stichting
mathematisch
centrum

MC

AFDELING MATHEMATISCHE STATISTIEK SN 5/76 APRIL

J.G. BETHLEHEM

HANDLEIDING VOOR HIËRARCHISCHE CLUSTERANALYSE

2e boerhaavestraat 49 amsterdam

Printed at the Mathematical Centre, 49, 2e Boerhaavestraat, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O), by the Municipality of Amsterdam, by the University of Amsterdam, by the Free University at Amsterdam, and by industries.

AMS(MOS) subject classification scheme (1970): 62H30

Handleiding voor hiërarchische clusteranalyse

door

J.G. Bethlehem

SAMENVATTING

Van een verzameling objecten is de onderlinge gelijkenis gemeten en vastgelegd in een associatie- of dissociatiematrix. Nadat een aantal mogelijke maten is behandeld om deze gelijkenis te meten, worden zeven verschillende methoden besproken om uitgaande van een dergelijke matrix een hiërarchische boom van groepen (clusters) op te bouwen, waarbij groepen op een hoger niveau zijn ontstaan door samenvoeging van de meest gelijkende groepen op een lager niveau. Een groep bevat dan steeds objecten die het meest op elkaar lijken.

TREFWOORDEN: *Hierarchical cluster analysis*

INHOUD:

1. CLUSTERANALYSE

Een van de meest primitieve en algemene aktiviteiten van de mens is het sorteren van dingen in categorieën. Ook in wetenschappelijk onderzoek komt het regelmatig voor dat men een (grote) hoeveelheid objecten aan de hand van onderzoeksresultaten wil opdelen in een (klein) aantal zo homogeen mogelijke groepen. Men hoopt dan dat zo de gemeenschappelijke kenmerken die sommige objecten bezitten, aan het licht komen. Clusteranalyse is een techniek om een verzameling objecten op te delen in een aantal homogene groepen (clusters).

We gaan uit van een situatie waarin aan n onderzoekseenheden, p kwalitatieve, kwantitatieve of dichotome kenmerken (variabelen) zijn gemeten. Als nu x_{ij} de waarde is die variabele j heeft voor onderzoekseenheid i, dan kunnen we de uitkomsten van het experiment samenvatten in onderstaande datamatrix:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

Er zijn nu twee benaderingswijzen mogelijk:

1. Aan de hand van de metingen aan de variabelen worden de onderzoekseenheden opgedeeld in clusters.
2. Aan de hand van de metingen aan de onderzoekseenheden worden de variabelen opgedeeld in clusters.

Beide benaderingen zijn mogelijk bij de speciale vorm van clusteranalyse (hiërarchische clusteranalyse) die in dit rapport besproken zal worden. Wanneer in het vervolg over clusteren van objecten wordt gesproken, dan kunnen daarmee zowel variabelen als onderzoekseenheden bedoeld zijn.

DEFINITIE 1.1. $O = \{o_1,\ldots,o_n\}$ is de verzameling van te clusteren objecten.

2. MATEN

Om objecten te kunnen clusteren zullen we op een of andere manier de gelijkenis van twee objecten moeten kunnen meten en die in een getalwaarde uitdrukken. Vele soorten maten zijn hiervoor in gebruik. Ze kunnen worden onderscheiden in twee groepen: dissociatie- of afstandsmaten en associatiematen.

DEFINITIE 2.1. Een *dissociatie- of afstandsmaat* is een functie $d: O \times O \to \mathbb{R}$, die aan het paar $(o_i,o_j) \in O \times O$ het getal d_{ij} toevoegt, waarvoor geldt:

1. $d_{ij} \geq 0 \qquad \forall i,j$
2. $d_{ij} = d_{ji} \qquad \forall i,j$
3. $d_{ii} = 0 \qquad \forall i$

Soms wordt nog een vierde eis opgelegd:

4. $d_{ij} \leq d_{ik} + d_{kj} \qquad \forall i,j,k$ (Driehoeksongelijkheid)

Deze eis wordt in het bijzonder opgelegd, wanneer men de gebruikte maat een afstand wil noemen.

DEFINITIE 2.2. Een *associatiemaat* is een functie $s: O \times O \to \mathbb{R}$ die aan het paar $(o_i,o_j) \in O \times O$ het getal s_{ij} toevoegt waarvoor geldt:

1. $|s_{ij}| \leq 1 \qquad \forall i,j$
2. $s_{ij} = s_{ji} \qquad \forall i,j$
3. $s_{ii} = 1 \qquad \forall i$

Voorbeelden van associatiematen zijn de diverse correlatie coëfficiënten.

In het vervolg zullen nog verscheidene maten de revue passeren. Ook zal nog worden gesproken over conversie van dissociatie- en associatiematen.

3. HIËRARCHISCHE CLUSTERANALYSE

Dit rapport zal alleen gaan over hiërarchische clusteranalyse. Dat is een stapsgewijs proces waarin een hiërarchische boom van clusters wordt opgebouwd, zoals hieronder nog verder uitgelegd zal worden. Voor andere clusteranalyse methoden kan men bijvoorbeeld kijken in EVERITT [2].

DEFINITIE 3.1. Een clustering is een opdeling van $O = \{o_1,\ldots,o_n\}$ in onderling disjuncte deelverzamelingen (clusters). Voor de clustering Γ_k in de $k^{\underline{de}}$ fase van het proces gebruiken we de volgende notatie:

$$\Gamma_k = \{C_i^{(k)} : \bigcup_{i=1}^{n-k+1} C_i^{(k)} = O,\ C_i^{(k)} \cap C_j^{(k)} = \emptyset \quad \text{als } i \neq j\}$$

Het clusteranalyse-proces start in fase 1 waarin er n clusters $C_1^{(1)},\ldots, C_n^{(1)}$ zijn, die ieder precies één object bevatten: $C_i^{(1)} = \{o_i\}$. In de $k^{\underline{de}}$ stap van het proces wordt nu de clustering Γ_{k+1} gevormd door van clustering Γ_k die twee clusters $C_i^{(k)}$ en $C_j^{(k)}$ te zoeken en samen te voegen die het meest op elkaar lijken. In iedere fase vermindert het aantal clusters dus met één, totdat we uiteindelijk na de $n-1^{\underline{ste}}$ stap clustering Γ_n krijgen die uit nog maar één cluster $C_1^{(n)}$ bestaat, die alle objecten bevat. Aangezien iedere cluster van Γ_{k+1} bevat is in een cluster van Γ_k kunnen we inderdaad spreken van een hiërarchische boom van clusters.
Behalve dat bij hiërarchische clusteranalyse de samenhang tussen objecten gemeten moet worden, is het ook noodzakelijk te weten in hoeverre twee clusters van een bepaalde clustering op elkaar lijken. Hiervoor worden ook weer dissociatie- en associatiematen gebruikt.

DEFINITIE 3.2. Een *dissociatie- of afstandsmaat op de clusters* is een reële functie D, die voor iedere $C_i, C_j \subset O$ met $C_i \cap C_j = \emptyset$ voldoet aan:

1. $D(C_i,C_j) \geq 0$
2. $D(C_i,C_j) = D(C_j,C_i)$
3. $D(C_i,C_i) = 0$

(4. $D(C_i,C_j) \leq D(C_i,C_k) + D(C_k,C_j)$ voor $C_k \cap C_i = \emptyset,\ C_k \cap C_j = \emptyset$)

DEFINITIE 3.3. Een *associatiemaat op de clusters* is een reële functie S, die voor iedere $C_i,C_j \subset O$ met $C_i \cap C_j = \emptyset$ voldoet aan:

1. $|S(C_i,C_j)| \leq 1$
2. $S(C_i,C_j) = S(C_j,C_i)$
3. $S(C_i,C_i) = 1$

We kunnen het verloop van het hiërarchische clusteringsproces nu in het volgende schema weergeven:

Schema 3.4.

3.4.1. Initialisatie:

$\Gamma_1 \overset{\text{def}}{=} \{\{o_i\} : 1 \leq i \leq n\}$

$D(C_i^{(1)}, C_j^{(1)}) \underline{\underline{\text{def}}} d_{ij}$ resp. $S(C_i^{(1)}, C_j^{(1)}) \underline{\underline{\text{def}}} s_{ij}$

In de eerste stap bestaan de clusters ieder precies uit één object. De afstanden tussen de clusters zijn gelijk aan de afstanden tussen de objecten.

3.4.2. Zoek op elkaar lijkende clusters:

Bepaal die clusters $C_i^{(k)}$ en $C_j^{(k)}$ uit Γ_k waarvoor $D(C_i^{(k)}, C_j^{(k)})$ de kleinste resp. $|S(C_i^{(k)}, C_j^{(k)})|$ de grootste waarde heeft. (Wat de consequenties zijn als meerdere paren voldoen, wordt op blz. 7. besproken).

3.4.3. Voeg samen:

Voeg $C_i^{(k)}$ en $C_j^{(k)}$ samen en vorm zo de nieuwe clustering

Γ_{k+1} : (zeg $i < j$)

$C_1^{(k+1)} = C_1^{(k)}$ voor $1 \leq 1 < i$

$C_1^{(k+1)} = C_i^{(k)} \cup C_j^{(k)}$ voor $1 = i$

$C_1^{(k+1)} = C_1^{(k)}$ voor $i < 1 < j$

$C_1^{(k+1)} = C_{1+1}^{(k)}$ voor $j \leq 1 \leq n - k$.

3.4.4. Bepaal nieuwe afstanden:

Bereken de afstanden van de nieuwe cluster tot de oude clusters. De overige afstanden blijven hetzelfde.

3.4.5. Herhaal:

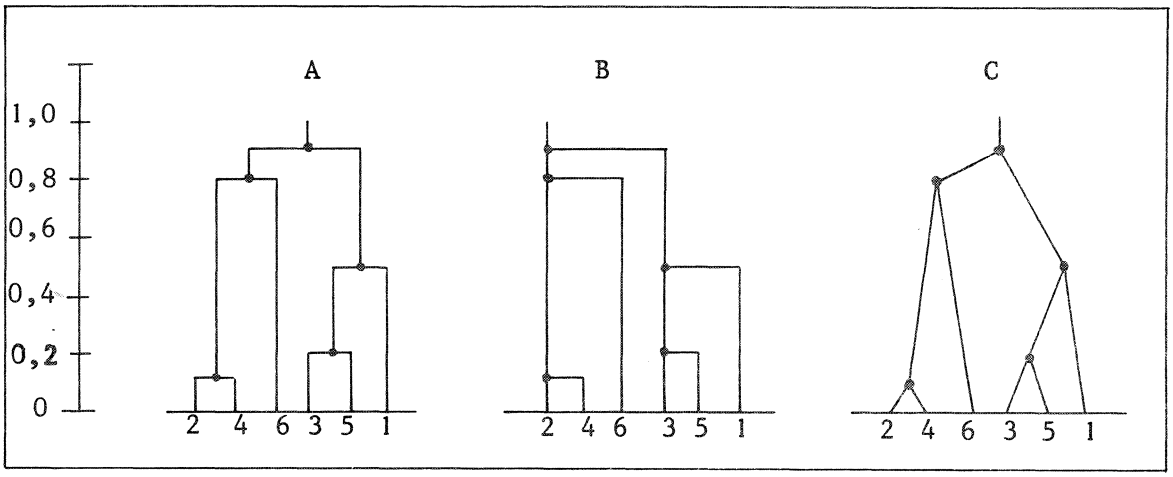
Herhaal de stappen 3.4.2., 3.4.3. en 3.4.4. zolang er nog minstens twee clusters zijn. Als er nog maar één cluster over is, dan is het proces klaar.

Het verloop van een hiërarchisch clusteringsproces kan heel illustratief worden weergegeven in een dendrogram. Dit is een diagram waarin horizontaal de objecten staan afgezet en vertikaal de dissociatie/associatiemaat. Een cluster wordt aangegeven door een vertikale lijn. Als twee clusters worden samengevoegd, dan worden de vertikale lijnen verbonden door een horizontale lijn. Hierboven loopt dan nog slechts één vertikale lijn die de samengevoegde cluster symboliseert.

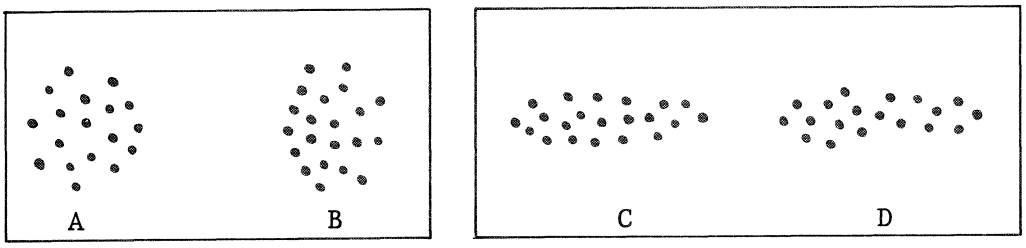
Voorbeeld 3.5. Stel dat we door het uitvoeren van hiërarchische clusteranalyse op 6 objecten $o_1,\ldots,o_6$ het volgende resultaat hebben gekregen:

STAP	MEEST GELIJKENDE CLUSTERS	AFSTAND TUSSEN DIE CLUSTERS	NIEUWE CLUSTERING
1	$\{o_2\}$, $\{o_4\}$	0,1	$\{o_1\},\{o_2,o_4\},\{o_3\},\{o_5\},\{o_6\}$
2	$\{o_3\}$, $\{o_5\}$	0,2	$\{o_1\},\{o_2,o_4\},\{o_3,o_5\},\{o_6\}$
3	$\{o_1\}$, $\{o_3,o_5\}$	0,5	$\{o_1,o_3,o_5\},\{o_2,o_4\},\{o_6\}$
4	$\{o_2,o_4\}$, $\{o_6\}$	0,8	$\{o_1,o_3,o_5\},\{o_2,o_4,o_6\}$
5	$\{o_1,o_3,o_5\}$, $\{o_2,o_4,o_6\}$	0,9	$\{o_1,o_2,o_3,o_4,o_5,o_6\}$

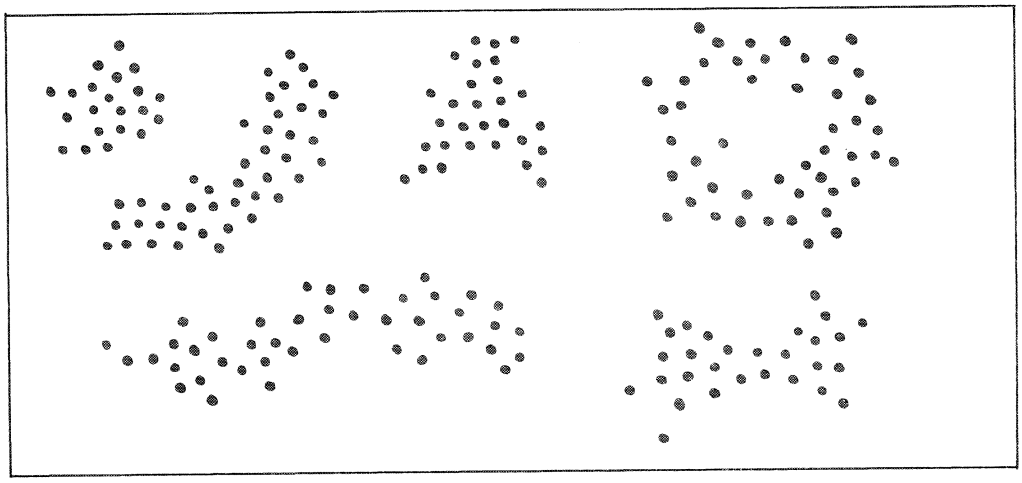
In figuur A staat het dendrogram van dit proces. Figuur B en C geven het proces weer in een andere stijl.



Er zijn vele manieren om dissociatie- of associatiematen tussen de clusters te definiëren. Iedere definitie bepaalt een andere methode van hiërarchische clusteranalyse. Ze leveren alle in mindere of meerdere mate verschillende resultaten op. Dit komt doordat niet iedere methode dezelfde vorm van cluster kan ontdekken. Het gebruik van een bepaalde methode houdt in feite een a priori veronderstelling in over de vorm van de clusters die men zoekt. Heeft men bijvoorbeeld een methode die twee clusters alleen maar samenvoegt als alle elementen van die twee clusters dicht genoeg bij elkaar liggen, dan is die methode in het algemeen geschikt om bolvormige clusters te vinden. In onderstaande figuur zullen dus clusters A en B eerder samengevoegd worden dan clusters C en D.



We kunnen stellen dat de huidige methoden over het algemeen niet in staat zijn al die patronen als cluster te onderkennen die de mens zich visualiseert als een cluster, namelijk als een continu gebied van hogere dichtheid in een p-dimensionale ruimte.
Voorbeelden van natuurlijke clusters zijn:



De methoden, die besproken zullen worden, zijn alle geïmplementeerd in computerprogramma's. Ze gaan er van uit dat er altijd een unieke kleinste afstand tussen de clusters te vinden is. In de praktijk kan het echter voorkomen dat er meerdere gelijke waarden zijn. De computer zal dan volgens een of ander willekeurig gekozen criterium er twee clusters uitkiezen en die samenvoegen. Dit betekent dus dat als oorspronkelijk de variabelen in een andere volgorde stonden, de resultaten ook anders zouden zijn geweest. De gebruiker zij dus gewaarschuwd voorzichtig te zijn met het toepassen van hiërarchische clusteranalyse als hij veel gelijke afstanden heeft tussen de objecten.

Hiërarchische clusteranalyse is oorspronkelijk afkomstig uit de biologie, waarin men bijvoorbeeld een hiërarchie van plantensoorten of diersoorten wilden bepalen. Men was dus geïnteresseerd in de hele hiërarchische structuur. Niettemin kan deze vorm van clusteranalyse ook goed toegepast worden wanneer men slechts belangstelling heeft voor één opdeling van de verzameling van objecten. De onderzoeker zal dan zelf een criterium moeten bepalen om een bepaalde clustering uit de hiërarchie te kiezen. Het kan bijvoorbeeld zijn dat hij van te voren al een idee heeft van het aantal clusters of hij kan zoeken naar een punt in het proces waarop de afstanden tussen de samen te voegen clusters plotseling veel groter worden. Hij kan ook van te voren al vastleggen dat de samen te voegen clusters niet verder dan een bepaalde waarde van elkaar af mogen liggen. Dit zijn slechts suggesties en andere criteria zijn denkbaar. Er is geen algemene regel aan te geven.

Voor de volledigheid moet nog worden opgemerkt dat er ook nog hiërarchische clustermethoden zijn die in omgekeerde richting werken. Ze beginnen met één cluster die alle objecten bevat en proberen achtereenvolgens de bestaande clusters op te delen in zo homogeen mogelijke nieuwe clusters. Dit proces gaat door tot er precies n clusters zijn, die ieder precies één object bevatten. Deze methoden zullen hier niet behandeld worden. De lezer wordt hiervoor verwezen naar EVERITT [2] en ANDERBERG [1].

Aangezien zich bij het clusteren van variabelen toch wel andere problemen voordoen dan bij het clusteren van onderzoekseenheden zullen beide

technieken in aparte paragrafen behandeld worden. Daarna zal een beschrijving worden gegeven van een aantal bestaande hiërarchische methoden.

4. CLUSTEREN VAN ONDERZOEKSEENHEDEN

We gaan uit van de datamatrix

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

waarin x_{ij} de waarde die variabele j heeft voor onderzoekseenheid i.

Als we de onderzoekseenheden willen clusteren, dan moeten we deze onderzoekseenheden kunnen vergelijken aan de hand van de waarden van de variabelen, m.a.w. als we de onderzoekseenheden i en k willen vergelijken dan moeten we nagaan in hoeverre de vectoren $(x_{i1} \ \cdots \ x_{ip})$ en $(x_{k1} \ \cdots \ x_{kp})$ op elkaar lijken. Daarbij kunnen zich verschillende problemen voordoen:

1. Moeten we elke variabele even zwaar mee laten tellen of vinden we bepaalde variabelen zo belangrijk dat we ze extra gewicht willen geven?
2. Wat doen we als we allerlei verschillende typen variabelen hebben?

Over het wegen van variabelen kan in het algemeen weinig worden gezegd. Dit hangt uitsluitend af van de situatie waarin de onderzoeker zich bevindt. Een variabele die hij belangrijk acht, zal hij zwaarder willen laten meetellen dan een andere variabele. Als hij eigenlijk het idee heeft, dat alle variabelen ongeveer even belangrijk zijn, dan kan hij er over denken zijn variabelen te standaardizeren, waardoor iedere variabele hetzelfde waardebereik krijgt.

Wat betreft het tweede punt kunnen we een opdeling van de variabelen maken in vier groepen:

1. *Nominale variabelen* classificeren de metingen in groepen, waarbij in een groep alle metingen, die het zelfde zijn, samengebracht worden.

Als twee metingen in verschillende groepen terechtkomen, dan betekent dit alleen dat de twee bijbehorende onderzoekseenheden, althans wat deze variabele betreft, verschillen; het zegt niets over de aard van het verschil. Vaak worden met de zo ontstane groepen getallen geassocieerd om de groepen aan te duiden. Deze getallen hebben geen ander doel dan te dienen als naamplaatje en het zou dus absurd zijn er berekeningen mee uit te voeren. De enige toegestane operaties zijn de (logische) operaties = en $\neq$. Een voorbeeld van een nominale variabele is de haarkleur van de proefpersoon.

Een speciaal geval van de nominale variabele is de *dichotome* of *binaire variabele*. Een dichotome variabele meet slechts aan- of afwezigheid van een kenmerk. Als waardebereik zijn dus slechts twee categoriën mogelijk, die meestal aangeduid worden met 1 (aanwezig) of 0 (afwezig): Een voorbeeld van een dichotome variabele is het al of niet roker zijn van de proefpersoon.

2. *Ordinale variabelen* gaan een stapje verder. Niet alleen worden gelijke metingen in dezelfde groep ondergebracht, maar bovendien is er een zekere ordening in deze groepen met betrekking tot een bepaalde karakteristieke eigenschap, die de groepen bezitten. Aan de groep kunnen getallen toegekend worden, die de ordening symboliseren. Een meting aan een onderzoekseenheid heeft dan tot resultaat het getal dat met de betreffende groep is geassocieerd. Op deze getallen mogen de operaties =, $\neq$, <, $\leq$, > en $\geq$ worden toegepast. Een voorbeeld van een ordinale variabele is het oordeel van de proefpersoon over een schilderij, met als mogelijke antwoorden: 1 (lelijk), 2 (onverschillig) en 3 (mooi).

3. *Interval-variabelen* meten meer dan ordinale variabelen, want niet alleen wordt een ordening aangebracht tussen de groepen, maar het is ook mogelijk de verschillen tussen de groepen in een getalwaarde uit te drukken. Om dit te kunnen doen moet er een of andere meeteenheid gedefinieerd zijn. Bij interval-variabelen mogen we dus de operaties =, $\neq$, <, $\leq$, >, $\geq$, + en - toepassen. Een voorbeeld van een intervalvariabele is de temperatuur in ^{o}F.

4. *Ratio-variabelen* zijn interval-variabelen, waarbij bovendien nog een nulpunt gedefinieerd is, zodat we kunnen zeggen dat een bepaalde meetuitkomst een bepaald aantal malen groter is dan een andere meetuitkomst. Buiten de bij de interval-variabelen genoemde operaties, kunnen dus ook nog de operaties × en / gebruikt worden. Een voorbeeld van een ratio-variabele is de lengte van de proefpersoon.

Als we nu een maat definiëren om de gelijkenis te meten tussen $(x_{i1} \ldots x_{ip})$ en $(x_{k1} \ldots x_{kp})$, dan moeten we precies aangeven wat er gebeurt met elk paar (x_{ij}, x_{kj}) uit de beide vectoren.

Als we bijvoorbeeld als maat nemen de som van de absolute waarden van de verschillen,

$$\sum_{j=1}^{p} |x_{ij} - x_{kj}|,$$

dan is dat alleen maar mogelijk indien alle variabelen interval- of ratio-variabelen zijn.
Hieronder zullen we nu voor elk type variabele een aantal maten bespreken.

A. Dichotome variabelen

We nemen aan dat alle p variabelen dichotoom zijn, waarbij 1 de aanwezigheid en 0 de afwezigheid van een bepaald kenmerk aangeeft. Voor de vergelijking van de onderzoekseenheden i en k wordt eerst de volgende tabel gemaakt:

i \ k	1	0	
1	a_{ik}	b_{ik}	$a_{ik} + b_{ik}$
0	c_{ik}	d_{ik}	$c_{ik} + d_{ik}$
	$a_{ik} + c_{ik}$	$b_{ik} + d_{ik}$	p

a_{ik} = aantal malen dat zowel onderzoekseenheid i als k een 1 scoort voor de p variabelen;

b_{ik} = aantal malen dat eenheid i een 1 scoort en tegelijk eenheid k een 0;

c_{ik} = aantal malen dat eenheid i een 0 scoort en tegelijk eenheid k een 1;

d_{ik} = aantal malen dat zowel eenheid i als k tegelijk een 0 scoren.

In formule:

$$a_{ik} = \sum_{j=1}^{p} x_{ij} \cdot x_{kj}; \quad b_{ik} = \sum_{j=1}^{p} x_{ij} - a_{ik}$$

$$c_{ik} = \sum_{j=1}^{p} x_{kj} - a_{ik}; \quad d_{ik} = p - (a_{ik}+b_{ik}+c_{ik})$$

De meeste maten voor dichotome variabelen zijn gebaseerd op de grootheden a_{ik}, b_{ik}, c_{ik} en d_{ik}. De verschillende mogelijkheden ontstaan al naar de antwoorden op de volgende twee vragen:

1. Moeten paren (1,1) (dubbel positief) even zwaar meetellen als paren (0,0) (dubbel negatief)?
2. Moeten ongelijke paren (0,1) en (1,0) even zwaar meetellen als gelijke paren ((1,1) en (0,0))?

De volgende associatiematen worden veel gebruikt:

4.1. $$s_{ik} = \frac{a_{ik}}{p}$$ (Russell & Rao)

De waarde wordt bepaald door verhouding van aantal dubbel positieve paren en totaal andere paren.

4.2. $$s_{ik} = \frac{a_{ik} + d_{ik}}{p}$$ (simple matching)

De waarde wordt bepaald door verhouding van aantal gelijke paren en totaal aantal paren.

4.3. $$s_{ik} = \frac{a_{ik}}{a_{ik} + b_{ik} + c_{ik}}$$ (Jaccard)

De waarde wordt bepaald door verhouding van aantal positieve paren en totaal aantal paren. Dubbel negatieve paren worden echter genegeerd.

4.4. $$s_{ik} = \frac{2(a_{ik}+d_{ik})}{2(a_{ik}+d_{ik}) + b_{ik} + c_{ik}}$$

De waarde wordt bepaald door verhouding van aantal gelijke paren en totaal aantal paren. In beide aantallen worden gelijke paren echter dubbel meegeteld.

4.5. $$s_{ik} = \frac{2a_{ik}}{2a_{ik} + b_{ik} + c_{ik}}$$ (Dice)

De waarde wordt bepaald door verhouding van aantal dubbel positieve paren en totaal aantal paren. Dubbel negatieve paren worden genegeerd en dubbel positieve paren tellen dubbel mee.

4.6. $$s_{ik} = \frac{a_{ik}}{a_{ik} + 2(b_{ik}+c_{ik})}$$

De waarde wordt bepaald door verhouding van aantal dubbel positieve paren en totaal aantal paren. Dubbel negatieve paren worden genegeerd en ongelijke paren tellen dubbel mee.

Nog meer voorstellen voor associatiematen tussen dichotome variabelen staan bijvoorbeeld in ANDERBERG [1].

B. Nominale variabelen

We kunnen nominale variabelen op dezelfde manier behandelen als dichotome variabelen. Daarbij interpreteren we een eventueel aanwezig zijnde categorie "niet van toepassing" op dezelfde manier als afwezigheid van het kenmerk bij dichotome variabelen. Voor de vergelijking van onderzoekseenheden i en k bepalen we dan eerst de volgende grootheden:

u_{ik} = aantal malen dat onderzoekseenheden i en k voor de p variabelen gelijke antwoorden opleveren. (correspondeert met $a_{ik} + d_{ik}$).

v_{ik} = aantal malen dat i en k tegelijk "niet van toepassing" opleveren. (correspondeert met d_{ik}).

w_{ik} = aantal malen dat eenheden i en k een verschillend resultaat opleveren. (correspondeert met $b_{ik} + c_{ik}$).

Analoog aan 4.1 tot en met 4.6 worden nu de volgende maten voorgesteld:

4.7. $$s_{ik} = \frac{u_{ik} - v_{ik}}{u_{ik} + w_{ik}}$$

4.8. $$s_{ik} = \frac{u_{ik}}{u_{ik} + w_{ik}}$$

4.9. $$s_{ik} = \frac{u_{ik} - v_{ik}}{u_{ik} - v_{ik} + w_{ik}}$$

4.10. $$s_{ik} = \frac{2u_{ik}}{2u_{ik} + w_{ik}}$$

4.11. $$s_{ik} = \frac{2(u_{ik}-v_{ik})}{2(u_{ik}-v_{ik}) + w_{ik}}$$

4.12. $$s_{ik} = \frac{u_{ik} - v_{ik}}{u_{ik} - v_{ik} + 2w_{ik}}$$

Een andere methode krijgt men door de nominale variabele (die m waarden aan kan nemen) op te splitsen in m dichotome variabelen. Als een meting dan in de i^{de} categorie terecht komt dan betekent dit dat de i^{de} dichotome variabele de waarde 1 krijgt en de overige m-1 dichotome variabelen de waarde 0. De bijdrage $s_{ik}^{(j)}$ die de j^{de} nominale variabele levert aan de 2×2-tabel is dan:

4.13. Als de onderzoekseenheid i en k hetzelfde scoren voor variabele j:

$s_{ik}^{(j)}$:

i \ k	1	0	
1	1	0	1
0	0	m-1	m-1
	1	m-1	m

Als onderzoekseenheid i en k verschillend scoren:

$s_{ik}^{(j)}$:

i \ k	1	0	
1	0	1	1
0	1	m-2	m-1
	1	m-1	m

Voor alle nominale variabelen worden zo de bijdragen in deze tabel verzameld en vervolgens kan men een van de associatiematen voor dichotome variabelen gebruiken. Opgemerkt dient te worden dat in feite alleen de dubbel positieve paren van belang zijn, aangezien het aantal dubbel negatieve paren plus het aantal dubbel positieve paren altijd het aantal categoriën van de nominale variabele oplevert. Verder is het zo, dat de combinatie (1,1) een bijdrage 1 aan a_{ik} levert, terwijl de combinatie (0,1) of (1,0) een bijdrage 2 levert aan $b_{ik} + c_{ik}$.

C. Ordinale variabelen

Voor ordinale variabelen kunnen in principe dezelfde methoden gebruikt worden als bij nominale variabelen. Helaas wordt dan niet meer gebruik gemaakt van de extra informatie die zich in de ordening van de uitkomsten bevindt. Om daaraan tegemoet te komen hebben SOKAL & SWEATH [18] het onderstaande additatieve coderingsschema opgesteld. Ze gaan er daarbij vanuit dat wanneer twee metingen veel in rang verschillen, ze ook in werkelijkheid meer verschillen dan wanneer ze in rang dichter bij elkaar liggen.

4.14. Stel de mogelijke waarden van de ordinale variabele j zijn geordend van 1 tot en met m. De ordinale variabele wordt nu omgezet in m-1 dichotome variabelen. Als de ordinale variabele de waarde i heeft dan betekent dit dat de eerste i-1 dichotome variabelen de waarde 1 krijgen en de overige m-i dichotome variabelen de waarde 0.

waarde van ordinale variabele	dichotome variabelen 1	2	3	...	m-1
1	0	0	0	...	0
2	1	0	0	...	0
3	1	1	0	...	0
⋮	⋮				⋮
m	1	1	1	...	1

Alle dichotome variabelen worden nu weer verwerkt in een 2×2-tabel. Als nu $r = |x_{ij} - x_{kj}|$ het absolute verschil in rang is van onderzoeks-eenheden i en k voor variabele j, dan is de bijdrage aan $(a_{ik} + d_{ik})$ gelijk aan m-r-1 en de bijdrage aan $(b_{ik} + c_{ik})$ gelijk aan r. Het is duidelijk dat hier de (1,1)-paren even zwaar moeten meetellen als de (0,0)-paren. Om de bijdrage onafhankelijk te maken van het aantal categorieën van de ordinale variabele, wordt nog gedeeld door m-1, zodat:

bijdrage aan $a_{ik} + d_{ik}$ van variabele j: $1 - \frac{r}{m-1}$
bijdrage aan $b_{ik} + c_{ik}$ van variabele j: $\frac{r}{m-1}$

D. Interval- en ratio-variabelen

Voor interval- en ratio-variabelen wordt traditioneel veel gewerkt met dissociatiematen. De belangrijkste dissociatiemaat is de Minkowski-afstand:

$$4.15. \qquad d_{ik} = \left(\sum_{j=1}^{p} |x_{ij} - x_{kj}|^r \right)^{1/r} , \quad r \geq 1$$

Voor het geval r = 1 krijgen we de absolute of cityblock-afstand:

$$4.16. \qquad d_{ik} = \sum_{j=1}^{p} |x_{ij} - x_{kj}|$$

en voor het geval r = 2 de Euclidische afstand:

4.17. $$d_{ik} = \sqrt{\sum_{j=1}^{p} (x_{ij}-x_{kj})^2}$$

Bij deze twee variabelen speelt het wegingsproces een rol. Immers als een variabele een groot waardebereik heeft, dan kan zo'n variabele een belangrijke bijdrage aan de afstand leveren.

Alhoewel het voor het clusteren van onderzoekseenheden weinig voor de hand liggend lijkt, wordt als associatiemaat ook nog wel eens de product-moment correlatiecoëfficiënt tussen de onderzoekseenheden genomen. Dit vindt zijn verklaring in het volgende:

Laten $\bar{x}_i = \frac{1}{p} \sum_{j=1}^{p} x_{ij}$ en $\bar{x}_k = \frac{1}{p} \sum_{j=1}^{p} x_{kj}$ de gemiddelde scores

zijn voor onderzoekseenheden i en k en stel:

$$L_i = \sqrt{\sum_{j=1}^{p} (x_{ij}-\bar{x}_i)^2} \quad \text{en} \quad L_k = \sqrt{\sum_{j=1}^{p} (x_{kj}-\bar{x}_k)^2} \quad \text{zijn}$$

de lengtes van de vectoren $(x_{i1} \ldots x_{ip})$ en $(x_{k1} \ldots x_{kp})$. Als we de scores nu zodanig schalen dat de gemiddelde score 0 en de lengte van de vector 1 is, dus $x_{ij} \rightarrow \frac{x_{ij} - \bar{x}_i}{L_i}$, dan geldt voor de Euclidische afstand d_{ik} van twee zo geschaalde onderzoekseenheden i en k:

$$d_{ik}^2 = \sum_{j=1}^{p} \left(\frac{x_{ij}-\bar{x}_i}{L_i} - \frac{x_{kj}-\bar{x}_k}{L_k} \right)^2 =$$

$$= \frac{1}{L_i^2 L_k^2} \left(L_i^2 L_k^2 - 2L_i L_k \sum_{j=1}^{p} (x_{ij}-\bar{x}_i)(x_{kj}-\bar{x}_k) + L_i^2 L_k^2 \right) =$$

$$= 2\left(1 - \frac{\sum_{j=1}^{p} (x_{ij}-\bar{x}_i)(x_{kj}-\bar{x}_k)}{L_i L_k} \right) = 2(1 - s_{ik}),$$

waarbij s_{ik} de product-moment correlatiecoëfficiënt is.

E. gemengde variabelen

GOWER [7] stelt een algemene associatiemaat voor, die voor alle typen variabelen gebruikt kan worden:

$$4.18. \qquad s_{ik} = \sum_{j=1}^{p} \sigma_{ikj} \; / \; \sum_{j=1}^{p} w_{ikj}$$

Het gewicht w_{ikj} is 0 of 1 al naar gelang de vergelijking van onderzoekseenheden i en k met behulp van variabele j geldig is. Behalve in het geval van dichotome variabelen is w_{ikj} alleen 0 wanneer de waarde van variabele j onbekend is voor één of beide onderzoekseenheden. Als $w_{ikj} = 0$ dan $\sigma_{ikj} = 0$ en als $w_{ikj} = 0$ voor alle j variabelen dan is de gelijkenis tussen onderzoekseenheid i en k niet gedefinieerd, zodat we hem niet kunnen meten.
De waarden van σ_{ikj} en w_{ikj} worden verder bepaald volgens onderstaand schema:

dichotome variabele :	eenheid i	+	+	−	−
	eenheid k	+	−	+	−
	σ_{ikj}	1	0	0	0
	w_{ikj}	1	1	1	0
normale, ordinale variabele:	$\sigma_{ikj} = 1$ als $x_{ij} = x_{kj}$ $\sigma_{ikj} = 0$ als $x_{ij} \neq x_{kj}$ $w_{ikj} = 1$				
interval-, ratio-variabele:	$\sigma_{ikj} = 1 - \frac{\lvert x_{ij} - x_{kj} \rvert}{R_j}$ (R_j is range van variabele j) $w_{ikj} = 1$				

Voorbeeld. Aan 3 onderzoekseenheden zijn 5 variabelen gemeten; 1 dichotome, 2 nominale en 2 ratio-variabelen:

onderzoekseenheid	roken	oogkleur	haarkleur	lengte	gewicht
1	ja	bruin	zwart	166	20
2	ja	groen	blond	172	30
3	nee	bruin	zwart	170	50

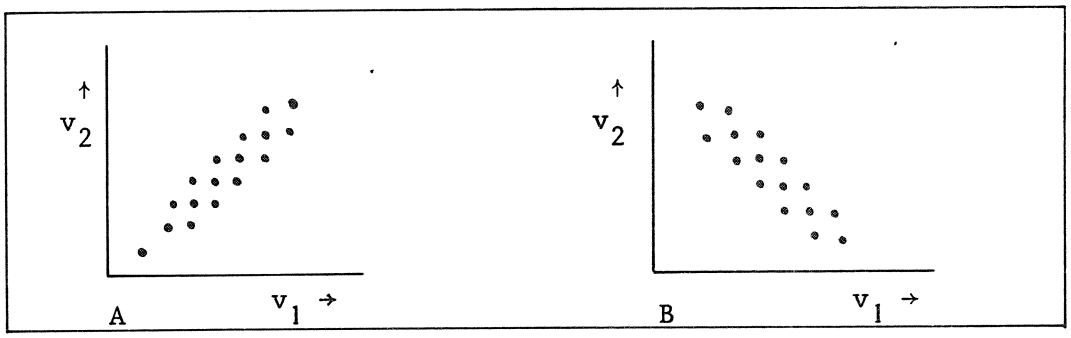
Dit levert:

i	k	σ_{ik1}	w_{ik1}	σ_{ik2}	w_{ik2}	σ_{ik3}	w_{ik3}	σ_{ik4}	w_{ik4}	σ_{ik5}	w_{ik5}	s_{ik}
1	2	1	1	0	1	0	1	0,000	1	0,667	1	0,333
1	3	0	1	1	1	1	1	0,333	1	0,000	1	0,467
2	3	0	1	0	1	0	1	0,667	1	0,333	1	0,200

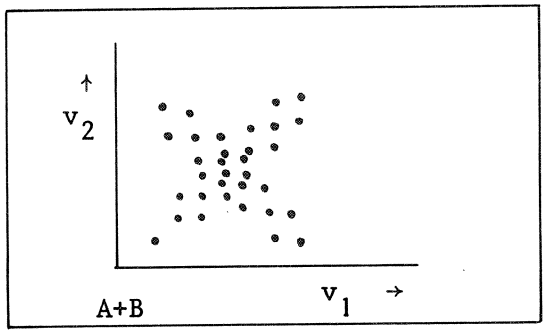
5. CLUSTEREN VAN VARIABELEN

Bij het clusteren van variabelen gaan we op analoge wijze te werk als bij het clusteren van onderzoekseenheden. Alleen proberen we nu variabelen te vergelijken aan de hand van hun waarden voor de verschillende onderzoekseenheden. Als we de variabelen clusteren dan nemen we in feite aan dat alle onderzoekseenheden uit een zelfde populatie komen, waarbij de keuze van de onderzoekseenheden niet bepalend is voor de gelijkenis van de variabelen. M.a.w. iedere aselecte steekproef uit de onderzoekseenheden zou tot ongeveer hetzelfde resultaat moeten leiden. Ter illustratie het volgende voorbeeld waarin dit niet het geval is:

Voorbeeld: Een populatie van onderzoekseenheden is onder te verdelen in twee deelpopulaties A en B. Voor iedere onderzoekseenheid zijn twee variabelen v_1 en v_2 gemeten. Het is dan mogelijk dat, indien we onderzoekseenheden in grafieken uitzetten voor A en B apart, we het volgende resultaat zien:



Hieruit zouden we moeten concluderen dat voor deelpopulatie A de variabelen v_1 en v_2 sterk op elkaar lijken en voor deelpopulatie B de variabele v_1 en de tegengestelde van v_2 sterk op elkaar lijken terwijl we, als we de populatie in zijn geheel hadden bekeken, we onderstaande grafiek gekregen zouden hebben waaruit geen enkele relatie tussen v_1 en v_2 blijkt:



Als we in het algemeen willen nagaan in hoeverre variabele j en 1 op elkaar lijken dan wordt dit meestal gedaan aan de hand van de vectoren $(x_{1j} \cdots x_{nj})$ en $(x_{11} \cdots x_{n1})$ uit de datamatrix

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

Alvorens nu over te gaan tot bespreking van een aantal maten om gelijkenis tussen variabelen te meten, moet opgemerkt worden, dat er nog andere methoden bestaan dan clusteranalyse voor het classificeren van variabelen.

Met name moet daarbij gedacht worden aan technieken als factoranalyse en principale componentenanalyse, waarbij geprobeerd wordt de variabelen samen te vatten in een kleiner aantal lineaire combinaties van deze variabelen. Zie hiervoor bijvoorbeeld MORRISON [14].

A. Dichotome variabelen

Net als bij het clusteren van onderzoekseenheden, wordt ook hier weer gebruik gemaakt van een 2×2-tabel. Voor de vergelijking van de variabelen j en 1 krijgen we:

j \ 1	1	0	
1	a_{j1}	b_{j1}	$a_{j1} + b_{j1}$
0	c_{j1}	d_{j1}	$c_{j1} + d_{j1}$
	$a_{j1} + c_{j1}$	$b_{j1} + d_{j1}$	n

$$a_{j1} = \sum_{i=1}^{n} x_{ij}\, x_{i1} \quad ; \qquad b_{j1} = \sum_{i=1}^{n} x_{ij} - a_{j1}$$

$$c_{j1} = \sum_{i=1}^{n} x_{i1} - a_{j1} \; ; \qquad d_{j1} = n - (a_{j1}+b_{j1}+c_{j1})$$

De volgende maten worden voorgesteld:

5.1. De cosinus van de hoek tussen de vectoren $(x_{1j} \ldots x_{nj})$ en $(x_{11} \ldots x_{n1})$:

$$s_{j1} = \frac{a_{j1}}{\sqrt{(a_{j1}+b_{j1})\,(a_{j1}+c_{j1})}}$$

Het is niet altijd op voorhand duidelijk aan welke uitkomst de waarde 0 en aan welke uitkomst de waarde 1 moet worden toegekend. Bij deze maat kan dit tot vervelende complicaties leiden, omdat omkering van 0 en 1 tot een andere waarde leidt:

De associatiemaat is niet symmetrisch m.b.t. 0 en 1.
Daarom wordt voorgesteld, ingeval zo'n onduidelijke situatie zich voordoet, in plaats van deze maat het kwadraat van het geometrisch gemiddelde van de mogelijke uitkomsten door verwisseling van 0 en 1 te nemen. Dit leidt tot:

5.2. $$s_{j1} = \frac{a_{j1}\, d_{j1}}{\sqrt{(a_{j1}+b_{j1})(a_{j1}+c_{j1})(b_{j1}+d_{j1})(c_{j1}+d_{j1})}}$$

De product-moment correlatiecoëfficiënt luidt voor gedichotomiseerde variabelen:

5.3. $$s_{j1} = \frac{a_{j1}\, d_{j1} - b_{j1}\, c_{j1}}{\sqrt{(a_{j1}+b_{j1})(a_{j1}+c_{j1})(b_{j1}+d_{j1})(c_{j1}+d_{j1})}}$$

De chi-kwadraat toetsingsgrootheid met correctie van Yates kan gebruikt worden, mits goed genormeerd zodanig dat $|s_{j1}| \leq 1$:

5.4. $$s_{j1} = \frac{n(|a_{j1}\, d_{j1} - b_{j1}\, c_{j1}| - n/2)^2}{(a_{j1}+b_{j1})(a_{j1}+c_{j1})(b_{j1}+d_{j1})(c_{j1}+d_{j1})}$$

Twee door Yule voorgestelde associatiematen zijn:

5.5. $$s_{j1} = \frac{a_{j1}\, d_{j1} - b_{j1}\, c_{j1}}{a_{j1}\, d_{j1} + b_{j1}\, c_{j1}}$$

5.6. $$s_{j1} = \frac{\sqrt{a_{j1}d_{j1}} - \sqrt{b_{j1}c_{j1}}}{\sqrt{a_{j1}d_{j1}} + \sqrt{a_{j1}c_{j1}}}$$

Men kan ook gebruik maken van de associatiematen, zoals voorgesteld in 4.1. t/m 4.6. :

5.7. $$s_{j1} = \frac{a_{j1}}{n}$$ (Russel & Rao)

5.8. $$s_{j1} = \frac{a_{j1} + d_{j1}}{n}$$ (Simple Matching)

5.9. $$s_{j1} = \frac{a_{j1}}{a_{j1} + b_{j1} + c_{j1}}$$ (Jaccard)

5.10. $$s_{j1} = \frac{2(a_{j1}+d_{j1})}{2(a_{j1}+d_{j1}) + b_{j1} + c_{j1}}$$

5.11. $$s_{j1} = \frac{2a_{j1}}{2a_{j1} + b_{j1} + c_{j1}}$$ (Dice)

5.12. $$s_{j1} = \frac{a_{j1}}{a_{j1} + 2(b_{j1}+c_{j1})}$$

Zie voor nog meer maten en nadere toelichting bijvoorbeeld ANDERBERG [1].

B. Nominale variabelen

De volgende drie associatiematen voor nominale variabelen zijn allen gebaseerd op de chi-kwadraat-toetsingsgrootheid zoals die kan worden berekend voor kruistabellen. Stel dat de nominale variabele j de waarden 1 t/m p kan aannemen en de nominale variabele 1 de waarden 1 t/m q.
Vorm de volgende p×q-tabel:

j \ 1	1	2	...	q	
1	n_{11}	n_{12}	...	n_{1q}	n_{1+}
2	n_{21}	n_{22}	...	n_{2q}	n_{2+}
⋮	⋮	⋮		⋮	⋮
p	n_{p1}	n_{p2}	...	n_{pq}	n_{p+}
	n_{+1}	n_{+2}	...	n_{+q}	n

n_{rs} = aantal onderzoekseenheden dat tegelijk r scoort op variabele j en s op variabele 1.

$$n_{r+} = \sum_{s=1}^{q} n_{rs} \quad ; \quad n_{+s} = \sum_{r=1}^{p} n_{rs}$$

De chi-kwadraat grootheid zelf zou een associatiemaat kunnen zijn:

$$\chi^2 = n\left(\sum_{r=1}^{p}\sum_{s=1}^{q}\frac{n_{rs}^2}{n_{r+}n_{+s}} - 1\right) ,$$

ware het niet dat deze maat afhankelijk is van n, en wel toeneemt met groter wordende n. Een correctie hierop levert het voorstel:

$$\phi^2 = \frac{\chi^2}{n} = \sum_{r=1}^{p}\sum_{s=1}^{q}\frac{n_{rs}^2}{n_{r+}n_{+s}} - 1 ,$$

maar deze maat blijkt nog steeds afhankelijk van n.

De volgende drie voorstellen proberen χ^2 te transformeren tot een echte associatiemaat:

5.13. $$s_{j1} = \sqrt{\frac{\phi^2}{\sqrt{(p-1)(q-1)}}}$$ (Tschuprow)

5.14. $$s_{j1} = \sqrt{\frac{\phi^2}{\min\{(p-1),(q-1)\}}}$$ (Cramér,Maung)

5.15. $$s_{j1} = \sqrt{\frac{\phi^2}{1+\phi^2}}$$ (Contingency coëfficiënt van Pierson)

Een moeilijkheid bij deze op χ^2 gebaseerde associatiematen is het ontbreken van een operationele interpretatie: Onderzoekers ondervinden problemen bij het op waarde schatten van verschillen in uitkomsten van deze associatiematen. (GOODMAN & KRUSKAL [3])

C. Ordinale variabelen

Voor ordinale variabelen worden vaak dezelfde maten gebruikt als voor nominale variabelen. Dit levert wel enig informatieverlies op, aangezien geen gebruik gemaakt wordt van de rangordening, maar aan de andere kant is het ook weer zo dat verschil in rang moeilijk is te interpreteren, omdat verschil in rang nog niets hoeft te zeggen over het werkelijke verschil. Voor meer informatie hierover en over maten speciaal voor ordinale variabelen zie bijvoorbeeld KENDALL [10] en KRUSKALL [12].

D. Interval- en ratio-variabelen

We noemen de drie belangrijkste associatiematen:

De cosinus van de hoek tussen de twee vectoren $(x_{1j} \ldots x_{nj})$ en $(x_{11} \ldots x_{n1})$:

$$5.16. \qquad s_{j1} = \frac{\sum_{i=1}^{n} x_{ij}\, x_{i1}}{\sqrt{\sum_{i=1}^{n} x_{ij}^2 \sum_{i=1}^{n} x_{i1}^2}}$$

Deze maat is invariant onder schaaltransformaties. De cosinus is gebaseerd op de originele scores. Het heeft daarom alleen zin deze maat te gebruiken als er een goed gedefinieerde oorsprong (nulpunt) is. Een plaatstransformatie van de oorsprong zou immers tot een andere waarde van de associatiemaat leiden. Daarom is de cosinus alleen te gebruiken, indien men beschikt over ratio-variabelen.

De covariantie is gedefinieerd als:

$$5.17. \qquad s_{j1} = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{i1} - \bar{x}_1),$$

waarin

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

en

$$\bar{x}_1 = \frac{1}{n} \sum_{i=1}^{n} x_{i1}$$

De covariantie is invariant onder plaatstransformaties en kan dus gebruikt worden voor zowel interval-variabelen als ratio-variabelen, mits deze variabelen een unieke schaal hebben. Schaaltransformaties zouden immers tot een andere waarde leiden.

De product-moment correlatiecoëfficiënt is gedefinieerd als:

$$5.18. \qquad s_{jl} = \frac{\sum_{i=1}^{n} (x_{ij}-\bar{x}_j)(x_{il}-\bar{x}_l)}{\sqrt{\sum_{i=1}^{n} (x_{ij}-\bar{x}_j)^2 \sum_{i=1}^{n} (x_{il}-\bar{x}_l)^2}},$$

waarin

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

en

$$\bar{x}_l = \frac{1}{n} \sum_{i=1}^{n} x_{il}$$

Aangezien deze associatiemaat invariant is onder zowel plaats- als schaaltransformaties, kan hij gebruikt worden voor interval-variabelen en ratio-variabelen.

6. CONVERSIE VAN ASSOCIATIE- EN DISSOCIATIEMATEN

Associatiematen en dissociatiematen meten beide de gelijkenis tussen objecten en/of clusters, alleen doen ze het op een tegengestelde manier. Het is goed mogelijk om associatie- en dissociatiematen in elkaar om te zetten. Hieronder volgen een aantal voorbeelden

A. dissociatie → associatie

$$6.1. \qquad s_{ij} = \frac{1}{1 + d_{ij}}$$

$$6.2. \qquad s_{ij} = e^{-d_{ij}}$$

$$6.3. \qquad s_{ij} = 1 - d_{ij}^2 \qquad (\text{als } d_{ij} \leq 1, \forall i,j)$$

$$6.4. \qquad s_{ij} = \frac{m - d_{ij}^2}{m + d_{ij}^2} \qquad (m \text{ is de mediaan der } d_{ij}\text{'s})$$

B. associatie → dissociatie

6.5. $d_{ij} = 1 - |s_{ij}|$

6.6. $d_{ij} = 1 - s_{ij}^2$

Sommige onderzoekers vinden het een prettige eigenschap als hun dissociatiemaat aan de driehoeksongelijkheid voldoet. (zie Def. 2.1.)
Als men daarop prijs stelt dan moet men er rekening mee houden dat bij conversie van een associatiemaat naar een dissociatiemaat de resulterende dissociatiemaat lang niet altijd deze eigenschap heeft. GOWER [4], [5] toont aan dat

6.7. $d_{ij} = \sqrt{2(1-s_{ij})}$

dienst kan doen als afstandsmaat met die eigenschap, indien de matrix der s_{ij}'s positief semidefiniet is. Al eerder was vermeld dat als de s_{ij} de product-moment correlatiecoëfficiënt is, dat dan $d_{ij} = \sqrt{2(1-s_{ij})}$ onder bepaalde voorwaarden wat betreft schaling, dienst kan doen als Euclidische afstandsmaat.

Voor wie meer wil weten over associatie- en dissociatiematen en hun problemen verwijzen we naar SOKAL & SNEATH [18], WILLIAMS & DALE [20] en MORRISON [15].

7. EEN AANTAL HIERARCHISCHE CLUSTERANALYSE METHODEN

In deze paragraaf zullen een aantal veel gebruikte hiërarchische clusteranalyse methoden ter sprake komen. We kunnen ze indelen in drie groepen:

A. "Linkage"-technieken.
Bij deze technieken worden de clusters gekarakteriseerd door hun inwendige structuur, met name hoe de objecten in de cluster ten opzichte van elkaar liggen. Het al of niet samenvoegen van twee clusters wordt bepaald door een criteriumfunctie, waarvan de waarde wordt vastgelegd door de gelijkenis van de objecten uit de ene cluster met de objecten uit de andere

cluster. De gelijkenis van twee clusters is per definitie gelijk aan de waarde van de criteriumfunctie. Vier "linkage"-technieken komen met name aan de orde: de minimum-methode (single linkage), de maximum-methode (complete linkage), de groepsgemiddelden-methode (average linkage) en de methode van ELSHOUT.

B. Zwaartepunt-technieken.
Deze methoden karakteriseren de clusters door hun zwaartepunt. Het al of niet samenvoegen van twee clusters hangt af van een criteriumfunctie, waarvan de waarde wordt bepaald door de gelijkenis van de zwaartepunten van de clusters en eventueel het aantal elementen in de beide clusters. De gelijkenis van twee clusters is per definitie gelijk aan deze criteriumfunctie. Twee zwaartepunt-technieken komen aan de orde: De centroïde-methode en de mediaan-methode.

C. Residuele kwadratensom-technieken.
Hier worden de clusters gekarakteriseerd door de toename in de residuele kwadratensom, die het onderbrengen van een aantal objecten in één groep (cluster), teweeg brengt. Het al of niet samenvoegen van twee clusters hangt af van een criteriumfunctie waarvan de waarde gelijk is aan de toename in de residuele kwadratensom, die het samenvoegen van beide clusters teweeg brengt. De gelijkenis van twee clusters is per definitie gelijk aan deze criteriumfunctie.
Uit deze groep behandelen we één techniek: De methode van WARD.

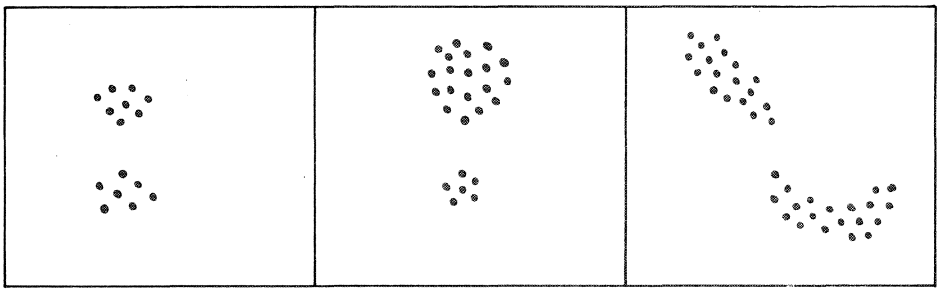
Bij iedere methode zal een recursieformule vermeld worden die het mogelijk maakt, gegeven de actuele associatie/dissociatiematrix van de clusters en de twee samen te voegen clusters, snel de associatie/dissociatiematrix voor de nieuwe clustering uit te rekenen.

7.1. Minimum-Methode (Single Linkage)

De minimum-methode wordt hier beschreven in termen van afstandsmaten, hoewel associatiematen ook zeker toegepast kunnen worden. De methode wordt o.a. vermeld door SNEATH [16] en JOHNSON [9].
De criteriumfunctie, dus de afstand tussen twee clusters C_i en C_j, wordt gedefinieerd als:

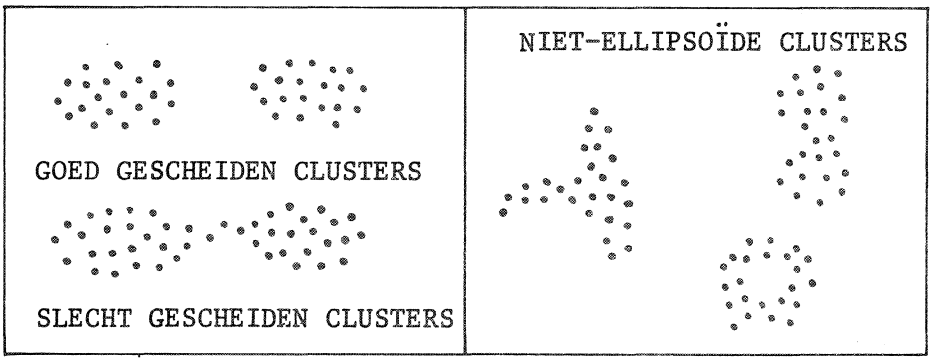
$$D(C_i,C_j) = \min \{d_{ij} : o_i \in C_i \text{ en } o_j \in C_j\}$$

De afstand tussen twee clusters is dus gelijk aan de afstand van de twee dichtstbij gelegen objecten uit de twee clusters. Ongeacht hoe de clusters er verder uitzien, als twee objecten uit de twee clusters maar dicht genoeg bij elkaar liggen, dan worden de clusters samengevoegd. Een gevolg van deze benadering is dat voor de minimum-methode de drie clusterparen uit het onderstaande voorbeeld allen even dicht bij elkaar liggen.



Dit brengt een nadeel met zich mee, want de methode zal niet in staat zijn clusters, die niet duidelijk van elkaar gescheiden zijn, bijvoorbeeld doordat er enkele geïsoleerde punten tussen in liggen, ook werkelijk te onderscheiden. Wanneer dit optreedt, dan heeft de methode de neiging lange slierten van objecten te vormen. Deze tendens wordt wel aangeduid met "ketenvorming". Het is aan de gebruiker zelf om deze eigenschap positief of negatief te beoordelen.

Aan de andere kant is het ook weer zo, dat als de clusters goed onderscheidbaar zijn, de minimum-methode ze zeker zal vinden en bovendien is de methode zodanig dat hij ook goede resultaten levert wanneer de clusters niet een mooie ellipsoïde vorm bezitten:



JOHNSON [9] beschrijft hoe we de waarde waarop twee clusters worden samengevoegd, kunnen interpreteren:
Stel C is de cluster waarvan we een vorm van samenhang willen meten.
Onder een keten van object x naar object y verstaan we iedere rij objecten $o_1, \ldots, o_k$ met $o_1 = x$ en $o_k = y$. Not.: ket(x,y). Voor een keten ket(x,y) wordt de ketenwaarde k(x,y) gedefinieerd als:

$$k(x,y) = \max \{d_{i-1\,i} : i = 2,\ldots,k\}$$

($d_{i-1\,i}$ is de afstand van o_{i-1} en o_i)

k(x,y) is dus de lengte van de grootste stap in de keten.
De ketenafstand dk(x,y) wordt nu gedefinieerd als:

$$dk(x,y) = \min \{k(x,y) : \text{Ket}(x,y) \text{ is keten van x naar y}\}$$

De grootte van de samenhang wordt vastgelegd door:

$$sam(C) = \max \{dk(x,y) : x,y \in C\}$$

sam(C) is dus de grootste stap die ooit gemaakt zou moeten worden om van een willekeurig punt in C naar een willekeurig ander punt in C te gaan.
sam(C) kan aangewend worden om de juiste clustering uit de hiërarchische boom te bepalen, in die zin dat gestopt wordt wanneer verdere uitbreiding van een cluster C tot een onevenredig veel grotere waarde van sam(C) zou leiden.

Recursieformule:
Als C_i en C_j worden samengevoegd tot een nieuwe cluster en C_k is een andere oude cluster, dan geldt

$$D(C_i \cup C_j, C_k) = \min \{D(C_i,C_k), D(C_j,C_k)\}$$

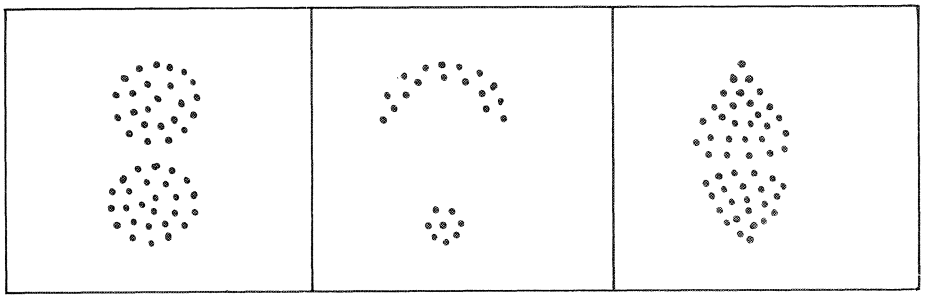
7.2. Maximum-Methode (Complete Linkage)

De maximum-methode zal worden beschreven in termen van afstandsmaten,

maar ook associatiematen kunnen toegepast worden. Deze methode, die o.a. beschreven wordt door JOHNSON [9], is in feite het tegengestelde van de minimum-methode. De criteriumfunctie, dus de afstand tussen twee clusters C_i en C_j, wordt gedefinieerd als:

$$D(C_i,C_j) = \max \{d_{ij} : o_i \in C_i, \quad o_j \in C_j\}$$

De afstand van twee clusters is dus gelijk aan de afstand van de twee verst weg gelegen objecten uit beide clusters. Dit betekent dat gedurende het hiërarchische clusteringsproces alleen dan twee clusters C_i en C_j samengevoegd worden wanneer alle elementen uit beide clusters maar dicht genoeg bij elkaar liggen. Dit in tegenstelling tot de minimum-methode waar twee clusters al worden samengevoegd als er maar twee objecten zijn die dicht genoeg bij elkaar liggen. In de praktijk betekent dit dat de maximum-methode alleen goed in staat is om bolvormige clusters te ontdekken. Ter illustratie staan hieronder drie clusterparen die voor de maximum-methode allen even dicht bij elkaar liggen:



Het is wel zo dat wanneer er alleen maar bolvormige clusters zijn, dat ze dan zeker zullen worden gevonden, ook al zijn ze slecht van elkaar te onderscheiden.

We kunnen in de maximum-methode de clusters karakteriseren door de grootste afstand die tussen objecten van de cluster voorkomt. Deze afstand wordt wel de diameter van de cluster genoemd:

$$\text{diam}(C) = \max \{d_{ij} : o_i, o_j \in C\}$$

Net zoals sam(C) bij de minimum-methode kunnen we diam(C) gebruiken voor het vinden van de meest geschikte clustering uit de hiërarchische boom, door te stoppen wanneer samenvoegen van clusters tot een onevenredig snelle stijging van de diameter zou leiden.

Beide methoden, de minimum- en de maximum-methode, kunnen zowel voor het clusteren van onderzoekseenheden als voor het clusteren van variabelen gebruikt worden. Men moet oppassen bij het gebruik van associatiematen die zowel positieve als negatieve waarden aan kunnen nemen. Als we bij gebruik van correlaties bijvoorbeeld vinden dat $s_{ij} = -0,9$, dan kunnen we dat op twee manieren interpreteren:
In de eerste plaats als "o_i en o_j lijken sterk op elkaar (op het teken na, maar dat is niet belangrijk)", of als "o_i en o_j wijken sterk van elkaar af."

Indien men gesteld is op de eerste interpretatie dan verdient het aanbeveling de associatiemaat zodanig te veranderen dat waarden als -0,9 en +0,9 op dezelfde manier behandeld worden. Dit kan men bijvoorbeeld doen door in plaats van s_{ij} gebruik te maken van $|s_{ij}|$ of s_{ij}^2.

Tot slot dan nog voor de maximum-methode de recursieformule:

Recursieformule:
Als C_i en C_j twee clusters zijn die samengevoegd worden en C_k is een andere cluster, dan geldt:

$$D(C_i \cup C_j, C_k) = \max \{D(C_i, C_k), D(C_j, C_k)\}$$

7.3. Groepsgemiddelden-Methode (Average Linkage)

De groepsgemiddelden-methode is te beschouwen als een tussenvorm van de minimum-methode en de maximum-methode. In termen van afstandsmaten kunnen we de criteriumfunctie, en dus de afstand van twee clusters C_i en C_j, als volgt definiëren:

$$D(C_i, C_j) = \frac{1}{n_i n_j} \sum_{\substack{o_i \in C_i \\ o_j \in C_j}} d_{ij},$$

waarin n_i en n_j het aantal elementen in C_i en C_j.

$D(C_i,C_j)$ is dus de gemiddelde afstand van de objecten van cluster C_i en cluster C_j. Als we associatiematen gebruiken, moeten we wat meer oppassen. Niet alleen moeten we er ons, net als bij gebruik van afstandsmaten, van overtuigen dat het concept van gemiddelde associatie zinvol is, maar bovendien kunnen moeilijkheden ontstaan bij gebruik van associatiematen die negatieve waarden aan kunnen nemen. Als we bijvoorbeeld variabelen willen clusteren met behulp van correlaties en we hebben $s_{12} = 0,8$, $s_{13} = -0,8$, dan is bij deze methode de gelijkenis van cluster $\{o_1\}$ met $\{o_2,o_3\}$ gelijk aan:

$$S(\{o_1\}, \{o_2,o_3\}) = \frac{0,8 - 0,8}{2} = 0$$

Hoewel er dus een zeer sterk verband is tussen de objecten o_1,o_2 en o_3 , zal dit bij deze methode niet blijken. Als men toch een associatiemaat wil gebruiken, dan verdient het de voorkeur alleen associatiematen toe te passen, die positieve waarden aan kunnen nemen, of ze zo te maken door het nemen van de absolutiewaarde of het kwadraat.

In de volgende sectie staat de methode van ELSHOUT, die specifiek gericht is op het clusteren van variabelen met behulp van correlaties en die ook op bovenstaande methode geënt is.

In het geval dat men de cosinus van de hoek tussen twee vectoren gebruikt als associatiemaat voor de objecten en men vindt het concept van gemiddelde cosinus niet aanvaardbaar als maat voor gelijkenis tussen de clusters, dan kan men eventueel iets hebben aan het voorstel van LANCE & WILLIAMS [13]. Zij gebruiken:

$$S(C_i,C_j) = \cos\left(\frac{1}{n_i n_j} \sum_{\substack{o_i \in C_i \\ o_j \in C_j}} \cos^{-1}(s_{ij})\right)$$

Dit wil zoveel zeggen als: "Neem niet de gemiddelde cosinus, maar de cosinus van de gemiddelde hoek."

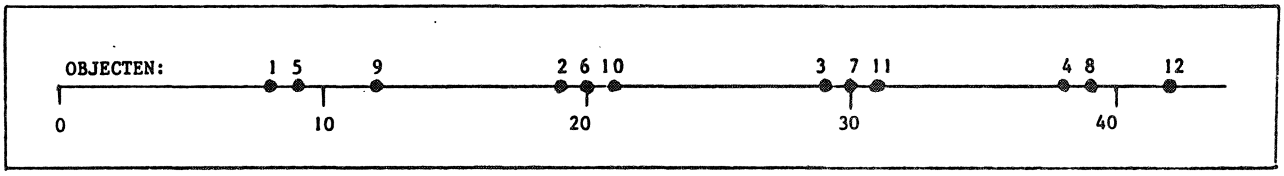
Rekening houdend met de hierboven gemaakte opmerkingen kan de methode zowel voor het clusteren van variabelen als voor het clusteren van onderzoekseenheden gebruikt worden. In termen van afstandsmaten krijgen we de volgende recursieve betrekking:

Recursieformule:
Als C_i en C_j twee clusters zijn die samengevoegd worden en C_k is een andere cluster, dan geldt:

$$D(C_i \cup C_j, C_k) = \frac{n_i D(C_i, C_k) + n_j D(C_j, C_k)}{n_i + n_j} \quad ,$$

waarin n_i en n_j het aantal objecten in C_i en C_j.

Ter illustratie zullen we nu de minimum-, maximum- en groepsgemiddeldenmethode toepassen op dezelfde data, die als volgt gemaakt zijn:
12 punten zijn afgezet op de getallenrechte. Deze punten vormen de te clusteren objecten. Hun afstand wordt gedefinieerd als de gewone Euclidische afstand:



De waarden behorende bij de objecten zijn:

object	waarde	object	waarde	object	waarde
1	8	5	9	9	12
2	19	6	20	10	21
3	29	7	30	11	31
4	38	8	39	12	42

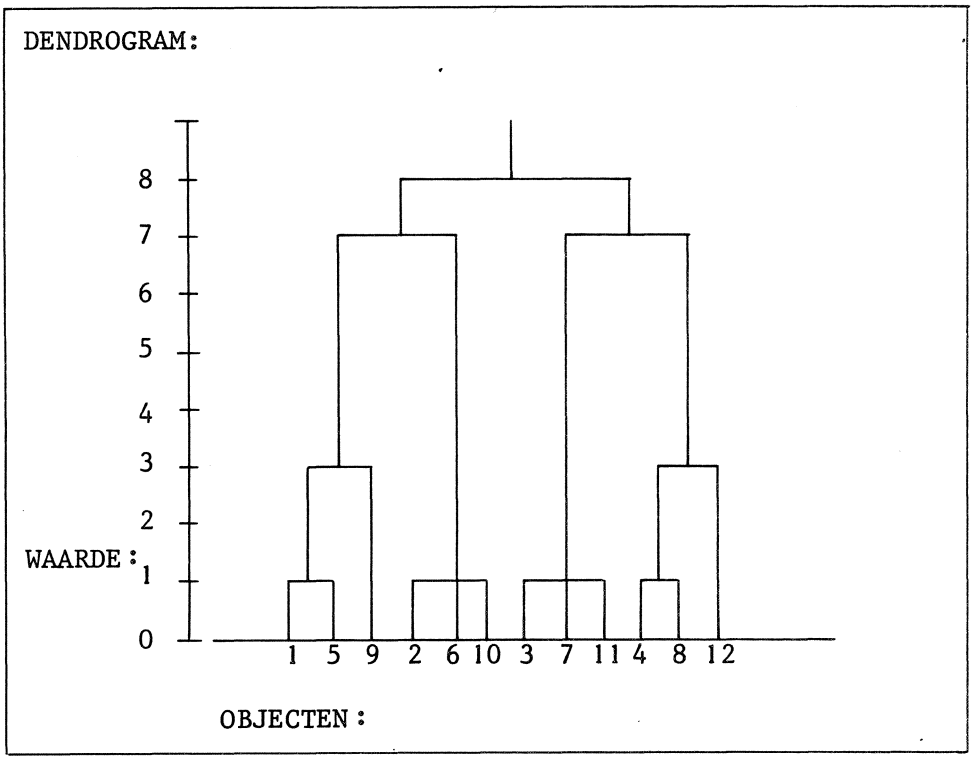
De afstandsmatrix van alle objecten ziet er als volgt uit:

	1	2	3	4	5	6	7	8	9	10	11	12
1	0											
2	11	0										
3	21	10	0									
4	30	19	9	0								
5	1	10	20	29	0							
6	12	1	9	18	11	0						
7	22	11	1	8	21	10	0					
8	31	20	10	1	30	19	9	0				
9	4	7	17	26	3	8	18	27	0			
10	13	2	8	17	12	1	9	18	9	0		
11	23	12	2	7	22	11	1	8	19	10	0	
12	34	23	13	4	33	22	12	3	30	21	11	0

Van alle drie de methoden zal nu in een schema stap voor stap het hiërarchische clusteringsproces beschreven worden. Bovendien zal een dendrogram van het proces vermeld worden.

A. Volgens minimum-methode:

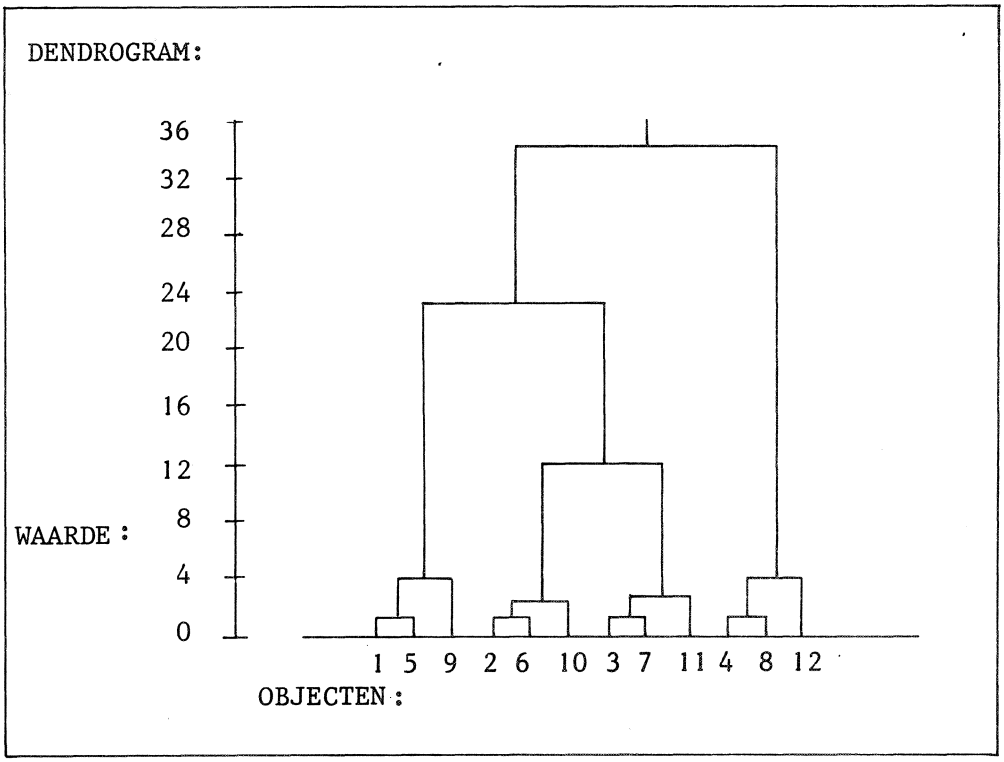
STAP	MEEST GELIJKENDE CLUSTERS	AFSTAND	NIEUWE CLUSTER
1	{1} , {5}	1	{1,5}
2	{2} , {6}	1	{2,6}
3	{3} , {7}	1	{3,7}
4	{4} , {8}	1	{4,8}
5	{2,6} , {10}	1	{2,6,10}
6	{3,7} , {11}	1	{3,7,11}
7	{1,5} , {9}	3	{1,5,9}
8	{4,8} , {12}	3	{4,8,12}
9	{2,6,10} , {1,5,9}	7	{1,2,5,6,9,10}
10	{3,7,11} , {4,8,12}	7	{3,4,7,8,11,12}
11	{1,2,5,6,9,10} , {3,4,7,8,11,12}	8	{1,2,3,4,5,6,7,8,9,10,11,12}



Aangezien na stap 3 de afstanden sterk toenemen zou men kunnen besluiten na stap 3 te stoppen, zodat er dan dus vier clusters zijn: {1,5,9} , {2,6,10}, {3,7,11} en {4,8,12}.

B. Volgens de maximum-methode

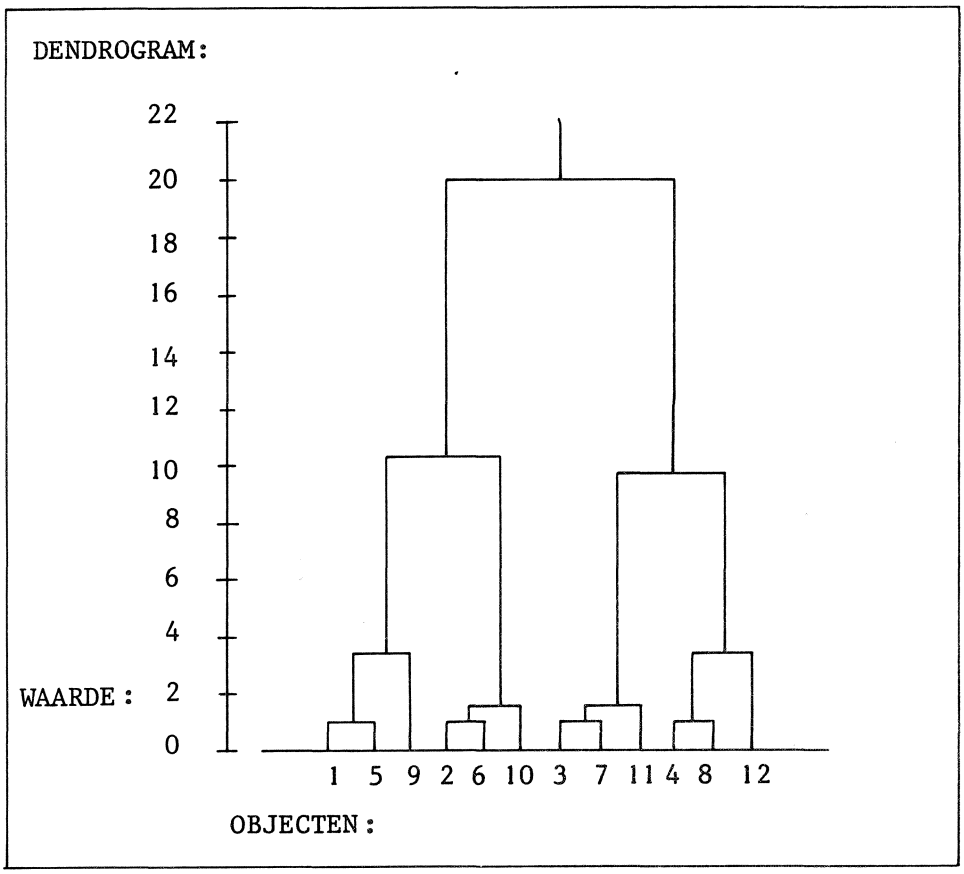
STAP	MEEST GELIJKENDE CLUSTERS	AFSTAND	NIEUWE CLUSTER
1	{1} , {5}	1	{1,5}
2	{2} , {6}	1	{2,6}
3	{3} , {7}	1	{3,7}
4	{4} , {8}	1	{4,8}
5	{2,6} , {10}	2	{2,6,10}
6	{3,7} , {11}	2	{3,7,11}
7	{1,5} , {9}	4	{1,5,9}
8	{4,8} , {12}	4	{4,8,12}
9	{2,6,10} , {3,7,11}	12	{2,3,6,7,10,11}
10	{1,5,9} , {2,3,6,7,10,11}	23	{1,2,3,5,6,7,9,10,11}
11	{1,2,3,5,6,7,9,10,11} , {4,8,12}	34	{1,2,3,4,5,6,7,8,9,10,11,12}



Alhoewel het clusteringsproces hier anders verloopt zou men tot dezelfde conclusie kunnen komen als bij de minimum-methode, namelijk dat er vier clusters zijn.

C. Gemiddelden-methode:

STAP	MEEST GELIJKENDE CLUSTERS	AFSTAND	NIEUWE CLUSTER
1	{1} , {5}	1,000	{1,5}
2	{2} , {6}	1,000	{2,6}
3	{3} , {7}	1,000	{3,7}
4	{4} , {8}	1,000	{4,8}
5	{2,6} , {10}	1,500	{2,6,10}
6	{3,7} , {11}	1,500	{3,7,11}
7	{1,5} , {9}	3,500	{1,5,9}
8	{4,8} , {12}	3,500	{4,8,12}
9	{3,7,11} , {4,8,12}	9,667	{3,4,7,8,11,12}
10	{1,5,9} , {2,6,10}	10,333	{1,2,5,6,9,10}
11	{1,2,5,6,9,10}, {3,4,7,8,11,12}	20,000	{1,2,3,4,5,6,7,8,9,10,11,12}



Ook bij de groepsgemiddelden-methode zouden we weer tot dezelfde conclusie kunnen komen; dat er vier clusters zijn. Het is niet zo verwonderlijk dat we steeds hetzelfde vinden, want dit één-dimensionale voorbeeld werkt goed voor zowel minimum-methode (goed gescheiden clusters) als maximum-methode (bolvormige clusters) en zal het dus ook wel doen voor de er tussen in liggende groepsgemiddelden-methode. In het algemeen zullen echter de resultaten niet zo mooi overeenstemmen.

7.4. De Methode van ELSHOUT

J. ELSHOUT van de Universiteit van Amsterdam heeft een methode geconstrueerd die lijkt op de groepsgemiddelden-methode, maar speciaal geschikt is voor het clusteren van variabelen met behulp van de product-moment correlatiecoëfficiënt. Voor zover bekend, wordt deze methode niet in publicaties genoemd.

Definieer de volgende grootheden:

a. De *tekenfunctie* is een functie $t_C : O \rightarrow \{-1, 0, +1\}$, met:

$$t_C(o_i) = \begin{cases} -1 \text{ of } +1 & \text{als } o_i \in C \\ 0 & \text{als } o_i \in C \end{cases}$$

Bij initialisatie van het hiërarchische clusteringsproces geldt:

$$t_{\{o_i\}}(o_j) = \begin{cases} 1 & \text{als } i = j \\ 0 & \text{als } i \neq j \end{cases}$$

b. De *cluster-correlatiecoëfficiënt* r voor twee clusters C_i en C_j, niet noodzakelijk disjunct, is gelijk aan:

$$r(C_i,C_j) = \frac{1}{|C_i \oplus C_j|} \sum_{(o_i,o_j) \in C_i \oplus C_j} t_{C_i}(o_i) \cdot t_{C_j}(o_j) \cdot s_{ij},$$

waarin

$$C_i \oplus C_j = \{(o_i,o_j) : o_i \in C_i,\ o_j \in C_j,\ o_i \neq o_j\}$$

en $|C_i \oplus C_j|$ het aantal elementen in $C_i \oplus C_j$.

c. Voor een cluster $C \subset O$ is de *introcluster-correlatiecoëfficiënt* gelijk aan $r(C,C)$.
$r(C,C)$ is een maat voor de gelijkenis van de objecten in C. Voor een cluster die een enkel object o_i bevat geldt per definitie $r(\{o_i\}, \{o_i\}) = 1$. Naarmate er meer objecten komen in de cluster C neemt $r(C,C)$ toe. $r(C,C)$ kan gebruikt worden voor het bepalen van een zinvolle clustering door het hiërarchische proces te stoppen wanneer de afname van $r(C,C)$ plotseling onevenredig snel zou plaats vinden.

d. De *intercluster-correlatiecoëfficiënt* van twee disjuncte clusters C_i en C_j is gelijk aan $r(C_i,C_j)$. De intercluster-correlatie wordt gebruikt als associatiemaat voor de clusters, dus:

$$S(C_i,C_j) = r(C_i,C_j)$$

Bij deze methode worden in elke stap steeds die twee clusters C_i en C_j samengevoegd die de in absolute waarde grootste intercluster-correlatie bezitten. Als echter de intercluster-correlatie in werkelijkheid negatief is, dan draaien in de tweede cluster C_j alle variabelen van teken om. Dit gebeurt door middel van de tekenfunctie:

$$t_{C_i \cup C_j}(o) = \begin{cases} t_{C_j}(o) & \text{als } o \in C_j \quad \text{en } S(C_i,C_j) \geq 0, \\ -t_{C_j}(o) & \text{als } o \in C_j \quad \text{en } S(C_i,C_j) < 0, \\ t_{C_i}(o) & \text{als } o \in C_i, \\ 0 & \text{als } o \notin C_i \cup C_j. \end{cases}$$

Door dit tekenmechanisme wordt bereikt dat alle variabelen binnen een cluster zoveel mogelijk gelijk gericht zijn, d.w.z. een grote waarde heeft voor alle variabelen binnen de cluster dezelfde onderliggende betekenis, waardoor de intracluster-correlatie inderdaad een (positieve) maat voor de samenhang van een cluster wordt en problemen als bij de groepsgemiddelden-methode vermeden worden.

Wanneer de methode van ELSHOUT toegepast wordt met dissociatiematen of met associatiematen die alleen positieve waarden aan kunnen nemen, dan reduceert de methode tot de groepsmiddelen-methode.

Recursieformule:
Wanneer clusters C_i en C_j worden samengevoegd tot een nieuwe cluster en C_k is een andere cluster, dan geldt:

$$S(C_i \cup C_j, C_k) = \frac{n_i \cdot S(C_i,C_k) + \text{sgn}(S(C_i,C_j)) \cdot n_j \cdot S(C_j,C_k)}{(n_i + n_j)}$$

waarin

$$\text{sgn}(x) = \begin{cases} +1 & \text{als } x > 0 \\ 0 & \text{als } x = 0 \\ -1 & \text{als } x < 0 \end{cases}$$

Aangezien de methode van ELSHOUT nergens in de literatuur voorkomt, volgt hieronder een eenvoudig voorbeeld waarin de werking uitvoerig wordt geïllustreerd.

Gegeven is onderstaande correlatiematrix van variabelen:

	1	2	3	4	5
1	1,000				
2	0,130	1,000			
3	0,830	-0,220	1,000		
4	-0,940	0,720	-0,720	1,000	
5	-0,350	-0,440	0,120	-0,750	1,000

Bij iedere stap zullen worden vermeld: meest gelijkende clusters, hun intercorrelatie, de nieuw gevormde cluster, zijn intracluster-correlatie en de nieuwe associatiematrix van de cluster.

stap 1: Grootste absolute associatie tussen {1} en {4} : S_{14} = -0,940
Nieuwe cluster: {1,-4}
Intracluster-correlatie: 0,940
Nieuwe associatiematrix:

	{1,-4}	{2}	{3}	{5}
{1,-4}	1,000			
{2}	-0,295	1,000		
{3}	0,775	-0,220	1,000	
{5}	0,200	-0,440	0,120	1,000

stap 2: Grootste absolute associatie tussen {1,-4} en {3} : S({1,-4},{3}) = = 0,775
Nieuwe cluster: {1,3,-4}
Intracluster-correlatie: 0,830
Nieuwe associatiematrix:

	{1,3,-4}	{2}	{5}
{1,3,-4}	1,000		
{2}	-0,270	1,000	
{5}	0,173	-0,440	1,000

stap 3: Grootste absolute associatie tussen {2} en {5} : S({2},{5}) = -0,440

Nieuwe cluster: {2, -5}

Intracluster-correlatie: 0,440

Nieuwe associatiematrix:

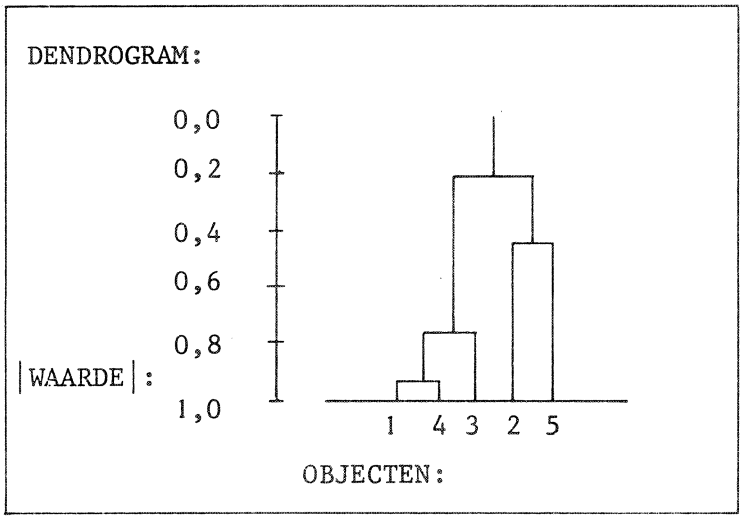
	{1,3,-4}	{2,-5}
{1,3,-4}	1,000	
{2,-5}	-0,222	1,000

stap 4: Grootste absolute associatie tussen {1,3,-4} en {2,-5} :

S({1,3,-4},{2,-5}) = -0,222

Nieuwe cluster: {1,-2,3,-4,5}

Intracluster-correlatie: 0,426



Concluderend kunnen we stellen dat het alleen zinvol is om de methode van ELSHOUT te gebruiken indien men variabelen wil clusteren met behulp van de product-moment correlatiecoëfficiënt of eventueel de cosinus van de hoek tussen de vectoren. Voor het clusteren van onderzoekseenheden en het gebruik van andere associatie- of dissociatiematen kan verwezen worden naar de groeps-gemiddelden-methode.

7.5. Centroïde-Methode

De centroïde-methode zal eerst worden beschreven in termen van afstandsmaten voor het clusteren van onderzoekseenheden en daarna in termen van correlaties voor het clusteren van variabelen.

a. Clusteren van onderzoekseenheden:

Het zwaartepunt van een cluster C_i is per definitie gelijk aan

$$z_i = (z_{i1},\ldots,z_{ip}),$$

met

$$z_{ij} = \frac{1}{n_i} \sum_{o_k \in C_i} x_{kj},$$

waarin n_i het aantal objecten in C_i en x_{kj} de score van object o_k op variabele j.

De afstand tussen twee clusters C_i en C_j wordt nu gedefinieerd als de Euclidische afstand tussen de zwaartepunten van C_i en C_j:

$$D(C_i,C_j) = \|z_i - z_j\| = \sqrt{\sum_{k=1}^{p} (z_{ik}-z_{jk})^2}$$

Als twee clusters C_i en C_j worden samengevoegd, dan is het zwaartepunt van de nieuwe cluster $C_i \cup C_j$ gelijk aan:

$$\frac{n_i z_i + n_j z_j}{n_i + n_j},$$

Immers:

Het zwaartepunt is z_l met

$$z_{il} = \frac{1}{|C_i \cup C_j|} \sum_{o_k \in C_i \cup C_j} x_{kl} = \frac{1}{n_i+n_j} \left(\sum_{o_k \in C_i} x_{kl} + \sum_{o_k \in C_j} x_{kl} \right) =$$

$$= \frac{1}{n_i+n_j} (n_i z_{il} + n_j z_{jl}).$$

Voor de recursieformule geldt:

Als C_i en C_j worden samengevoegd en C_k is een andere cluster, dan is

$$D^2(C_i \cup C_j, C_k) = \frac{n_i}{n_i+n_j} D^2(C_i,C_k) + \frac{n_j}{n_i+n_j} D^2(C_j,C_k) - \frac{n_i n_j}{(n_i+n_j)^2} D^2(C_i,C_j)$$

Immers:

$$D^2(C_i \cup C_j, C_k) = \left\| \frac{n_i z_i + n_j z_j}{n_i + n_j} - z_k \right\|^2 =$$

$$= \sum_{l=1}^{p} \frac{1}{(n_i+n_j)^2} (n_i z_{il} + n_j z_{jl} - (n_i+n_j) z_{kl})^2 =$$

$$= \frac{1}{(n_i+n_j)^2} \sum_{l=1}^{p} (n_i^2 z_{il}^2 + n_j^2 z_{jl}^2 + (n_i+n_j)^2 z_{kl}^2 + 2n_i n_j z_{il} z_{jl} +$$

$$- 2n_i(n_i+n_j) z_{il} z_{kl} - 2n_j(n_i+n_j) z_{jl} z_{kl}) =$$

$$= \frac{1}{(n_i+n_j)^2} \sum_{l=1}^{p} ((n_i^2+n_i n_j) z_{il}^2 - 2(n_i^2+n_i n_j) z_{il} z_{kl} + (n_i^2+n_i n_j) z_{kl}^2 +$$

$$+ (n_j^2+n_i n_j) z_{jl}^2 - 2(n_j^2+n_i n_j) z_{jl} z_{kl} + (n_j^2+n_i n_j) z_{kl}^2 +$$

$$- n_i n_j z_{il}^2 - n_i n_j z_{jl}^2 + 2n_i n_j z_{il} z_{jl}) =$$

$$= \frac{n_i}{n_i+n_j} \sum_{l=1}^{p} (z_{il}-z_{kl})^2 + \frac{n_j}{n_i+n_j} \sum_{l=1}^{p} (z_{jl}-z_{kl})^2 - \frac{n_i n_j}{(n_i+n_j)^2} \sum_{l=1}^{p} (z_{il}-z_{jl})^2 =$$

$$= \frac{n_i}{n_i+n_j} D^2(C_i,C_k) + \frac{n_j}{n_i+n_j} D^2(C_j,C_k) - \frac{n_i n_j}{(n_i+n_j)^2} D^2(C_i,C_k).$$

b. Clusteren van variabelen:

Analoog aan het zwaartepunt bij het clusteren van onderzoekseenheden wordt bij het clusteren van variabelen de cluster gekarakteriseerd door het gemiddelde van de variabelen in de cluster. Als we de product-moment correlatiecoëfficiënt aangeven met r dan wordt de associatie tussen twee clusters C_i en C_j gedefinieerd door:

$$S(C_i,C_j) = r\left(\frac{1}{n_i}\sum_{o_i \in C_i} o_i \;,\; \frac{1}{n_j}\sum_{o_j \in C_j} o_j\right)$$

waarin n_i en n_j het aantal variabelen in C_i en C_j zijn.

KING [11], aan wie deze methode wordt toegeschreven, gebruikte in plaats van het gemiddelde van de variabelen de som van de variabelen. Dit maakt in de praktijk geen enkel verschil, aangezien de correlatie-coëfficiënt invariant is onder schaaltransformaties.
Ook voor dit geval is een recursieformule te bepalen, die echter ingewikkelder is. In feite moeten we steeds weten wat de standaardafwijking (of variantie) is van de gemiddelde variabele die de cluster representeert.

Recursieformule:
Stel C_i en C_j zijn twee clusters die samengevoegd worden en C_k is een andere cluster. Laten de representerende variabelen van C_i,C_j en C_k aangeduid worden met y_i,y_j en y_k en hun standaardafwijkingen met σ_i, σ_j en σ_k. Als de aantallen variabelen van C_i,C_j en C_k gelijk zijn aan n_i,n_j en n_k, dan geldt:

$$S(C_i \cup C_j, C_k) = \frac{n_i\sigma_i S(C_i,C_k) + n_j\sigma_j S(C_j,C_k)}{\sqrt{n_i^2\sigma_i^2 + n_j^2\sigma_j^2 + 2n_in_j\sigma_i\sigma_j S(C_i,C_j)}}$$

Immers

$$S(C_i \cup C_j, C_k) = r\left(\frac{n_iy_i+n_jy_j}{n_i+n_j}\;,\; y_k\right) = r(n_iy_i+n_jy_j, y_k) =$$

$$= \frac{n_i\,\mathrm{cov}(y_i,y_k) + n_j\,\mathrm{cov}(y_j,y_k)}{\sqrt{(\mathrm{var}(n_ix_i) + \mathrm{var}(n_jx_j) + 2\mathrm{cov}(n_ix_i,n_jx_j))\,\mathrm{var}(y_k)}} =$$

$$= \frac{n_i\sigma_i\sigma_k\,S(C_iC_k) + n_j\sigma_j\sigma_k\,S(C_j,C_k)}{\sqrt{(n_i^2\sigma_i^2 + n_j^2\sigma_j^2 + 2n_in_j\sigma_i\sigma_j\,S(C_i,C_j))\sigma_k^2}} =$$

Voor de recursieformule geldt:

Als C_i en C_j worden samengevoegd en C_k is een andere cluster, dan is

$$D^2(C_i \cup C_j, C_k) = \frac{n_i}{n_i+n_j} D^2(C_i,C_k) + \frac{n_j}{n_i+n_j} D^2(C_j,C_k) - \frac{n_i n_j}{(n_i+n_j)^2} D^2(C_i,C_j)$$

Immers:

$$D^2(C_i \cup C_j, C_k) = \left\| \frac{n_i z_i + n_j z_j}{n_i + n_j} - z_k \right\|^2 =$$

$$= \sum_{l=1}^{p} \frac{1}{(n_i+n_j)^2} (n_i z_{il} + n_j z_{jl} - (n_i+n_j) z_{kl})^2 =$$

$$= \frac{1}{(n_i+n_j)^2} \sum_{l=1}^{p} (n_i^2 z_{il}^2 + n_j^2 z_{jl}^2 + (n_i+n_j)^2 z_{kl}^2 + 2n_i n_j z_{il} z_{jl} +$$

$$- 2n_i(n_i+n_j) z_{il} z_{kl} - 2n_j(n_i+n_j) z_{jl} z_{kl}) =$$

$$= \frac{1}{(n_i+n_j)^2} \sum_{l=1}^{p} ((n_i^2+n_i n_j) z_{il}^2 - 2(n_i^2+n_i n_j) z_{il} z_{kl} + (n_i^2+n_i n_j) z_{kl}^2 +$$

$$+ (n_j^2+n_i n_j) z_{jl}^2 - 2(n_j^2+n_i n_j) z_{jl} z_{kl} + (n_j^2+n_i n_j) z_{kl}^2 +$$

$$- n_i n_j z_{il}^2 - n_i n_j z_{jl}^2 + 2n_i n_j z_{il} z_{jl}) =$$

$$= \frac{n_i}{n_i+n_j} \sum_{l=1}^{p} (z_{il}-z_{kl})^2 + \frac{n_j}{n_i+n_j} \sum_{l=1}^{p} (z_{jl}-z_{kl})^2 - \frac{n_i n_j}{(n_i+n_j)^2} \sum_{l=1}^{p} (z_{il}-z_{jl})^2 =$$

$$= \frac{n_i}{n_i+n_j} D^2(C_i,C_k) + \frac{n_j}{n_i+n_j} D^2(C_j,C_k) - \frac{n_i n_j}{(n_i+n_j)^2} D^2(C_i,C_k).$$

b. Clusteren van variabelen:

Analoog aan het zwaartepunt bij het clusteren van onderzoekseenheden wordt bij het clusteren van variabelen de cluster gekarakteriseerd door het gemiddelde van de variabelen in de cluster. Als we de product-moment correlatiecoëfficiënt aangeven met r dan wordt de associatie tussen twee clusters C_i en C_j gedefinieerd door:

$$S(C_i,C_j) = r\left(\frac{1}{n_i}\sum_{o_i \in C_i} o_i \;,\; \frac{1}{n_j}\sum_{o_j \in C_j} o_j\right)$$

waarin n_i en n_j het aantal variabelen in C_i en C_j zijn.

KING [11], aan wie deze methode wordt toegeschreven, gebruikte in plaats van het gemiddelde van de variabelen de som van de variabelen. Dit maakt in de praktijk geen enkel verschil, aangezien de correlatiecoëfficiënt invariant is onder schaaltransformaties.
Ook voor dit geval is een recursieformule te bepalen, die echter ingewikkelder is. In feite moeten we steeds weten wat de standaardafwijking (of variantie) is van de gemiddelde variabele die de cluster representeert.

Recursieformule:
Stel C_i en C_j zijn twee clusters die samengevoegd worden en C_k is een andere cluster. Laten de representerende variabelen van C_i,C_j en C_k aangeduid worden met y_i,y_j en y_k en hun standaardafwijkingen met σ_i, σ_j en σ_k. Als de aantallen variabelen van C_i,C_j en C_k gelijk zijn aan n_i,n_j en n_k, dan geldt:

$$S(C_i \cup C_j, C_k) = \frac{n_i\sigma_i S(C_i,C_k) + n_j\sigma_j S(C_j,C_k)}{\sqrt{n_i^2\sigma_i^2 + n_j^2\sigma_j^2 + 2n_in_j\sigma_i\sigma_j S(C_i,C_j)}}$$

Immers

$$S(C_i \cup C_j, C_k) = r\left(\frac{n_iy_i+n_jy_j}{n_i+n_j}\;,\; y_k\right) = r(n_iy_i+n_jy_j, y_k) =$$

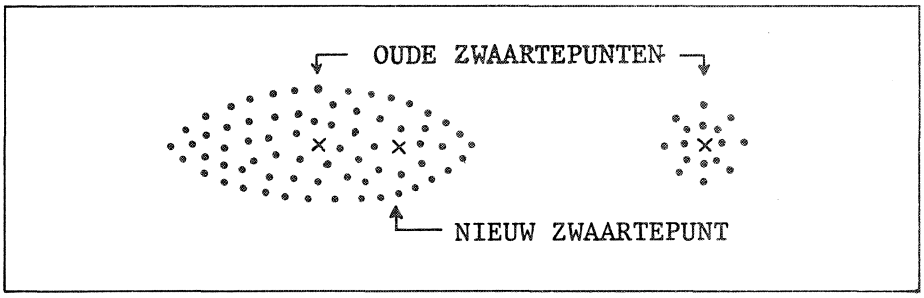
$$= \frac{n_i\,\mathrm{cov}(y_i,y_k) + n_j\,\mathrm{cov}(y_j,y_k)}{\sqrt{(\mathrm{var}(n_ix_i) + \mathrm{var}(n_jx_j) + 2\mathrm{cov}(n_ix_i,n_jx_j))\,\mathrm{var}(y_k)}} =$$

$$= \frac{n_i\sigma_i\sigma_k\,S(C_iC_k) + n_j\sigma_j\sigma_k\,S(C_j,C_k)}{\sqrt{(n_i^2\sigma_i^2 + n_j^2\sigma_j^2 + 2n_in_j\sigma_i\sigma_j\,S(C_i,C_j))\sigma_k^2}} =$$

$$= \frac{n_i\sigma_i\, S(C_i,C_k) + n_j\sigma_j\, S(C_j,C_k)}{\sqrt{n_i^2\sigma_i^2 + n_j^2\sigma_j^2 + 2n_in_j\sigma_i\sigma_j\, S(C_i,C_k)}}.$$

Dat de beide recursieformules zo weinig op elkaar lijken ligt daaraan dat het niet zonder meer mogelijk is afstanden en correlaties in elkaar om te zetten. Zoals al beschreven is in paragraaf 4 (na 4.17) spelen hierbij transformatieproblemen een rol.

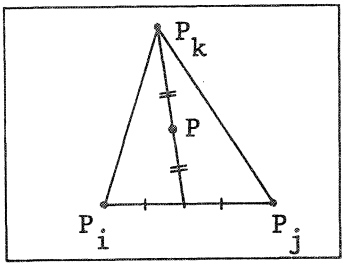
De centroïde-methode heeft het nadeel dat grote clusters kleine clusters kunnen opslokken op een zodanige wijze dat maar weinig van de karakteristieke eigenschappen van de kleine cluster bewaard blijven.
Het grote aantal objecten in de grote cluster zorgt er namelijk voor dat bij eventuele samensmelting van de grote en kleine cluster het zwaartepunt van de nieuwe cluster vlak in de buurt komt te liggen van het zwaartepunt van de oude grote cluster:



7.6. Mediaan-Methode

De mediaan-methode is ontwikkeld door GOWER [6] in een poging het nadeel van het opslokken van kleine door grote clusters bij de centroïde-methode op te heffen. Een cluster C wordt nu niet meer gekarakteriseerd door zijn zwaartepunt, maar door een ander punt, dat ontstaat door het midden te nemen van de twee kenmerkende punten van de clusters C' en C'' waar de cluster C uit is ontstaan. ($C=C'\cup C''$). In de praktijk betekent dit dat we de centroïde-methode toepassen, maar bij het samenvoegen net doen als of de samen te voegen clusters evenveel objecten bevatten.

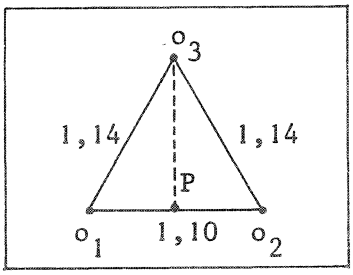
De methode ontleent zijn naam aan het volgende: Als clusters C_i,C_j en C_k gekarakteriseerd worden door respectievelijk de punten P_i,P_j en P_k en eerst worden C_i en C_j samengevoegd en daarna $C_i \cup C_j$ en C_k, dan ligt het karakteristieke punt P van $C_i \cup C_j \cup C_k$ op de mediaan (zwaartelijn) uit P_k op $P_i P_j$ van de driehoek $P_i P_j P_k$:



Recursieformule:
Als clusters C_i en C_j worden samengevoegd en C_k is een andere cluster, dan geldt in termen van afstandsmaten:

$$D^2(C_i \cup C_j, C_k) = \tfrac{1}{2}D^2(C_i, C_k) + \tfrac{1}{2}D^2(C_j, C_k) - \tfrac{1}{4}D^2(C_i, C_j)$$

De centroïde-methode en de mediaan-methode hebben een kenmerkende eigenschap, die de overige hier beschreven methoden niet bezitten. Het kan namelijk voorkomen dat tijdens het clusteringsproces de afstanden tussen de clusters plotseling kleiner worden dan ze ooit tevoren zijn geweest, doordat de zwaartepunten zich verplaatsen en daardoor dicht in elkaars buurt kunnen komen te liggen. Neem bijvoorbeeld eens drie objecten o_1,o_2 en o_3, die op de hoekpunten van een driehoek liggen zodanig, dat:
$d_{12} = 1{,}10$, $d_{13} = 1{,}14$ en $d_{23} = 1{,}14$



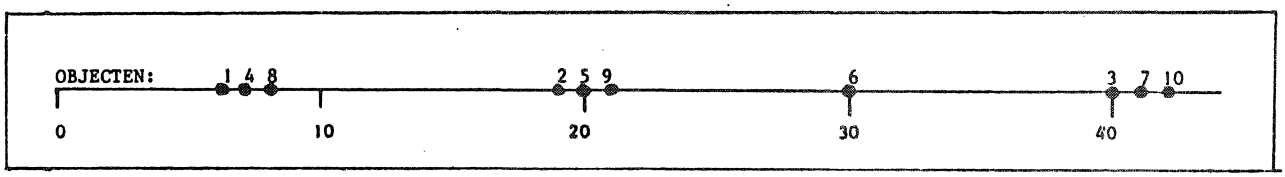
Als eersten worden o_1 en o_2 samengevoegd aangezien hun afstand de kleinste is (d_{12}=1,10). $\{o_1,o_2\}$ wordt nu gekarakteriseerd door het punt op het midden op de verbindingslijn. Dit is voor zowel de centroïde-methode als de mediaan-methode hetzelfde punt P. Voor de afstand $\{o_1,o_2\}$ en $\{o_3\}$ geldt nu:

$$D(\{o_1,o_2\}, \{o_3\}) = \sqrt{1,14^2 - 0,55^2} = 1,00$$

Deze afstand is kleiner dan alle oorspronkelijke afstanden. De hierboven beschreven eigenschap heeft tot gevolg, dat het soms voor beide methoden niet meer mogelijk is het hiërarchische clusteringsproces op een nette wijze weer te geven in een dendrogram. Bij een dendrogram wordt immers aangenomen dat de afstanden tussen de clusters niet kleiner worden naarmate het proces vordert.

Tot slot van de zwaartepunt-methoden volgt hieronder een voorbeeld van beide methoden, werkend op dezelfde getallen:

Voorbeeld: Op de getallenrechte zijn de volgende 10 punten afgezet die de te clusteren objecten vormen:



De waarden, behorende bij de objecten zijn:

object	waarde	object	waarde	object	waarde
1	6,1	5	20,0	8	7,9
2	19,0	6	30,0	9	21,0
3	39,9	7	41,0	10	42,1
4	7,0				

Als maat voor de gelijkenis nemen we de Euclidische afstand.
Dit leidt tot de volgende afstandsmatrix van de objecten:

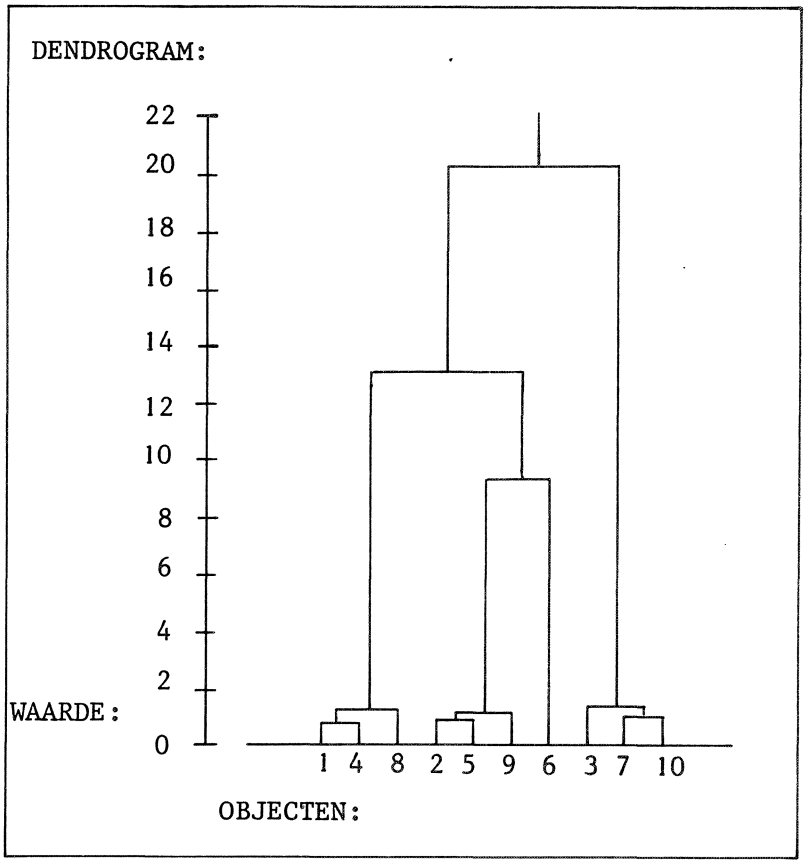
	1	2	3	4	5	6	7	8	9	10
1	0,0									
2	12,9	0,0								
3	33,8	20,9	0,0							
4	0,9	12,0	32,9	0,0						
5	13,9	1,0	19,9	13,0	0,0					
6	23,9	11,0	9,9	23,0	10,0	0,0				
7	34,9	22,0	1,1	34,0	21,0	11,0	0,0			
8	1,8	11,1	32,0	0,9	12,1	22,1	33,1	0,0		
9	14,9	2,0	18,9	14,0	1,0	9,0	20,0	13,1	0,0	
10	36,0	23,1	2,2	35,1	22,1	12,1	1,1	34,2	21,1	0,0

A. Volgens centroïde-methode:

STAP	MEEST GELIJKENDE CLUSTERS	AFSTAND	NIEUWE CLUSTER
1	{1} , {4}	0,900	{1,4}
2	{2} , {5}	1,000	{2,5}
3	{7} , {10}	1,100	{7,10}
4	{1,4} , {8}	1,125	{1,4,8}
5	{2,5} , {9}	1,250	{2,5,9}
6	{3} , {7,10}	1,375	{3,7,10}
7	{2,5,9} , {6}	9,556	{2,5,6,9}
8	{1,4,8} , {2,5,6,9}	12,975	{1,2,4,5,6,8,9}
9	{1,2,4,5,6,8,9}, {3,7,10}	20,091	{1,2,3,4,5,6,7,8,9,10}

Op grond van de informatie in de tabel ligt het voor de hand te stoppen na stap 6, zodat we tot de conclusie komen dat er vier clusters zijn: {1,4,8}, {2,5,9}, {3,7,10} en {6}.

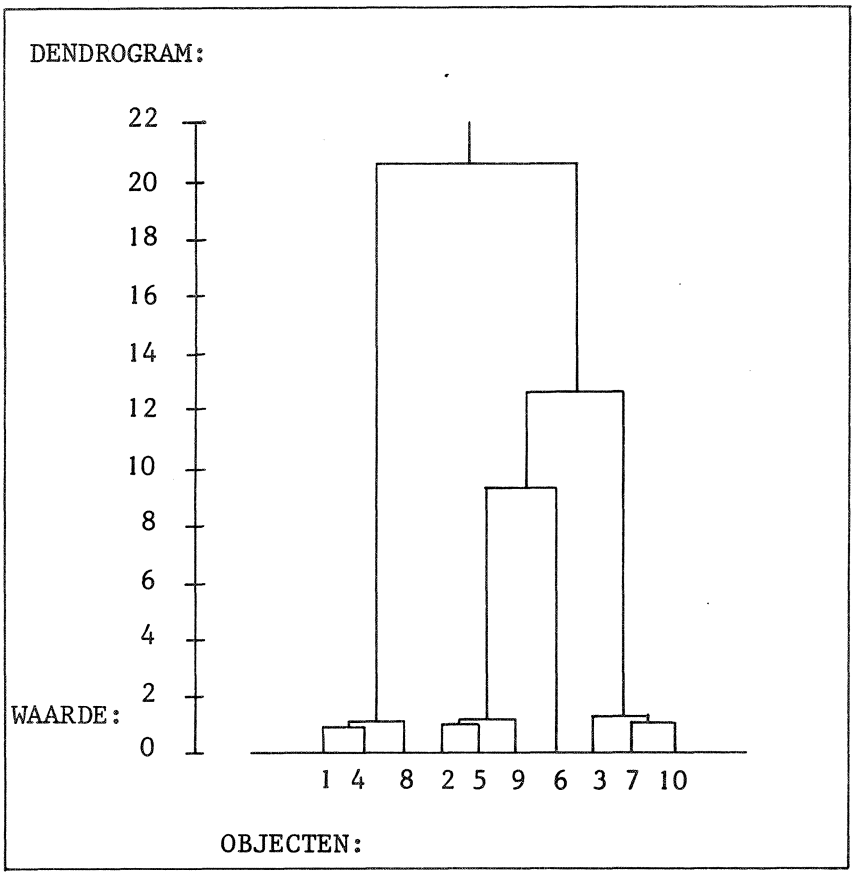
Het dendrogram illustreert dit nog eens duidelijk:



B. Voor de mediaan-methode:

STAP	MEEST GELIJKENDE CLUSTERS	AFSTAND	NIEUWE CLUSTER
1	{1}, {4}	0,900	{1,4}
2	{2}, {5}	1,000	{2,5}
3	{7}, {10}	1,100	{7,10}
4	{1,4}, {8}	1,125	{1,4,8}
5	{2,5}, {9}	1,250	{2,5,9}
6	{3}, {7,10}	1,375	{3,7,10}
7	{2,5,9}, {6}	9,313	{2,5,6,9}
8	{2,5,6,9}, {3,7,10}	12,572	{2,3,5,6,7,9,10}
9	{1,4,8}, {2,3,5,6,7,9,10}	20,649	{1,2,3,4,5,6,7,8,9,10}

Uiteraard lijken aanvankelijk beide methoden in dit voorbeeld op elkaar. De verschillen treden pas op wanneer de omvang van de clusters die samengevoegd worden sterk gaat verschillen.



De keuze van het al of niet wegen van de clusters met het aantal objecten blijft aan de gebruiker. Het voorbeeld laat in ieder geval zien dat beide methoden tot verschillende resultaten kunnen leiden.

7.7. Methode van WARD

De laatste methode die besproken zal worden is de methode van WARD. Het intuïtieve eventueel te visualiseren, idee van de afstand tussen twee objecten is hier los gelaten.
De methode is ontworpen voor het clusteren van onderzoekseenheden, maar zou toegepast kunnen worden voor het clusteren van variabelen, mits het gebruik van de Euclidische afstand voor het meten van de gelijkenis tussen variabelen aanvaardbaar gevonden wordt. Dit zal echter weinig gebeuren aangezien er geen duidelijk eenvoudig verband is tussen de product-moment correlatiecoëfficiënt en de Euclidische afstand. Dezelfde schalingsproblemen, beschreven in Sectie 7.5., spelen een rol. De methode van WARD zal daarom alleen behandeld worden voor het clusteren van onderzoekseenheden.

WARD [19] gaat er van uit dat, als een groep objecten een cluster vormt, de variantie van de objecten binnen de groep zo klein mogelijk moet zijn. In plaats van de variantie gebruikt hij echter de residuele kwadratensom: De residuele kwadratensom van een cluster C_i is gelijk aan

$$SSE(C_i) = \sum_{j=1}^{p} \sum_{o_k \in C_i} (x_{kj} - \bar{x}_j^{(i)})^2,$$

waarin

$$\bar{x}_j^{(i)} = \frac{1}{n_i} \sum_{o_k \in C_i} x_{kj},$$

en n_i het aantal objecten in C_i.

De afstand tussen twee clusters C_i en C_j wordt nu gedefinieerd als de toename in de residuele kwadratensom, wanneer C_i en C_j samengevoegd zouden worden:

$$D(C_i,C_j) = SSE(C_i \cup C_j) - SSE(C_i) - SSE(C_j)$$

Als C_i en C_j worden samengevoegd, dan is de toename in de residuele kwadratensom gelijk aan:

$$SSE(C_i \cup C_j) - SSE(C_i) - SSE(C_j) =$$

$$= \sum_{l=1}^{p} \sum_{o_k \in C_i \cup C_j} (x_{kl} - \frac{1}{n_i+n_j} \sum_{o_k \in C_i \cup C_j} x_{kl})^2 -$$

$$- \sum_{l=1}^{p} \sum_{o_k \in C_i} (x_{kl} - \frac{1}{n_i} \sum_{o_k \in C_i} x_{kl})^2 - \sum_{l=1}^{p} \sum_{o_k \in C_j} (x_{kl} - \frac{1}{n_j} \sum_{o_k \in C_j} x_{kl})^2 =$$

$$= \sum_{l=1}^{p} \left(\sum_{o_k \in C_i \cup C_j} (x_{kl} - \frac{n_i \bar{x}_l^{(i)} + n_j \bar{x}_l^{(j)}}{n_i + n_j} \right)^2 -$$

$$\left(\sum_{o_k \in C_i} (x_{kl} - \bar{x}_l^{(i)})^2 - \sum_{o_k \in C_j} (x_{kl} - \bar{x}_l^{(j)})^2 \right) =$$

$$= \frac{1}{n_i+n_j} \sum_{l=1}^{p} (n_i(n_i+n_j)(\bar{x}_l^{(i)})^2 + n_j(n_i+n_j)(\bar{x}_l^{(j)})^2 - (n_i\bar{x}_l^{(i)}+n_j\bar{x}_l^{(j)})^2) =$$

$$= \frac{1}{n_i+n_j} \sum_{l=1}^{p} (n_in_j(\bar{x}_l^{(i)})^2 + n_in_j(\bar{x}_l^{(j)})^2 - 2n_in_j\bar{x}_l^{(i)}\bar{x}_l^{(j)}) =$$

$$= \frac{n_in_j}{n_i+n_j} \sum_{l=1}^{p} (\bar{x}_l^{(i)} - \bar{x}_l^{(j)})^2$$

De toename in de residuele kwadratensom is dus evenredig met het kwadraat van de Euclidische afstand tussen de centroïden van C_i en C_j.

De hiërarchische clusteringsmethoden, die hier worden beschreven, zijn over het algemeen niet optimaal. D.w.z. wanneer men voor een gegeven aantal van m_0 clusters de betreffende clustering bekijkt, dan is die clustering niet de best mogelijke oplossing voor m_0 clusters. Dit geldt ook voor de methode van WARD. De clustering met m_0 clusters, die deze methode vindt, is over het algemeen niet die clustering met m_0 clusters met de kleinste residuele kwadratensom. Hoewel de methode van WARD niet optimaal is, blijkt in de praktijk de gevonden oplossing vaak dicht in de buurt te liggen van de optimale oplossing.

Recursieformule:
Als clusters C_i en C_j worden samengevoegd en C_k is een andere cluster, dan geldt:

$$D(C_i \cup C_j, C_k) = \frac{n_i+n_k}{n_i+n_j+n_k} D(C_i,C_k) + \frac{n_j+n_k}{n_i+n_j+n_k} D(C_j,C_k) - \frac{n_k}{n_i+n_j+n_k} D(C_i,C_j)$$

Immers

$$D(C_i \cup C_j, C_k) = \frac{(n_i+n_j)n_k}{(n_i+n_j+n_k)} \sum_{l=1}^{p} \left(\frac{n_i\bar{x}_l^{(i)}+n_j\bar{x}_l^{(j)}}{n_i+n_j} - \bar{x}_l^{(k)} \right)^2 =$$

$$= \frac{n_k}{(n_i+n_j)(n_i+n_j+n_k)} \sum_{l=1}^{p} (n_i^2(\bar{x}_l^{(i)})^2 + n_j^2(\bar{x}_l^{(j)})^2 + 2n_i n_j \bar{x}_l^{(i)}\bar{x}_l^{(j)} -$$

$$- 2(n_i+n_j)(n_i\bar{x}_l^{(i)}+n_j\bar{x}_l^{(j)}) + (n_i+n_j)^2(\bar{x}_l^{(k)})^2) =$$

$$= \frac{n_k}{(n_i+n_j)(n_i+n_j+n_k)} \sum_{l=1}^{p} (n_i(n_i+n_j)(\bar{x}_l^{(i)}-\bar{x}_l^{(k)})^2 + n_j(n_i+n_j)(\bar{x}_l^{(j)}-\bar{x}_l^{(k)})^2 -$$

$$- n_i n_j(\bar{x}_l^{(i)}-\bar{x}_l^{(j)})^2) =$$

$$= \frac{n_i n_k(n_i+n_j)}{(n_i+n_j)(n_i+n_j+n_k)} \sum_{l=1}^{p} (\bar{x}_l^{(i)}-\bar{x}_l^{(k)})^2 + \frac{n_j n_k(n_i+n_j)}{(n_i+n_j)(n_i+n_j+n_k)} \sum_{l=1}^{p} (\bar{x}_l^{(j)}-\bar{x}_l^{(k)})^2 -$$

$$- \frac{n_i n_j n_k}{(n_i+n_j)(n_i+n_j+n_k)} \sum_{l=1}^{p} (\bar{x}_l^{(i)}-\bar{x}_l^{(j)})^2 =$$

$$= \frac{(n_i+n_k)}{(n_i+n_j+n_k)} D(C_i,C_k) + \frac{(n_j+n_k)}{(n_i+n_j+n_k)} D(C_j,C_k) - \frac{n_k}{(n_i+n_j+n_k)} D(C_i,C_j)$$

In onderstaand voorbeeld wordt de methode van WARD geïllustreerd. Wanneer we van de recursieformule gebruik willen maken dan moet in de oorspronkelijke afstandsmatrix der objecten de afstand van de objecten gedefinieerd zijn als de helft van het kwadraat van de Euclidische afstand, aangezien moet gelden $d_{ij} = D(\{o_i\},\{o_j\})$. Dit impliceert

$$d_{ij} = SSE(\{o_i\},\{o_j\}) - SSE(\{o_i\}) - SSE(\{o_j\}) =$$

$$= \sum_{l=1}^{p} ((x_{il} - \frac{x_{il}+x_{jl}}{2})^2 + (x_{jl} - \frac{x_{il}+x_{jl}}{2})^2) - 0 - 0 =$$

$$= 2 \sum_{l=1}^{p} (\frac{x_{il}-x_{jl}}{2})^2 = \tfrac{1}{2} \sum_{l=1}^{p} (x_{il}-x_{jl})^2$$

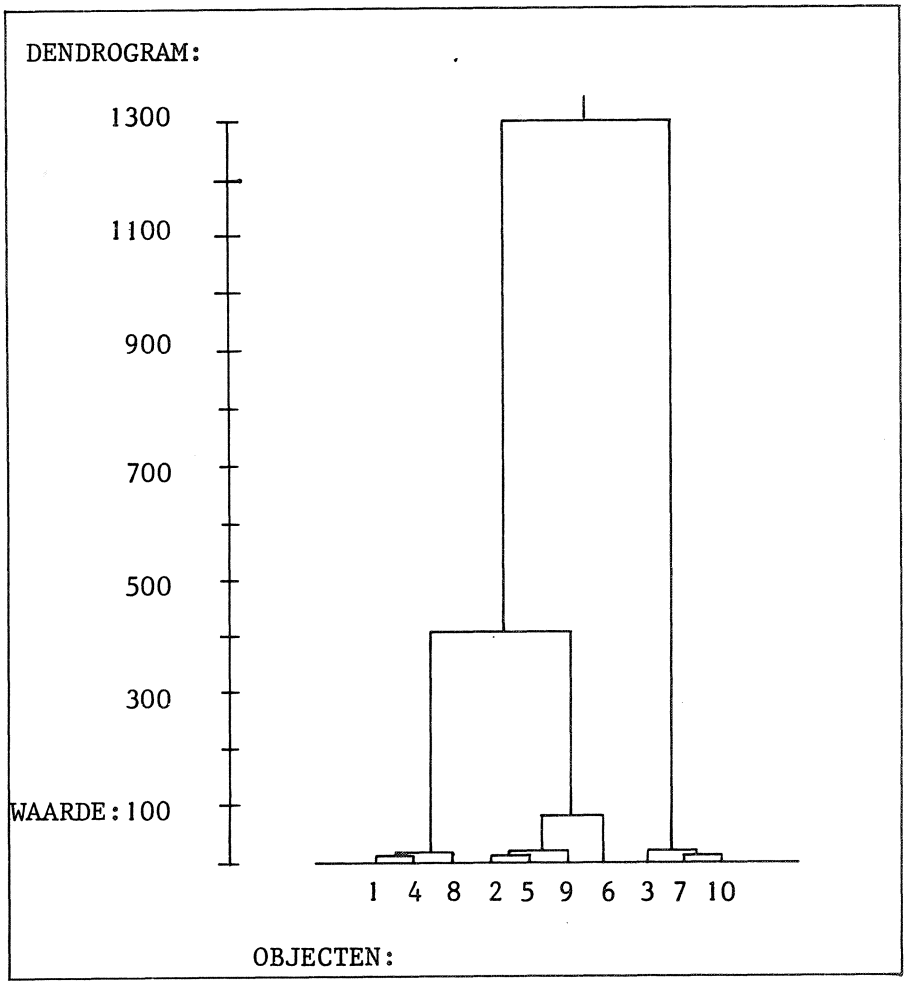
Als we uitgaan van dezelfde getallen als bij de centroïde- en mediaan-methode dan krijgen we de volgende dissociatiematrix der objecten:

	1	2	3	4	5	6	7	8	9	10
1	0,0									
2	83,2	0,0								
3	571,2	218,4	0,0							
4	0,4	72,0	541,2	0,0						
5	96,6	0,5	198,0	84,5	0,0					
6	285,6	60,5	49,0	264,5	50,0	0,0				
7	609,0	242,0	0,6	578,0	220,5	60,5	0,0			
8	1,6	61,6	512,0	0,4	73,2	244,2	547,8	0,0		
9	111,0	2,0	178,6	98,0	0,5	40,5	200,0	85,8	0,0	
10	648,0	266,8	2,4	616,0	244,2	73,2	0,6	548,8	222,6	0,0

We krijgen nu het volgende clusteringsproces:

STAP	MEEST GELIJKENDE CLUSTERS	AFSTAND	NIEUWE CLUSTER
1	{1},{4}	0,4	{1,4}
2	{2},{5}	0,5	{2,5}
3	{7},{10}	0,6	{7,10}
4	{1,4},{8}	1,2	{1,4,8}
5	{2,5},{9}	1,5	{2,5,9}
6	{3},{7,10}	1,8	{3,7,10}
7	{2,5,9},{6}	75,0	{2,5,6,9}
8	{1,4,8},{2,5,6,9}	411,9	{1,2,4,5,6,8,9}
9	{1,2,4,5,6,8,9},{3,7,10}	1327,5	{1,2,3,4,5,6,7,8,9,10}

De methode van WARD geeft toevalligerwijs dezelfde oplossing als de centroïde-methode, maar de afstanden lopen hier veel sterker uiteen, wat de keuze voor het moment van afbreken van het proces gemakkelijker maakt.



8. ALGEMENE RECURRENTE BETREKKING

LANCE & WILLIAMS [13] en WISHART [21] hebben een algemene recurrente betrekking gevonden, die uitgaande van de matrix van associaties/dissociaties tussen de clusters in een bepaalde clustering en de twee samen te voegen clusters, de associaties/dissociaties van de nieuwe cluster en de oude cluster bepaalt. Met behulp van deze formule kunnen, door invullen van de parameters, alle hiervoor beschreven methoden samengevat worden.

Laten C_i en C_j twee clusters zijn, die samengevoegd worden en zij C_k een andere cluster met welke de associatie/dissociatie berekend moet worden. Zij verder:

n_i = aantal objecten in C_i,
n_j = aantal objecten in C_j,
n_k = aantal objecteb in C_k,

$$\text{sgn}(i,j) = \begin{cases} +1, \text{ als } & S(C_i,C_j) > 0, \\ 0, \text{ als } & S(C_i,C_j) = 0. \\ -1, \text{ als } & S(C_i,C_j) < 0. \end{cases}$$

Dan luidt de formule in termen van associatiematen:

$$S(C_i \cup C_j, C_k) = \alpha.S(C_i,C_k) + \beta.S(C_j,C_k) + \gamma.S(C_i,C_j) + \delta|S(C_i,C_k) - S(C_j,C_k)|$$

De waarden van α,β,γ en δ staan voor de diverse methoden in onderstaand schema, met daarbij de eventuele beperkingen

METHODE	α	β	γ	δ	BEPERKINGEN
minimum	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$	-
maximum	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$	-
groepsgemiddelde	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	0	0	ass.maat moet positief zijn.
ELSHOUT	$\frac{n_i}{n_i+n_j}$	$\frac{n_j.\text{sgn}(i,j)}{n_i+n_j}$	0	0	voor variabelen met correlatie als associatiemaat.
centroïde	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	$-\alpha\beta$	0	alleen onderzoeks-eenheden met kwadraat Eucl. afstand.
mediaan	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0	
WARD	$\frac{n_i+n_k}{n_i+n_j+n_k}$	$\frac{n_j+n_k}{n_i+n_j+n_k}$	$\frac{-n_k}{n_i+n_j+n_k}$	0	alleen onderzoeks-eenheden met helft van kwadraat Eucl. afstand.

Het is natuurlijk mogelijk om nog andere methoden van hiërarchische clusteranalyse te bedenken door simpelweg zelf waarden voor de parameters α,β,γ en δ te definiëren. Voor nadere overwegingen hierover zie bijvoorbeeld LANCE & WILLIAMS [13].

In veel van de bestaande computerprogrammatuur voor hiërarchische clusteranalyse wordt bovenstaande recurrente betrekking gebruikt. De gebruiker moet op zijn hoede zijn, dat niet elke willekeurige associatie- of dissociatiematrix gebruikt zal kunnen worden. Zie voor de beperkingen het voorgaande schema.

9. COMPUTERPROGRAMMA'S

Op het Mathematisch Centrum is op het ogenblik een pakketje programma's beschikbaar voor het uitvoeren van hiërarchische clusteranalyse. De twee belangrijkste programma's zijn DISTANC en CLUSTER. Het programma DISTANC kan uitgaande van een ruwe datamatrix een groot aantal typen associatie- of dissociatiematrix maken; Het programma CLUSTER kan uitgaande van een associatie- of dissociatiematrix een van de zeven beschreven methoden van hiërarchische clusteranalyse uitvoeren.

10. LITERATUUR

[1] ANDERBERG, M.R., *Cluster Analysis for Applications*, Academic Press, (1973).

[2] EVERITT, B., *Cluster Analysis*, Heineman Educational Books, (1974).

[3] GOODMAN, L.A. & W.H. KRUSKAL, *Measurement of Association for Cross Classifications*, J. Amer. Statist. Assoc. 49 (1954) p. 732-764.

[4] GOWER, J.C., *Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis*, Biometrika 53 (1966) p. 325-338.

[5] GOWER, J.C., *Multivariate Analysis and Multidimensional Geometry*, The Statistician 17 (1967) p. 13-25.

[6] GOWER, J.C., *A Comparison of Some Methods of Cluster Analysis*, Biometrics 23 (1967) p. 623-628.

[7] GOWER, J.C., *A General Coefficient of Similarity and Some of Its Properties*, Biometrics 27 (1971) p. 857-872.

[8] JARDINE, N. & R. SIBSON, *Mathematical Taxonomy*, John Wiley & Sons, 1971.

[9] JOHNSON, S.C., *Hierarchical Clustering Schemes*, Psychometrika, 32 (1967) p. 241-254.

[10] KENDALL, M.G., *Rank Correlation Methods*, Griffin, (1955).

[11] KING, B., *Stepwise Clustering Procedures*, J. Amer. Statist. Assoc. 62 (1967) p. 86-101.

[12] KRUSKAL, W.H., *Ordinal Measures of Association*, J. Amer. Statist. Assoc. 53 (1958) p. 814-861.

[13] LANCE, G.N. & W.T. WILLIAMS, *A General Theory of Classificatory Sorting Schemes*, 1 Hierarchical Systems, Comp. J. 9 (1967) p. 373-380.

[14] MORRISON, D.F., *Multivariate Statistical Methods*, McGraw-Hill, (1967).

[15] MORRISON, D.G., *Measurement Problems in Cluster Analysis*, Manag. Science 13 (1967) p. 775-780.

[16] SNEATH, P.H.A., *The Applications of Computers to Taxonomy*, J. Gen. Microbiol. 17 (1957) p. 201-226.

[17] SOKAL, R.R. & C.D. MICHENER, *A Statistical Method for Evaluating Systematic Relationships*, Univ. Kansas Science Bull. 38 (1958) p. 1409-1438.

[18] SOKAL, R.R. & P.H.A. SNEATH, *Principles of Numerical Taxonomy*, Freeman, 1963.

[19] WARD, J.H., *Hierarchical Grouping to Optimize an Objective Function*, J. Amer. Statist. Assoc. 58 (1963) p. 236-244.

[20] WILLIAMS, W.T. & M.B. DALE, *Fundamental Problems in Numerical Taxonomy*, in: Advances in Botanical Research, Academic Press, 1965.

[21] WISHART, D., *An Algorithm for Hierarchical Classifications*, Biometrics 25 (1969) p. 165-170.