters. Twee vragen dienen nu beantwoord te worden:

1. Hoe groot is het optimale aantal clusters?
2. Wanneer dit optimale aantal clusters geselecteerd is, kunnen we dan een zekerheid inbouwen dat de bekomen clusters zinvol zijn? Met andere woorden: hoe betrouwbaar is de bekomen oplossing?

A. Bepaling van het optimale aantal clusters

Het bepalen van het optimale aantal clusters is niet gemakkelijk. Ontelbare procedures om dit aantal te bepalen zijn in de literatuur voorhanden (Milligan en Cooper, 1985, blz. 159).

Men spreekt in dit verband van "stopcriteria". Twee types fouten kunnen zich voordoen. De eerste fout komt voor wanneer het stopcriterium k clusters aangeeft waar er in werkelijkheid minder dan k clusters zijn. De tweede fout komt voor wanneer het stopcriterium k clusters aangeeft en er in werkelijkheid meer dan k zijn. In het eerste geval heeft de oplossing te veel clusters, in het tweede te weinig. Het tweede geval is

ernstiger dan het eerste, daar informatie verloren gaat door duidelijk verschillende clusters samen te voegen.

In de literatuur worden veel stopcriteria beschreven, doch veel ervan zijn relatief onbetrouwbaar. Voor een uitgebreid overzicht en vergelijking wordt verwezen naar Milligan en Cooper (1985, blz. 163). Een in de praktijk veelgebruikt stopcriterium is de toename in de "error sum of squares". Dergelijke stopregel berekent de totale "error sum of squares" op elk clusterniveau (bijv. 10, 9, 8, 7, ..., 3, 2 clusters). Op een grafiek wordt de toename in de "error sum of squares" uitgezet t.o.v. het aantal clusters. Het correcte aantal clusters wordt bepaald op basis van de sterke knik in de figuur bij overgang van het vermoedelijk juiste aantal clusters naar een niveau met te weinig clusters. Bij die overgang is er een belangrijke toename in de "error sum of squares", wat erop zou wijzen dat de knik het optimum is. Uit experimenten (Milligan en Cooper, 1985, blz. 172) blijkt dit stopcriterium niet altijd het juiste aantal clusters weer te geven. Volgens hun vergelijkend onderzoek leidt dit stopcriterium in slechts 28% van de gevallen tot het juiste clusteraantal. Volgens hetzelfde onderzoek is de index van Calinski en Harabasz het beste stopcriterium. Probleem daarbij is dat dergelijke index slechts in een beperkt aantal softwarepakketten is opgenomen. In het voorbeeld geven de index van Calinski en Harabasz en de "error sum of squares" allebei een optimum van 4 clusters aan. Voor een voorbeeld van berekening en toepassing van beide indices wordt verwezen naar De Pelsmacker en Van Kenhove (1994, blz. 451-457).

B. Evaluatie van de clusterstructuur

Het toepassen van clusteranalyse op een dataset zal altijd aanleiding geven tot clusters. De vraag is nu of die gegenereerde oplossing ook stabiel en consistent is. Met andere woorden: zijn de clusters betrouwbaar? Verschillende vormen van betrouwbaarheid kunnen worden onderzocht.

In een eerste vorm van betrouwbaarheid wordt de interne consistentie van een oplossing onderzocht via de "split-run"-procedure (Punj en Stewart, 1983, blz. 145). Daarbij wordt een clusteranalyse op een steekproef uitgevoerd en wordt vervolgens een optimaal aantal clusters geselecteerd. Dan wordt de steekproef toevalsmatig in tweeën gesplitst, waarna dezelfde clusteranalyse op beide deelsteekproeven wordt overgedaan. Vervolgens wordt de clustersamenstelling (welke individuen bevinden zich in welke cluster) voor beide deelsteekproeven vergeleken met de clustersamenstelling uit de oorspronkelijke clusteranalyse. Als er een grote overeenkomst is tussen de oorspronkelijke clusters en de clusters uit de "split-run"-test, dan is de clustersamenstelling stabiel. Dergelijke meting is in de praktijk zeer gemakkelijk uit te voeren. In het voorbeeld is de overeenkomst tussen de oorspronkelijke clusteranalyse en de twee deelsteekproeven zeer hoog: 89% van de individuen uit de eerste deelsteekproef bevinden zich in dezelfde cluster als in de oorspronkelijke clusteranalyse. Voor de tweede deelsteekproef is dat zelfs 94%, wat wijst op een hoge mate van stabiliteit.

Als tweede vorm van betrouwbaarheid wordt de homogeniteit van een oplossing onderzocht. De homogeniteit geeft een beeld van de interne samenhang van de individuen binnen elke cluster. Een hoge homogeniteit binnen een cluster wijst erop dat alle individuen binnen die cluster een vrij gelijksoortig antwoordpatroon hebben t.o.v. de attributen waarop alle respondenten gescoord hebben. Dat kan ten eerste nagegaan worden via de binnenvariantie in elke cluster. Wanneer die te groot wordt, zijn de clusters onvoldoende homogeen. Ook dit kan in de praktijk gemakkelijk worden nagegaan. Het enige probleem met deze methode is dat er geen referentiewaarden bestaan: wat is te groot? De onderzoeker kan zich enkel behelpen met een aantal vuistregels, zoals het onderling vergelijken van de binnenvariantie van elk van de clusters: die mogen onderling niet te veel afwijken. Voor een overzicht hiervan wordt verwezen naar Milligan (1981, blz. 187) en Klastorin (1983, blz. 92). In het voorbeeld is de berekende binnenvariantie voor elk van de clusters ongeveer even groot.

7. Visuele voorstelling van de resultaten via FCA

A. Inleiding

Tot nog toe zijn de resultaten vrij abstract. Wat de clusters betekenen zullen we nu visueel trachten voor te stellen via FCA. Een visuele voorstelling heeft het voordeel interpretaties te vergemakkelijken. De output van clusteranalyse is meteen de input voor FCA. In het voorbeeld is dat een datamatrix die bestaat uit 46 rijen (de attributen) en 4 kolommen (de bekomen clusters). Enkel de attributen waarvoor tussen de

clusters significante verschillen bestaan ($\alpha = 0{,}05$) zijn in tabel 1 opgenomen. Op die manier blijven er 46 van de oorspronkelijke 64 attributen over.

In elke cel staat de associatie tussen een attribuut en een cluster in de vorm van een penetratiecijfer. Het getal 40 in de eerste rij en derde kolom wil dus zeggen dat binnen cluster 3 40 individuen (op een totaal van 81) dit attribuut hebben geassocieerd met hun voorkeurwinkelcentrum. Een dergelijke tabel wordt een contingentietabel genoemd.

B. Voordelen van FCA

Welke voordelen biedt FCA nu bij de interpretatie van dergelijke tabel?

Ten eerste worden de data multivariaat beschouwd, daar alle cellen uit de dataset (rijen en kolommen) simultaan vergeleken worden. De multivariate benadering heeft als voordeel dat verbanden tussen variabelen (bijv. clusters onderling, attributen onderling en clusters én attributen onderling) te voorschijn kunnen komen die onzichtbaar blijven bij paarsgewijze vergelijking.

Ten tweede legt FCA ook het verband tussen de variabelen (rijen en kolommen). Met andere woorden: niet alleen het feit dat variabelen gerelateerd zijn of niet maar ook hoe ze gerelateerd zijn, wordt onderzocht. Dat gebeurt door een weergave in een gemeenschappelijke visuele voorstelling van clusters en attributen, "joint space" genoemd.

Ten derde stelt de techniek weinig vereisten aan de data. De enige strikte voorwaarde is in principe een rechthoekige matrix met niet-negatieve cellen.

Met betrekking tot dat laatste punt zijn er een aantal bijkomende voorwaarden. De datamatrix moet groot genoeg zijn, zodat de structuur niet met het blote oog of via simpele statistische technieken achterhaald kan worden. De variabelen moeten homogeen zijn. Met variabelen worden opnieuw de elementen in de rijen (bijv. attributen) en in de kolommen (bijv. clusters) bedoeld. Dit wil zeggen dat de gegevens in de rijen respectievelijk kolommen onderling vergelijkbaar moeten zijn. Het is bijv. niet zinvol om een attribuut dat gemeten is via een semantische differentiaal in dezelfde matrix naast een attribuut te zetten dat gemeten

Tabel 1. Contingentietabel na clusteranalyse.

Attribuut	Cluster 1	2	3	4
1 Alledaags	0	2	40	2
2 Duur	30	3	3	6
3 Efficiënt	22	33	27	2
4 Exclusief	29	1	2	2
5 Kwaliteit	71	40	7	24
6 Gediversifieerd aanbod	3	47	18	3
7 Gelijksoortige winkels	2	17	49	2
8 Gewoonte	6	42	18	6
9 Goede prijs-kwaliteitverhouding	4	40	23	5
10 Kledij voor meer dan 1 seizoen	4	18	32	5
11 Kleine winkeltjes	17	9	25	55
12 Om te kopen	1	33	35	2
13 Parkeergelegenheid	14	36	41	14
14 Rondsnuffelen/om te kijken	19	19	40	64
15 Koopjes	3	38	26	8
16 Trendsettend	48	5	7	24
17 Winkeltrouw	42	7	13	19
18 Bekende merken	55	14	13	25
19 Genieten	46	1	2	49
20 Gezellig en sfeervol	24	3	49	85
21 Lang winkelen	39	4	19	63
22 Modebewust	43	7	15	6
23 Ontspannend	5	12	19	70
24 Oppervlakkig	2	29	40	0
25 Persoonlijk	49	2	4	35
26 Rumoerig	1	1	1	22
27 Rustig	4	20	50	4
28 Standing	63	2	6	9
29 Tijdverdrijf	13	6	26	72
30 Voor iedereen	2	15	64	43
31 Braaf	0	12	46	0
32 Business class	60	0	2	10
33 Chique	61	2	5	11
34 Druk	6	2	32	52
35 Eenvoudig	2	10	57	18
36 Gewoon	2	10	49	2
37 Jong	9	0	44	72
38 Klassiek	59	24	4	9
39 Sociaal	0	12	22	2
40 Sportief	29	0	5	29
41 Sympathiek	20	2	9	86
42 Veeleisend	45	0	4	5
43 Vlot	14	2	37	69
44 Volks	0	11	42	0
45 Vriendelijk	8	1	46	42
46 Zelfzeker	52	0	17	2
Aantal respondenten	77 25%	53 18%	81 27%	91 30%

is via de "pick any"-methode. De data daarentegen moeten heterogeen zijn. Met data wordt datgene bedoeld dat in de cellen van de matrix komt te staan. Als er binnen de dataset geen variantie aanwezig is, dan wordt een visuele voorstelling ook niet meer zinvol. De voorstelling is er juist op gericht om de variantie binnen de dataset visueel tot uiting te brengen.

Toegepast op dit voorbeeld biedt FCA een antwoord op de vier volgende vragen:

1. Wat zijn de similariteiten en verschillen tussen de 4 clusters m.b.t. de 46 attributen?
2. Wat zijn de similariteiten en verschillen tussen de 46 attributen m.b.t. de 4 clusters?
3. Wat is de relatie tussen de 4 clusters en de 46 attributen?
4. Kunnen die relaties (vraag 1-3) worden voorgesteld in een gemeenschappelijke ruimte, met zo weinig mogelijk dimensies?

Het is, m.a.w., de expliciete doelstelling om een inzicht te verwerven in de interrelaties tussen de clusters (segmenten) onderling, elk van de attributen onderling en alle elementen samen (clusters en attributen).

C. Werking van FCA en interpretatie van de oplossing

In de praktijk wordt de meeste aandacht geschonken aan de visuele voorstelling. Meestal wordt standaard een tweedimensionele oplossing weergegeven. Enkel deze visuele voorstelling interpreteren is echter in de meeste gevallen onvoldoende om een complete en correcte interpretatie van de resultaten te maken (Hoffman en Franke, 1986, blz. 219). Immers, uit de visuele voorstelling weten we niet:

(1) of een weergave in twee dimensies wel een "goed" beeld geeft van de data,
(2) wat het belang is van elke as,
(3) welke punten de meeste impact hebben bij de oriëntatie van de assen,
(4) hoe "goed" een punt weergegeven is in de ruimte, en
(5) wat er kan gebeuren met additionele externe informatie.

Op elk van deze vragen zal kort worden ingegaan. Voor een wiskundig overzicht wordt verwezen naar Greenacre (1984).

Vraag (1) heeft betrekking op de vraag of een visuele weergave in twee assen (zie verder: figuur 1) wel een "goed" beeld geeft van de datamatrix. Hierbij aansluitend kan vraag (2) gesteld worden naar het belang van elke individuele as.

Hiertoe berekent de algoritme de "inertie" van de oplossing, ook wel "principal inertias" genoemd . De inertie geeft een algemeen beeld van de kwaliteit van de oplossing, t.t.z. wat de bijdrage is van elke as in de totale oplossing (Hoffman en Franke, 1986, blz. 219). De inertie van elke as geeft aan hoeveel van de totale variantie in de data verklaard wordt door elke as. Ze wordt relatief uitgedrukt t.o.v. de totale inertie. We merken op dat 100% van de inertie steeds verklaard wordt door het minimum van het aantal rijen en kolommen in de matrix -1. Dit betekent dat een matrix bestaande uit 6 rijen en 4 kolommen steeds perfect kan worden voorgesteld in 3 dimensies.

In het voorbeeld verklaren drie assen dus 100% van de inertie (variantie). Tabel 2 toont de verklaarde variantie van elke as.

Tabel 2. Inertie van de assen.

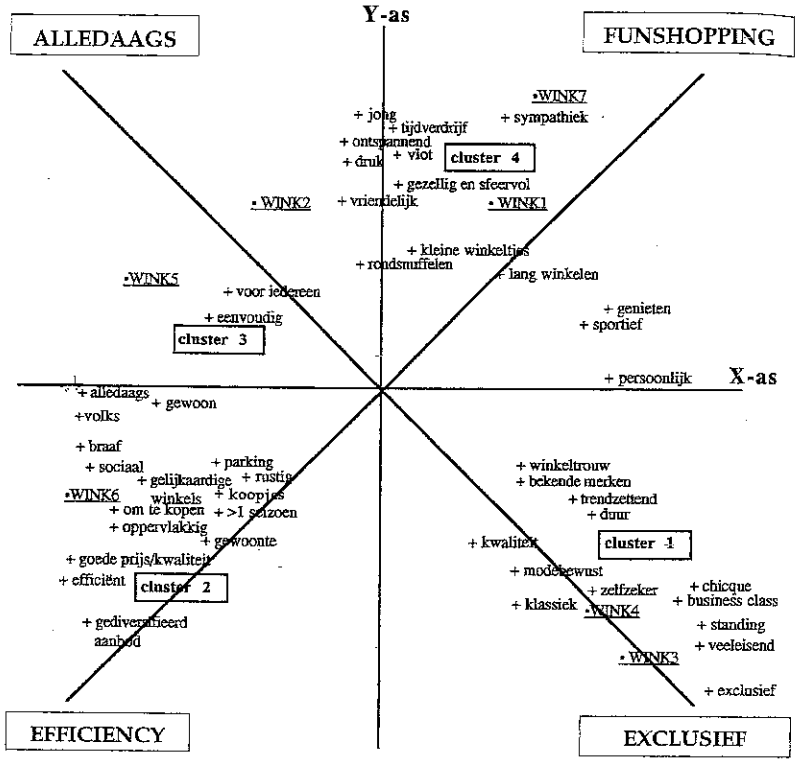
As 1	53%
As 2	34%
As 3	13%
Totaal	100%

Wanneer we enkel de tweedimensionele oplossing beschouwen, verliezen we dus iets aan informatie. Een additionele derde as zou in ons voorbeeld echter slechts 13% bijkomende variantie hebben verklaard, wat betekent dat de datamatrix "goed" is weergegeven in twee dimensies.

Wat onder "goed" dient te worden verstaan, zal van geval tot geval bekeken moeten worden. Als een as met een relatief hoge verklaarde variantie (20% of meer) wordt weggelaten, zal nauwkeurig onderzocht moeten worden wat het effect daarvan is op de visuele voorstelling.

Vervolgens kunnen de coördinaten van elk van de rij- en kolomelementen visueel worden voorgesteld. In het voorbeeld geeft dat het beeld van figuur 1.

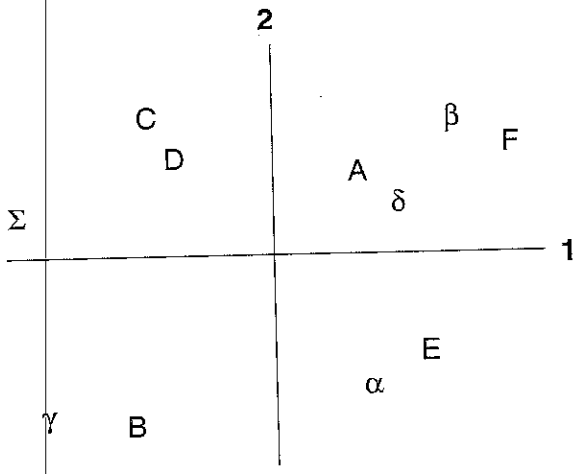
Figuur 1. Visuele voorstelling van clusters en attributen (FCA).



Nogal wat controverse bestaat over de interpretatie van een dergelijke visuele voorstelling (zie o.a. Volle, 1985, blz. 143; Greenacre, 1984, blz. 65). Deze controverse kan het best worden geïllustreerd aan de hand van figuur 2. Hierin zijn de rij-elementen van een datamatrix in hoofdletters en de kolomelementen in Griekse letters afgebeeld.

In oorspronkelijke versies van FCA, voornamelijk gebruikt in Frankrijk (verder de Franse school genoemd), is het verboden een onmiddellijke associatie te maken tussen de rij- en de kolomelementen. De op het eerste gezicht kleine afstand tussen A en δ is mogelijk zeer groot. Men mag wel iets zeggen over de afstand tussen de rijelementen of de afstand tussen de kolomelementen, maar niet tussen beide.

Figuur 2. Voorstelling van FCA.



Om het vervelende probleem van niet-vergelijkbaarheid tussen twee sets (rijen en kolommen) op te lossen, hebben Carroll, Green en Schaffer (1986, 1987) een nieuwe methode beschreven waarmee wel vergelijkingen gemaakt kunnen worden tussen alle punten binnen een set én tussen sets. Wij zullen die methode omschrijven als de CGS-methode.

Naast de mogelijkheid alle punten (ook tussen sets) met elkaar te vergelijken heeft de CGS-methode een tweede voordeel, namelijk dat het assenstelsel geroteerd en de oorsprong van plaats gewisseld kunnen worden, zonder dat de euclidische afstand tussen de punten onderling verandert. Dat is niet zo in de andere oplossingen (Franse school), daar afstanden tussen sets (rijen en kolommen) zonder betekenis zijn. Mogelijkheid tot rotatie van de assen vergemakkelijkt de interpretatie. In de Franse versie gebeurt de interpretatie op basis van projectie op de assen.

In het voorbeeld is de CGS-methode toegepast, opgenomen in het pakket SPSS. De interrelaties tussen de 4 clusters onderling, de attributen onderling en de clusters én attributen kunnen aldus bestudeerd worden.

Vooraleer de visuele voorstelling te interpreteren dienen we echter nog stil te staan bij de derde en de vierde vraag. Vraag (3) (welke punten de meeste impact hebben bij de oriëntatie van de assen) wordt bepaald

via de bijdrage van elk punt tot de inertie in elke as (Hoffman en Franke, 1986, blz. 220). De punten met hoge contributie spelen de grootste rol in de bepaling van de oriëntatie van de assen.

Dergelijke informatie is dus uiterst relevant om de stabiliteit van de oplossing te bepalen. Immers, wordt de contributie op een as praktisch volledig bepaald door 1 punt, dan kan het weglaten van de betreffende rij of kolom uit de datamatrix waarvoor dit punt staat en de heranalyse van de resterende rijen en kolommen een totaal andere configuratie opleveren. Draagt een punt anderzijds weinig bij tot de contributie op een as, dan zal het weglaten van de respectieve rij of kolom weinig invloed hebben op de uiteindelijke configuratie. Daarnaast is dit gegeven relevant om de as te benoemen. Immers, de punten met de hoogste contributie op een as bepalen de richting van de as en kunnen dus dienen om de assen te benoemen en te interpreteren. In ons voorbeeld geeft dit de resultaten van tabel 3.

Clusters 1 en 4 hebben blijkbaar een grote rol gespeeld bij de oriëntatie van de X-as (cluster 1) en de Y-as (cluster 4). Belangrijk voor de oriëntatie van de X-as zijn de attributen 12, 24, 28, 32 en 33 en voor de Y-as 20, 23, 29, 37, 38, 41 en 43. Elk van deze punten neemt een relatief extreme positie in in figuur 1.

Vraag (4) heeft betrekking op de vraag hoe goed een rij of kolom door een punt weergegeven wordt in de ruimte door middel van de geselecteerde assen. Dat wordt de contributie van de as tot de inertie van het punt genoemd. Dit bepaalt de kwaliteit van de oplossing (Hoffman en Franke, 1986, blz. 220). In tabel 4 worden de resultaten voor het voorbeeld weergegeven.

Het enige punt dat iets minder goed is voorgesteld in de tweedimensionele ruimte, is attribuut 1 met 53% verklaring door een derde as. Dit betekent dat de plaats in de visuele voorstelling van dit punt met de nodige voorzichtigheid moet worden geïnterpreteerd.

Vier groepen attributen worden onderscheiden aan de hand waarvan 2 (subjectieve) assen kunnen worden benoemd (zie figuur 1):

Cluster 1 is op zoek naar klasse, exclusiviteit en is zeer modebewust. De relatie tot winkelen en winkelcentra wordt gekenmerkt door aantrekking tot het chique en dure en afstoting van alles wat gewoon en minder

Tabel 3. Contributie op de X- en Y-as.

Attribuut	Contributie op X-as	Contributie op Y-as	Contributie op Z-as
1	0,022	0,001	0,102
2	0,019	0,013	0,002
3	0,035	0,018	0,019
4	0,023	0,019	0,006
5	0,011	0,045	0,039
6	0,037	0,037	0,109
7	0,035	0,003	0,031
8	0,026	0,027	0,081
9	0,031	0,023	0,054
10	0,020	0,003	0,002
11	0,001	0,023	0,005
12	0,041	0,015	0,006
13	0,019	0,009	0,005
14	0,000	0,017	0,006
15	0,031	0,015	0,041
16	0,026	0,004	0,000
17	0,014	0,005	0,001
18	0,017	0,010	0,000
19	0,037	0,004	0,005
20	0,001	0,052	0,001
21	0,014	0,016	0,002
22	0,012	0,021	0,010
23	0,000	0,054	0,037
24	0,041	0,013	0,000
25	0,035	0,000	0,000
26	0,001	0,029	0,020
27	0,032	0,003	0,020
28	0,047	0,029	0,012
29	0,001	0,053	0,009
30	0,017	0,016	0,012
31	0,035	0,001	0,051
32	0,056	0,025	0,009
33	0,047	0,024	0,009
34	0,000	0,045	0,000
35	0,022	0,003	0,046
36	0,029	0,000	0,066
37	0,000	0,067	0,002
38	0,016	0,051	0,010
39	0,022	0,002	0,006
40	0,019	0,002	0,000
41	0,012	0,074	0,035
42	0,038	0,022	0,013
43	0,000	0,052	0,000
44	0,032	0,001	0,047
45	0,002	0,028	0,021
46	0,024	0,027	0,047
Totaal	1,000	1,000	1,000
Cluster 1	0,470	0,238	0,032
Cluster 2	0,208	0,219	0,420
Cluster 3	0,278	0,004	0,424
Cluster 4	0,045	0,539	0,124
Totaal	1,000	1,000	1,000

duur is. Dit type consumente wenst zich in haar relatie tot haar winkelcentrum duidelijk te distantiëren van de "gewone" consument.

Cluster 2 distantieert zich compleet van alles wat met winkelen te maken heeft. Winkelen moet zo snel, efficiënt en goedkoop mogelijk zijn. Het winkelcentrum moet een breed aanbod hebben en goede parkeergelegenheid, zodat alle boodschappen ineens en vlug gedaan kunnen worden. Winkelen is niet plezierig voor dit type consument, het is een plicht. Daarom ligt cluster 2 in de map ver van alle "positieve" eigenschappen. Merk wel op dat het hier ook modegevoelige vrouwen betreft.

Cluster 3 is op zoek naar het gewone, alledaagse. Die mensen hebben een positieve ingesteldheid ten opzichte van winkelen, maar zijn in tegenstelling tot cluster 1 zeker niet op zoek naar klasse en exclusiviteit. Zij wensen zich duidelijk niet te onderscheiden van de anderen maar voelen zich goed in de massa.

Voor *cluster 4* is winkelen "funshopping". Zij slenteren en snuffelen graag rond en bekijken alles. Zij houden ervan lang te winkelen. Winkelen is voor hen duidelijk een aangename ontspanning en een leuk tijdverdrijf. Ze houden van een vriendelijke, jonge, dynamische en gezellige winkelomgeving.

Op basis van de interrelaties tussen de attributen in de visuele voorstelling kan een poging ondernomen worden om assen te benoemen. De eerste as loopt van cluster 4 naar cluster 2. De attributen rond cluster 2 zijn tegenovergesteld aan die rond cluster 4. De tweede as loopt van cluster 3 naar cluster 1. De attributen rond cluster 2 en de afstand tot alle andere attributen stellen ons in staat te concluderen dat die cluster in zijn relatie tot winkelen en winkelcentra op zoek is naar efficiency. De andere pool van de as wordt gevormd door cluster 4 en kan worden omschreven via de aldaar gesitueerde attributen. Die cluster is op zoek naar het tegenovergestelde van cluster 2, namelijk "funshopping". Een eerste as zouden we dus kunnen noemen: efficiency versus "funshopping".

Cluster 1 is op zoek naar exclusiviteit en distantieert zich van de massa. Aan de tegenovergestelde kant bevindt zich cluster 3, die zich juist comformeert en op zoek is naar het gewone, alledaagse. De tweede as zouden we kunnen noemen: exclusiviteit versus alledaags.

Tabel 4. Contributie van de X-, Y- en Z-as.

Attribuut	Contributie van X-as	Contributie van Y-as	Contributie van Z-as	Totaal
1	0,460	0,007	0,532	1,000
2	0,689	0,296	0,014	1,000
3	0,678	0,228	0,094	1,000
4	0,637	0,320	0,044	1,000
5	0,225	0,575	0,200	1,000
6	0,426	0,266	0,309	1,000
7	0,788	0,036	0,175	1,000
8	0,407	0,272	0,321	1,000
9	0,526	0,247	0,227	1,000
10	0,883	0,096	0,021	1,000
11	0,032	0,893	0,075	1,000
12	0,791	0,180	0,029	1,000
13	0,722	0,232	0,046	1,000
14	0,036	0,852	0,112	1,000
15	0,612	0,186	0,202	1,000
16	0,916	0,083	0,000	1,000
17	0,791	0,192	0,017	1,000
18	0,720	0,275	0,005	1,000
19	0,898	0,069	0,034	1,000
20	0,022	0,972	0,005	1,000
21	0,568	0,407	0,024	1,000
22	0,428	0,478	0,094	1,000
23	0,000	0,785	0,215	1,000
24	0,828	0,172	0,000	1,000
25	0,997	0,000	0,003	1,000
26	0,058	0,742	0,200	1,000
27	0,819	0,052	0,129	1,000
28	0,690	0,266	0,043	1,000
29	0,025	0,914	0,061	1,000
30	0,570	0,334	0,097	1,000
31	0,721	0,017	0,263	1,000
32	0,758	0,211	0,032	1,000
33	0,729	0,238	0,033	1,000
34	0,006	0,992	0,003	1,000
35	0,625	0,057	0,319	1,000
36	0,637	0,006	0,357	1,000
37	0,001	0,990	0,009	1,000
38	0,315	0,636	0,049	1,000
39	0,894	0,044	0,062	1,000
40	0,930	0,067	0,003	1,000
41	0,182	0,690	0,128	1,000
42	0,685	0,255	0,060	1,000
43	0,008	0,992	0,000	1,000
44	0,722	0,017	0,261	1,000
45	0,086	0,706	0,208	1,000
46	0,462	0,319	0,219	1,000
Cluster 1	0,749	0,239	0,013	1,000
Cluster 2	0,461	0,307	0,232	1,000
Cluster 3	0,719	0,007	0,274	1,000
Cluster 4	0,107	0,818	0,074	1,000

Vraag (5) stelt de vraag wat er kan gebeuren met additionele externe informatie. Een heel interessante eigenschap van factoriële correspondentieanalyse bestaat erin dat additionele externe informatie in de visuele voorstelling kan worden ingebracht zonder invloed op de bestaande visuele voorstelling. Dat gebeurt via de "transition formulae" en wordt de passieve-celoptie genoemd. Aangezien de rij- en de kolomelementen van de matrix in de visuele voorstelling zijn opgenomen, kan elke andere derde variabele, opgenomen in een matrix die de associaties bevat tussen die derde variabele en de rij- of kolomelementen waarvan hiervoor sprake, in de visuele voorstelling worden voorgesteld (zie voor meer details Greenacre, 1984, blz. 83; en vooral Hoffman en Franke, 1986, blz. 217).

In ons voorbeeld zouden aldus de verschillende winkelcentra als punten in de bestaande visuele voorstelling kunnen worden ingebracht. Deze coördinaten worden berekend aan de hand van de kruistabel tussen de clusters en de voorkeurwinkelcentra (zie tabel 5). Hierin zien we bijv. dat 43 van de 77 respondenten in cluster 1 winkelcentrum 3 als voorkeur hebben opgegeven. Dit winkelcentrum zal dus in de bestaande oplossing dicht bij cluster 1 moeten liggen. Uit de tabel blijkt overigens dat de voorkeurpatronen van de clusters sterk verschillen. Voor 7 van de 10 winkelcentra zijn er significante verschillen in voorkeur tussen de 4 clusters ($\alpha = 0{,}05$). Vanwege de confidentialiteit zijn de namen van de winkelcentra vervangen door een cijfer.

Tabel 5. Kruistabel clusters versus voorkeurwinkelcentrum.

	Cluster					
	1	2	3	4	Totaal	Kwaliteit
Voorkeurcentrum						
WINK1	5	7	5	20	37	75
WINK2	1	2	13	26	42	95
WINK3	43	1	1	1	46	99
WINK4	28	2	1	8	39	99
WINK5	0	4	32	3	39	54
WINK6	0	37	26	3	66	87
WINK7	0	0	3	30	33	98
Totaal cluster	77	53	81	91	302	

Significantie (gebaseerd op χ^2): $\alpha<0{,}05$

Opnieuw kan worden onderzocht hoe "goed" of "slecht" de additionele informatie is weergegeven in de visuele voorstelling (zie de laatste kolom van tabel 5). Alleen winkelcentrum 5 is iets minder goed weergegeven in de tweedimensionele oplossing. Slechts 54% van de variantie in dit punt wordt verklaard door de X- en de Y-as.

Additioneel worden de coördinaten in de map ingebracht. Dat geeft het volgende resultaat (zie opnieuw figuur 1).

Cluster 1 heeft een duidelijk uitgesproken voorkeur voor winkelcentra 3 en 4, waartoe de andere clusters veel minder aangetrokken worden. Die twee winkelcentra zijn heel exclusief en duur. Cluster 2 heeft met geen enkel winkelcentrum een uitgesproken relatie, wat eigenlijk logisch is. De enige voorkeur gaat iets meer uit naar winkelcentrum 6, dat inderdaad een algemeen centrum is met een groot parkeerterrein. Cluster 3 heeft een voorkeur voor winkelcentra 5 en 6, die tamelijk volks zijn. Cluster 4 tenslotte winkelt graag in winkelcentra 1, 2 en 7, die zeer gezellig en jong zijn. Zij hebben vanuit hun graag winkelen de meeste centra opgegeven.

De visuele voorstelling stelt ons in staat de onderlinge concurrentie tussen de verschillende winkelcentra in te schatten. De conclusie die uit de map kan worden getrokken, is dat de verschillende winkelcentra in Brussel toch vrij aardig op de gedifferentieerde behoeftenpatronen inspelen. Geen enkel winkelcentrum neemt een echte middenpositie ("stuck in the middle") in.

Vervolgens kunnen kruistabellen worden gemaakt tussen de clusters en alle andere vragen uit de vragenlijst (zoals leeftijd, taal, gekochte merken, bezoek - vroeger en nu - aan winkelcentra). De gegevens dienen om de gevonden segmenten beter te kunnen identificeren. Een aantal significante resultaten wordt teruggevonden. Zo blijken de clusters significant te verschillen op socio-demografische karakteristieken zoals leeftijd en taal. Cluster 1 is relatief minder jong (18-29 jaar: slechts 28%), Franssprekend (66%) en bevat slechts 19% kopers van het merk van de opdrachtgever, cluster 2 is eveneens relatief minder jong (18-29 jaar: 19%), Franssprekend (66%) en bevat het grootste aantal kopers van het merk van de opdrachtgever (36%), cluster 3 is relatief jong (18-24 jaar: 40%), minder Franstalig (50%) en bevat 30% kopers van het merk van de opdrachtgever, cluster 4 is eveneens relatief jong (18-24 jaar: 43%, zeer Franstalig (84%) en niet sterk geïnteresseerd in het merk van de

opdrachtgever (16%). Om redenen van confidentialiteit zijn de resultaten uit andere kruistabellen niet weergegeven in dit artikel.

Besluit

Marktsegmentatie is een van de belangrijkste concepten in de hedendaagse marketing. Wegens de steeds complexere consumentenvraag (vraagzijde) en de steeds fijner wordende produktiemogelijkheden (aanbodzijde) neemt het belang van marktsegmentatie nog steeds toe. Clusteranalyse en factoriële correspondentieanalyse zijn twee exploratieve multivariate statistische technieken die bijzonder bruikbaar zijn voor marktsegmentatie. Beide technieken zijn vandaag opgenomen in de meeste statistische pakketten. In dit artikel worden beide technieken aan de hand van een concreet voorbeeld uit de modesector stap voor stap verkend. Er wordt veel aandacht besteed aan het doel van de technieken, de interpretatiemogelijkheden en de mogelijke valkuilen. Ook wordt aangetoond hoe beide technieken makkelijk gecombineerd kunnen worden.

Bibliografie

ANDERBERG, M.R. (1973), *Cluster analysis for applications*, New York, Academic Press.

ARABIE, P. en J. CARROLL (1980), "Mapclus: a mathematical programming approach to fitting the adclus model", *Psychometrika*, 45, blz. 211-235.

ARABIE, P. J. CARROLL, W. DE SARBO en J. WIND (1981), "Overlapping clustering: a new method for product positioning", *Journal of Marketing Research*, 18, blz. 310-317.

BENZÉCRI, J.P. (1969), "Statistical analysis as a tool to make patterns emerge from data", in: S. WATANABE, ed., *Methodologies of pattern recognition*, blz. 35-74, Academic Press, New York.

CARROLL, J. en P. GREEN (1988), "An Indscal-bases approach to multiple correspondence analysis", *Journal of Marketing Research*, 25, blz. 193-203.

CARROLL, J., P. GREEN en C. SCHAFFER (1986), "Interpoint distance comparisons in correspondence analysis", *Journal of Marketing Research*, 23, blz. 271-280.

CARROLL, J., P. GREEN en C. SCHAFFER (1987), "Comparing interpoint distances in correspondence analysis: A clarification", *Journal of Marketing Research*, 24, blz. 445-450.

DE PELSMACKER, P. en P. VAN KENHOVE (1994), *Marktonderzoek: methoden en toepassingen*, Garant, Leuven.

EVERITT, B. (1974), *Cluster analysis*, Londen, Heinemann Educational Books.

GREENACRE, M. (1984), *Theory and applications of correspondence analysis*, Londen, Academic Press.

HOFFMAN, D. en G. FRANKE (1986), "Correspondence analysis: graphical representation of categorical data in market research", *Journal of Marketing Research*, 23, blz. 213-227.

ITCB (1988), *Marktsegmentatie van de dames- en herenkledingmarkt*, Brussel, Instituut voor Textiel en Confectie, 7 delen.

KAUFMAN, L. en P. ROUSSEEUW (1989), *Finding groups in data: An introduction to cluster analysis*, New York, John Wiley & Sons.

KLASTORIN, T.D. (1983), "Assessing cluster analysis results", *Journal of Marketing Research*, 20, blz. 92-98.

KRUSKAL, J. en M. WISH (1978), *Multidimensional scaling*, Londen, Sage Publications.

LONG, J.S., (1983), *Confirmatory factor analysis: A preface to Lisrel*, Londen, Sage Publications.

MILLIGAN, G.W. (1980), "An examination of the effect of six types of error perturbation on fifteen clustering algorithms", *Psychometrika*, 45, blz. 325-342.

MILLIGAN, G.W. (1981), "A Monte Carlo study of thirty internal criterion measures for cluster analysis", *Psychometrika*, 46, blz. 187-199.

MILLIGAN, G.W. en M. COOPER (1985), "An examination of procedures for determining the number of clusters in a data set", *Psychometrika*, 50, blz. 159-179.

PUNJ, G.N. en D.W. STEWART (1983), "Cluster analysis in marketing research: review and suggestions for application", *Journal of Marketing Research*, 20, blz. 134-148.

SLM (1989), Marktonderzoeksproject uitgevoerd voor een kledingfabrikant, Vakgroep Marketing, Universiteit Gent.

SLM (1992), Marktonderzoeksproject uitgevoerd voor een koekjesfabrikant, Vakgroep Marketing, Universiteit Gent.

VAN KENHOVE, P. (1989), *Marktstructuuranalyse, een integratie van marktsegmentatie en marktafbakening: een constitutief en operationeel onderzoeksopzet*, Vakgroep Marketing, Universiteit Gent.

VOLLE, M. (1985), *Analyse des données*, Parijs, Economica.

WISHART, D. (1987), *Clustan user manual*, University of St. Andrews, Schotland.

Abstract

Cluster Analysis and Correspondence Analysis as Tools for Market Segmentation Research: An Illustration

In this paper a practical example taken from reality illustrates the application of two important multivariate analysis techniques (cluster analysis and correspondence analysis) in market segmentation research. This kind of multivariate analysis consists of different steps to be taken. These steps are first discussed in theory and then applied. Attention is being paid to possible pitfalls.