Meer-sets analyse voor kwalitatieve gegevens

Doktoraalskriptie Renée Verdegaal
Vakgroep Datatheorie
Rijksuniversiteit Leiden

april 1985

Deze doktoraalskriptie voor de studie Psychologie met afstudeerrichting Methoden en Technieken omvat een gedeelte van mijn werkzaamheden bij de vakgroep Datatheorie van de Fakulteit de Sociale Wetenschappen aan de Rijksuniversiteit te Leiden.

Ik zou de volgende personen graag willen bedanken : Jan de Leeuw voor begeleiding en bijdragen en Eeke van der Burg voor uitleg en ondersteuning. Ook wil ik Leo van der Kamp, Wim van der Kloot en Ab Mooyaart bedanken voor hun kommentaar op de eerste versie van dit rapport.

Renée Verdegaal
Leiden, april 1985

INHOUDSOPGAVE

Overzicht

Binnen de psychometrie wordt een tak van technieken onderscheiden, die onder de noemer multivariate analyse vallen. Het basisdoel van deze technieken is datareduktie m.a.w. er wordt geprobeerd de geobserveerde variabelen zo spaarzaam mogelijk te herschrijven. Hiertoe worden de ruwe gegevens zo gekombineerd, dat er meer inzicht verschaft wordt in de struktuur van de data. Een van de leden van deze familie van multivariate technieken is de kanoniese korrelatie analyse. In deze skriptie zal op deze techniek worden ingegaan. Eerst wordt een korte uiteenzetting gegeven van de lineaire kanoniese korrelatie analyse zoals deze ontwikkeld is door Hotelling (1936) en hoe een mogelijke uitbreiding naar meer dan twee groepen (sets) variabelen verwezenlijkt kan worden (Horst, 1961; Carroll, 1968; Kettenring, 1971; Van der Geer, 1984).
Daarna wordt dieper ingegaan op een generalisatie van lineaire kanoniese korrelatieanalyse, waarbij naast lineaire ook niet-lineaire transformaties van de variabelen zijn toegestaan (De Leeuw, 19 3; Gifi, 1981). De uitwerking gekozen door Gifi en De Leeuw van meer-sets analyse krijgt vooral de aandacht. Hier komen de belangrijke aspekten waarop de techniek gebaseerd is aan de orde zoals de kleinste kwadraten verliesfunktie, de rang- en kegelrestrikties om de enkelvoudige kwantifikaties te definiëren en de orthogonaliteitseisen en normalisaties alsook de gekozen behandeling van ontbrekende gegevens (missing data). Verder zullen de verbanden met principale komponenten analyse en homogeniteitsanalyse duidelijk gemaakt worden en bovendien wat de mogelijkheden van een dergelijke analyse techniek zijn en de voordelen van zo'n analyse boven andere analyse technieken. Om een dergelijke niet-lineaire meer-sets analyse uit te kunnen voeren heb ik in samenwerking met Eeke van der Burg het komputerprogramma OVERALS geschreven. In het laatste hoofdstuk worden twee voorbeelden behandeld waarin alles nog eens verhelderd wordt.

Notatie en terminologie.

n aantal objekten/individuen

m totaal aantal variabelen

K aantal sets/groepen variabelen

m_j aantal variabelen behorend tot de k-de set, $\sum_{k=1}^{K} m_j = m$

H (n x m) matriks van ruwe data

H_k (n x m_j) matriks van ruwe data behorend tot set k

h_j (n x 1) ruwe observaties van variabele j

k_j aantal kategorieën van variabele j

G_j (n x k_j) de indikatormatriks van variabele j

De indikatormatriks is een binaire matriks, die op de volgende manier gedefiniëerd is :

$(G_j)_{ir} = 1$, als het i^{de} objekt in de r^{de} kategorie van variabele j valt.

$(G_j)_{ir} = 0$, in de andere gevallen.

Deze indikatormatriksen kunnen in een supermatriks verenigd worden. Een klein voorbeeld hiervan

$$H = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 1 \\ 2 & 2 \end{bmatrix} \qquad G = (G_1 : G_2) = \left[\begin{array}{ccc:cc} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{array}\right]$$

G (n x Σk_j) de superindikatormatriks van alle variabelen j

G^k (n x $\sum_{j \varepsilon I_k} k_j$) de superindikatormatriks voor variabelen behorend tot I_k

I_k indeksverzameling voor alle variabelen j, die tot groep k behoren

u een vektor met elementen gelijk aan 1

I de eenheidsmatriks

D_j de diagonaalmatriks bevattend de univariate marginalen van variabele j

$C_{j\ell}$ de kruistabel van variabele j en ℓ, bevattend de bivariate marginalen

p aantal dimensies (oplossingen)

a_j (p x 1) de gewichten per dimensie voor variabele j

a_k (m_j x 1) de gewichten voor de variabelen behorend tot set k voor één oplossing

A_k (m_j x p) de matriks van gewichten behorend tot set k (p dimensies)

t^j de niet-lineaire transformatie van variabele j

q_j niet-lineaire transformatie van h_j

Q_k (n x m_j) matriks van getransformeerde data behorend tot set k

X (n x p) matriks van objektskores

M_k (n x n) binaire diagonale matriks, die aangeeft welke observaties er binnen set k ontbreken

M_* (n x n) geeft gemiddelde van M_k, $M_* = \sum_{k=1}^{K} M_k / K$

In eerste instantie wordt uitgegaan van de komplete indikatormatriks. Dit houdt in dat elke rij één 1 bevat en voor de rest nullen. Dus de som van één rij van een indikatormatriks G_j is één. Dit geldt voor alle rijen, dus er kan ook geschreven worden $G_j u = u$, waar u een eenheidvektor voorstelt. De kolomtotalen $G_j' u$ worden d_j genoemd. Dit zijn de marginale frekwenties van de variabelen in de datamatriks. Deze d_j zijn de elementen van de diagonale matriks $D_j = G_j' G_j$. De som van de univariate marginale d_j is gelijk aan het aantal objekten. $d_j' u = n$.
Zijn er ontbrekende gegevens, dan is de indikatormatriks niet meer kompleet. Er zullen dan ook rijen voorkomen, die enkel nullen bevatten.

Y_j (k_j x p) De meervoudige kategoriekwantifikaties voor variabele j

y_j (k_j x 1) De enkelvoudige kategoriekwantifikaties voor variabele j

Γ^j (p x p) De matriks van komponentladingen (voor meervoudige variabelen)

b_j (p x 1) De komponentladingen voor enkelvoudige variabelen

Ξ^j (p x p) Totale dipersie binnen een set zonder variabele j

Δ^j (p x p) Diskriminatiematriks van variabele j

tr = trace

diag = diagonaal

Hoofdstuk 1 : Metriese relaties tussen K groepen variabelen.

1.1 Kanoniese korrelatie analyse voor twee groepen variabelen (K=2)

Een van de multivariate statistiese technieken is de kanoniese korrelatie analyse. Deze techniek wordt gebruikt om de samenhang tussen twee groepen variabelen te bepalen. Dit zou bijvoorbeeld het geval kunnen zijn als schoolprestatiegegevens van een aantal kinderen gerelateerd moeten worden aan hun achtergrondgegevens, zoals opleidingsnivo en beroep van de vader en/of moeder, het wel/niet kijken naar televisie enz.
In Hotelling (1935) wordt een asymmetriese benadering voorgesteld : De ene set bestaat uit prediktor variabelen en de andere set uit kriteria. Hotelling geeft vervolgens verschillende manieren aan om de kriteria variabelen te verenigen tot één variabele, waarna multiple regressie toegepast kan worden. Een van de methoden die hij suggereert is om de kriteria te vervangen door de eerste principale komponent.
Later in Hotelling (1936) geeft hij een uitwerking van wat nu onder kanoniese korrelatie analyse wordt verstaan, en worden beide sets op een meer symmetriese manier behandeld.
In woorden is het kanoniese korrelatie probleem als volgt te omschrijven : Zoek van twee groepen variabelen, elk met een theoretiese betekenis als groep dié lineaire kombinatie van de ene groep variabelen, die maksimaal korreleert met met een andere lineaire kombinatie in de tweede groep variabelen. Zo'n lineaire kombinatie staat ook bekend als kanoniese variabele of kanoniese variaat. Is dit eerste paar kanoniese variaten bepaald, dan wordt op dezelfde manier een tweede paar kanoniese variaten afgeleid, dus er wordt weer een zo hoog mogelijke interkorrelatie tussen de twee groepen gezocht, maar alleen voor het deel dat nog niet voor rekening van het eerste paar is gekomen. M.a.w. er wordt geëist dat de tweede richting orthogonaal is met de eerste richting.
Op deze manier kunnen er net zoveel paren worden afgeleid als het aantal variabelen in de kleinste set.
In matriks notatie : Van een matriks H_1 met afmetingen n x m_1 bevattend de gegevens van n observaties voor de m_1 variabelen uit de eerste set en een matriks H_2 (n x m_2) voor de tweede set variabelen, spannen de kolommen beide een lineaire deelruimte, zeg L_1 voor de eerste set en L_2 voor

de tweede set, op. Zoek nu richtingen in L_1 en L_2 die maksimaal korresponderen. Deze richtingen zijn één-dimensionale deelruimten van L_1 en L_2 en worden kanoniese variabelen genoemd. Ze zullen hier aangeduid worden als H_1a_1 en H_2a_2, waarbij de a_1 en de a_2 de gewichten van de m_1 en de m_2 variabelen uit respektievelijk de eerste en de tweede set bevatten. Met inachtneming van de orthogonaliteitsrestrikties kunnen nog p (met $p \leq$ minimum (m_1, m_2)) lineaire kombinaties van H_1 en H_2 gevonden worden. De gewichten zijn nu geen vektoren meer, maar matriksen, genoemd A_1 $(m_1 \times p)$ A_2 $(m_2 \times p)$.

Dus minimaliseer

$$\text{trace } (H_1A_1 - H_2A_2)'(H_1A_1 - H_2A_2)/p. \qquad (1)$$

met orthogonaliteitsrestrikties

$$A_1'H_1'H_1A_1 = nI \quad \text{of} \quad A_2'H_2'H_2A_2 = nI$$

en normalisatie

$$h_j'h_j = n \quad \text{en} \quad h_j'u = 0$$

De gelijkenis tussen de sets wordt uitgedrukt in de diagonaal van de matriks $A_1'H_1'H_2A_2/n$. Deze diagonaalelementen worden de kanoniese korrelatie koëfficienten genoemd.

De gevonden kanoniese variaten zijn in essentie gelijk aan de principale komponenten zoals ze geproduceerd worden door principale komponenten analyse (PCA) te doen, met uitzondering dat het kriterium voor de selektie veranderd is. Waar beide technieken lineaire kombinaties van de originele variabelen berekenen, doet kanoniese korrelatie analyse dit niet door zoveel mogelijk variantie binnen groepen te bevatten, maar met als doel de maksimale relatie tussen de beide groepen te vinden. De korrelatie tussen elk korresponderend paar kanoniese variaten is de kanoniese korrelatie. Het kwadraat van de korrelatie representeert de hoeveelheid variantie in een kanoniese variaat, die door een andere kanoniese variaat wordt verklaard.

1.2 Kanoniese korrelatie analyse voor meer dan twee groepen variabelen ($K \geq 3$).

1.2.1 Het meer-sets probleem : lineair.

Vaak is het mogelijk dat het te analyseren probleem meer dan twee groepen variabelen omvat. Bijvoorbeeld in de eerder genoemde situatie waar schoolprestatiegegevens en achtergrondgegevens van een aantal leerlingen zijn verzameld. Het zou gezien de vraagstelling wenselijk kunnen zijn deze achtergrondgegevens te splitsen in een set van variabelen met gegevens over de vader en een set met gegevens over de moeder, waardoor de beschikking is verkregen over drie groepen variabelen. Ook is het mogelijk dat er nog andere gegevens over deze groep leerlingen verzameld zijn zoals attituden t.o.v. de school, de lessen en de leraren . Het kan interessant zijn deze variabelen te relateren aan andere groepen variabelen.
Naast dit probleem zijn er nog vele situaties te bedenken waarin we te maken hebben met metingen, die vanuit theoreties oogpunt op te splitsen zijn in meer dan twee groepen variabelen.
Een meer algemene formulering van het probleem kan als volgt gegeven worden : Een steekproef van n observaties/objekten laat zich op inhoudelijke grond indelen in K groepen variabelen. Zodoende zijn er K matriksen H_k (k=1,...,K) van n rijen en m_j kolommen, namelijk de k^{de} groep bevat m_j variabelen. De breedte m_j kan voor iedere matriks H_k verschillend zijn. Verder is de som van de variabelen in de verschillende sets gelijk aan het totale aantal variabelen: $\Sigma m_j = m$.
Laat a_k (m_j x 1) een vektor van gewichten voor H_k zijn, en definiëer $q_k = H_k a_k$. Zoek dan onder van te voren gekozen optimaliteitskriteria die K transformaties van de matriksen H_k, die een zo groot mogelijke overeenkomst tussen de q_k geven. Op deze manier wordt een zogenoemde enkelvoudige oplossing gevonden. Het kan zijn dat de stationaire vergelijkingen meer dan één oplossing hebben. Of het is mogelijk additionele oplossingen te vinden door het probleem nogmaals op te lossen, maar dan bijv. met toegevoegde orthogonaliteitseisen die zorgen dat eerdere oplossingen niet opnieuw verkregen worden. Er kan onderscheid gemaakt worden tussen sterke en zwakke orthogonaliteitseisen (Dauxois et Pousse, 1976). Bij de sterke orthogonaliteitseisen wordt er geëist:

$$q_k' q_k^{(s)} = a_k' D_k a_k^{(s)} = 0, \quad \text{waarbij } D_k = H_k' H_k$$

voor alle k en alle eerdere oplossingen $a_k^{(s)}$. Bij de zwakke orthogonaliteitseisen wordt geëist dat

$$\sum_{k=1}^{K} q_k' q_k^{(s)} = \sum_{k=1}^{K} a_k' D_k a_k^{(s)} = 0,$$

voor alle eerdere oplossingen $a_k^{(s)}$.

Bij toepassing van de sterke orthogonaliteitseisen zijn er m_j mogelijke oplossingen voor een set met m_j variabelen. Daarna wordt $a_k = 0$ voor sets waar p (het aantal oplossingen) groter is dan m_j.

Deze *successieve* manier voor het vinden van additionele oplossingen kan onder andere aangetroffen worden bij Horst (1936), Edgerton and Kolbe (1936), Wilks (1938), Guttman (1941), Kettenring (1971) en Dauxois en Pousse (1976) en Van der Geer (1984).

Door Eckhart and Young (1936) werd een andere manier geïntroduceerd om additionele oplossingen te berekenen, namelijk *simultaan*. Het Eckhart-Young theorema zegt in de kontekst van principale komponenten analyse dat een n x m data matriks H geformuleerd kan worden in termen van een verliesfunktie

$$\sigma(X,A) = SSQ(H - XA').$$

waar X een n x p matriks is van rang p, en waar A is m x p. Gifi (1981, appendix A) laat zien dat de optimale X en A, die de hierboven gedefinieerde verliesfunktie minimaliseren, gevonden kunnen worden door een singuliere waarde dekompositie (SVD) op de matriks H toe te passen.

Een andere mogelijke berekeningsprocedure is die van de Alternating Least Squares (ALS) waarbij de stappen

$$X \leftarrow HA(A'A)^{+}$$

en

$$A \leftarrow H'X(X'X)^{+}$$

afgewisseld worden. Het superskript + indiceert de Moore-Penrose inverse van een matriks (Gifi 1981, appendix A).

Om in de meer-sets kontekst de simultane berekening te introduceren wordt de matriks $Q_k = H_k A_k$ gedefinieerd, met A_k een m_j x p matriks. Verder worden de matriksen $R_{k\ell} = A_k' C_{k\ell} A_\ell$ (met $C_{k\ell} = H_k' H_\ell$) verzameld in een supermatriks R(Q) van dimensie (Kp) x (Kp). Meestal zijn de optimaliteitskriteria funkties van deze supermatriks R(Q) of van zijn eigenwaarden. Ook kunnen nu de sterke of zwakke orthogonaliteitseisen opgelegd worden. De sterke orthogonaliteitseisen vereisen dan dat $R_{kk} = I$ voor alle k en de zwakke versie eist dat $\Sigma R_{kk} = KI$.

Deze simultane berekeningsprocedures geven een natuurlijkere manier om multidimensionale oplossingen te introduceren dan de successieve. Als de oplossingen voor de verschillende dimensionaliteiten *genest* zijn d.w.z dat de eerste kanoniese variaat van een p=1 oplossing identiek is aan de eerste kanoniese variaat van een p>1 oplossing, dan geven de simultane en de successieve procedures dezelfde oplossingen.

1.2.2 De optimaliteitskriteria.

Voor een probleem waarbij er meer dan twee groepen variabelen zijn zal gekozen moeten worden welke kriteria de voorkeur hebben om dit probleem op te lossen. Hiermee hebben zich o.a. Steel (1951), Horst (1961a,b;1965), Carroll (1968), Kettenring (1971), Dauxois et Pousse (1976), Van de Geer (1984) en Gifi (1981) mee bezig gehouden. Als regel kiezen zij p=1, en werken zij successief.
Steel stelt voor om de determinant van de korrelatie matriks R(Q), welke gelijk is aan het produkt van zijn eigenwaarden, te minimaliseren. Bij Kettenring (1971) komen we deze benadering weer tegen onder de naam GENVAR, en bovendien stelt Kettenring een algoritme voor voor de oplossing van de stationaire vergelijkingen, welke veel overeenkomst met het 'Alternating Least Squares" (ALS) principe uit Gifi (1981) vertoond.
Horst (1961a,b) brengt vier verschillende modellen naar voren.
In het eerste model wordt op die manier een vektor transformatie van elke set gezocht, dat de som van de interkorrelaties tussen de getransformeerde variabelen maksimaal is. $\Sigma R(Q)$ maksimaal. Dit kriterium is dus geen funktie van de eigenwaarden. Bovendien is het afhankelijk van het teken van de elementen in R(Q).
Dit model wordt door Kettenring met het acroniem SUMCOR aangeduid. En in de kontekst van het in overeenstemming brengen van konfiguraties met orthogonale rotaties is deze methode ook bediskussieerd door Van de Geer (1980) onder de naam ORTHOCAN.
Het tweede door Horst voorgestelde model specificeert dat de interkorrelaties van de eerste getransformeerde variabele voor de K sets, de beste kleinste kwadraten benadering aan een rang-één matriks zal geven. Ditzelfde geldt voor alle volgende getransformeerde variabelen en tegelijkertijd moeten deze getransformeerde variabelen allen orthogonaal zijn. Voor dit model gebruikt Kettenring het acroniem MAXVAR. Hij interpreteert het als het maksimaliseren van de grootste eigenwaarde van R(Q).

Ook bij McKeon (1966), McDonald (1968) en Carroll (1968) vinden we deze methode terug. De benadering die Carroll gebruikt is de moeite waard om apart vermeld te worden. Hij gebruikt namelijk een extra parameter x om het kriterium te definiëren en legt de orthogonaliteitsrestrikties aan deze parameter op.

Carroll definieert het probleem op de volgende manier: Zoek een vektor x (met n elementen, die optellen tot nul) en K transformatievektoren a_k op zo'n manier dat de som van de gekwadrateerde korrelaties tussen x en $H_k a_k$ maksimaal is. In formule

$R^2 = \sum_{k=1}^{K} w_j \{r(x, H_k a_k)\}^2$ moet maksimaal zijn. r symboliseert de produkt-moment korrelatie en w_k is een wegingsfaktor voor de k^{de} set van variabelen.

En de beste oplossing van x, zorgend voor een maksimale R^2 wordt gegeven door de eigenvektor die behoort bij de grootste eigenwaarde van S, met $S = \sum_{k=1}^{K} w_k H_k' (H_k' H_k)^{-1} H_k$, een positief definiete of semi-definiete matriks. (als alle w_k positief zijn). Een volgende kanoniese x variaat orthogonaal aan de eerste wordt gevonden door de tweede eigenvektor van S, en zo kun je evenveel kanoniese variaten bepalen als de rang van S. De korresponderende eigenwaarden geven de waarden van R^2.

Gifi (1981) en De Leeuw (1984) maken ook gebruik van deze extra parameter X, maar i.p.v. dat zij de gewogen gemiddelde som van de gekwadrateerde korrelaties tussen X en de lineaire kombinaties van de sets maksimaliseren, minimaliseren zij het verlies tussen X en $H_k A_k$. De matriks X (n x p) bevat de kanoniese variaten x. En de oplossingen a_k zijn per set verzameld in matriksen A_k (m_j x p). Beide benaderingen leveren wel dezelfde oplossing op. In formule kan dit probleem als volgt worden opgeschreven:

$$\sigma(X,A) = \sum_{k=1}^{K} \text{trace}(X - H_k A_k)'(X - H_k A_k)$$

onder de kondities u'X = 0 en X'X = nI. De vektor u is een kolomvektor bestaande uit énen en I is de identiteitsmatriks. u'X = 0 garandeert dat X in afwijking van de kolomgemiddelden staat en X'X = nI verzekert orthogonaliteit.

Een volgend door Horst voorgesteld model begint bij de beschouwing van de beste rang-m_v benadering van alle K groepen variabelen. m_v is het aantal variabelen uit de kleinste set. De beste kleinste kwadraten or-

thonormale faktorskorematriks van breedte m_v wordt uit de superset samengesteld uit de K deelsets bepaald. Daarna wordt een transformatie voor elke subset gevonden, die de beste kleinste kwadraten fit aan de faktorskorematriks geeft.

Het vierde model van Horst is gelijk aan het derde behalve dan dat de additionele eis van orthonormaliteit wordt opgelegd aan de transformatie matriksen.

Naast de al genoemde modellen die Kettenring (1971) behandelt, introduceert hij ook nog twee andere methoden. Een hiervan noemt hij SSQCOR, waarbij het kriterium het maksimaliseren van de kwadratensom van de korrelaties in R(Q) is oftewel de kwadratensom van de eigenwaarden van R(Q). De andere wordt door hem met het acroniem MINVAR aangeduid en deze minimaliseert de kleinste eigenwaarde van R(Q).

Van de Geer (1984) geeft een aantal mogelijke oplossingen voor verschillend gekozen kanoniese korrelatie modellen. De modellen hangen af van drie basis keuzes, namelijk

1) Wordt er alleen aandacht besteed aan de relatie tussen de groepen of ook aan de variantie binnen de groepen.
2) Hoeveel nadruk zal er gelegd worden op de verklaring van de variantie binnen de groepen.
3) Moeten alle groepen variabelen een zo gelijk mogelijke bijdrage leveren aan de oplossing of is het ook goed als de oplossing door enkele sets gedomineerd wordt. Dit punt is gerelateerd aan de orthogonaliteitseisen.

Hoofdstuk 2 : Niet-lineaire multivariate technieken.

2.1 Inleiding

In de ontwikkeling van de niet-lineaire wegingstechnieken, voor de kwantifikatie van kategoriese variabelen, verschijnen al rond 1900 bijdragen van Pearson (1901, 1904, 1906). Echte methoden om deze kwantifikaties te realiseren komen pas veel later tot ontwikkeling. De methode gebruikt door Gifi (1981) is gebaseerd op het werk van Guttman (1941) en Burt (1950), die onafhankelijk van Guttman de toepassing van principale komponenten analyse op kategoriese gegevens ontdekte. Naast Guttman en Burt hebben ook anderen zoals Fisher (1940), Hayashi (1950, 1956), de Leeuw (1973), Benzécri (1973) en Nishisato (1980) zich met technieken voor meervoudige kwantifikatie bezig gehouden. Deze meervoudigheid houdt in dat voor iedere dimensie van de oplossing nieuwe niet-lineaire transformaties worden berekend.

Naast de meervoudige benadering is er ook een enkelvoudige benadering te onderscheiden. In het enkelvoudige geval moeten de diverse transformaties op een bepaalde manier aan elkaar gerelateerd zijn. Deze enkelvoudige benadering is onder andere terug te vinden bij Kruskal and Shepard (1974) en bij Takane, Young and de Leeuw (1980).

De Gifi technieken hebben de mogelijkheid de meervoudige en de enkelvoudige benadering door elkaar voor de verschillende variabelen te gebruiken (PRINCALS en OVERALS). De aanpak kan geïnterpreteerd worden als het gebruik van de meervoudige benadering, waarbij voor de enkelvoudige variabelen de restriktie wordt opgelegd dat de transformaties in de verschillende oplossingen proportioneel moeten zijn.

De gebruikte technieken zijn dezelfde als die uit de lineaire analyse. Door de kategorieën van de variabelen te wegen worden de niet-lineaire transformaties gevonden. Hiervoor kan een 1 - 0 kodering gemaakt worden die aangeeft of een observatie wel/niet in een bepaalde kategorie voorkomt. Dus de informatie die in de kategorieën van de kwalitatieve variabelen vervat zit, wordt gerepresenteerd door de kategorieën van een aantal dummy-variabelen.

Zo'n niet-lineaire transformatie definiëert zelf een lineaire ruimte waarvan de lineaire transformaties een deelruimte vormen.

2.2 Niet-lineaire twee-sets analyse versus meer-sets.

Heel kort wordt hier apart het twee-sets probleem beschreven, omdat de aanpak zoals uitgebreid besproken wordt door Van der Burg en De Leeuw (1983) en Van der Burg (1983) afwijkt van de meer-sets aanpak (De Leeuw, 1983; Van der Burg, De Leeuw & Verdegaal, 1984). De niet-lineaire twee-sets theorie, gerealiseerd in het programma CANALS zoekt net als klassieke kanoniese korrelatie analyse voor beide sets gewogen sommen van de variabelen (kanoniese variabelen), die een zo'n hoog mogelijke korrelatie tussen beide sets moet geven. De niet-lineairiteit zit in het feit dat tegelijkertijd optimale transformaties van de gegevens worden gezocht. Deze optimale transformaties kunnen naast de klassieke *enkelvoudige numerieke* transformatie ook *enkelvoudig nominaal* of *ordinaal* zijn. Zie voor een beschrijving van deze transformaties §2.3.3 of (meer uitgebreid) Young, De Leeuw & Takane (1976) of Gifi (1981).
Deze gewogen sommen worden afzonderlijk voor beide sets berekend en zijn ook in de output van CANALS terug te vinden. In het programma OVERALS voor meer-sets kanoniese korrelatie analyse worden deze lineaire kombinaties gesommeerd over sets. Zodoende worden er "gemiddelde" kanoniese variabelen in de output gegeven.
Verder zijn alle variabelen in CANALS enkelvoudig. Zij zijn de nominale, ordinale of numerieke transformaties van de oorspronkelijke kategorieën van de variabele h_j. Noem q_j de niet-lineaire transformatie van h_j en verzamel ze per set in de matriksen Q_1 en Q_2, dan kan vergelijking (1) ook opgeschreven worden als

$$\mathrm{tr}(Q_1A_1 - Q_2A_2)'(Q_1A_1 - Q_2A_2)$$

met $A_1'Q_1'Q_1A_1=nI$ of $A_2'Q_2'Q_2A_2=nI$. Minimaliseren van deze verliesfunktie komt niet helemaal overeen met de OVERALS verliesfunktie, die in §2.3.2 aan de orde komt. Bovendien wordt in OVERALS voor alle variabelen eerst een *meervoudig nominale* oplossing berekend voor elke dimensie. Daarna kan eventueel geeist worden dat deze kwantifikaties voor de verschillende dimensies proportioneel moeten zijn.
Ook de behandeling van ontbrekende gegevens (missing data) verschilt. CANALS gebruikt de '*meervoudige kategorie*' aanpak. Dit betekent dat elke missend gegeven van een variabele als een aparte extra kategorie geanalyseerd wordt. In OVERALS worden de ontbrekende gegevens *'passief'* meegenomen in de analyse. Hiermee wordt bedoeld dat het ontbrekende waarde binnen de set niet gekwantifiseerd wordt en niet bijdraagt aan de verliesfunktie.

2.3 Niet-lineaire relaties tussen K groepen variabelen.

2.3.1 De basis.

De meer sets kanoniese korrelatie analyse zoals gepresenteerd in Gifi (1981) is gebaseerd op de homogeniteits analyse, ook bekend als meervoudige korrespondentie analyse (Guttman 1941, De Leeuw 1973, Benzécri et al 1973, Gifi 1981). Volgens een van te voren gedefinieerde kleinste kwadraten verliesfunktie worden de optimale schalingen voor de objekten X en voor de kategorieën van de variabelen Y_j berekend met behulp van wat ook wel het "eerste centroïd-principe" genoemd wordt. Dit is een van de "principes baricentriques" (Benzécri, 1973). Het houdt in dat X genormaliseerd wordt en dat de kwantifikatie van een kategorie dan gevonden wordt door het zwaartepunt van de objekten die tot die kategorie behoren te berekenen. Met andere woorden: Probeer een speciale representatie X van de objekten en een speciale representatie Y_j van de kategorieën te vinden in een laag dimensionale ruimte R^p te vinden, zodat een objekt relatief dichtbij een kategoriepunt komt te liggen waarin het skoort en relatief veraf van een kategoriepunt waarin het niet skoort. De p duidt de gekozen dimensionaliteit van de ruimte waarin geprojekteerd moet worden aan.
Het verlies kan gemeten worden met de verliesfunktie van homogeniteitsanalyse : (Gifi, 1981b; De Leeuw 1984) :

$$\sigma(X;Y_1,\ldots,Y_m) = \frac{1}{m}\sum_{j=1}^{m} \operatorname{tr}\,(X - G_jY_j)'(X - G_jY_j) \qquad (1)$$

waarbij de objektskores genormaliseerd worden door $u'X = 0$ en $X'X = nI$. u is een kolomvektor met elementen gelijk aan 1 en I is de identiteitsmatriks.

2.3.2 Additiviteitsrestrikties.

Om de stap van homogeniteitsanalyse naar <u>meervoudige</u> meer sets analyse te maken moeten restrikties worden ingevoerd om groepen variabelen te definiëren. Met meervoudig wordt hier bedoeld dat voor elke dimensie een afzonderlijke kwantifikatie voor de kategorieën wordt berekend. De matriks Y_j (k_j x p) is dus van rang p.
De verliesfunktie gaat er dan als volgt uit zien :

$$\sigma(X;Y_1,\ldots,Y_K) = \frac{1}{K}\sum_{s=1}^{K} \operatorname{tr}\,(X - \sum_{j\epsilon I_K} G_jY_j)'(X - \sum_{j\epsilon I_K} G_jY_j) \qquad (2)$$

Waar I_K een indeksverzameling is voor alle variabelen j, die tot groep K behoren.

Er wordt nu dus binnen en over groepen gesommeerd. Als er maar één variabele per groep is dan is eenvoudig te zien dat deze verliesfunktie (2) dezelfde is als die van homogeniteits analyse (1).

Deze restriktie heeft konsekwenties voor het centroïd principe. Dit komt omdat in de meeste gevallen de indikator matriksen G_j binnen set K niet orthogonaal zijn.

Door de meer sets verliesfunktie

$$\sigma(X;Y_1,\ldots,Y_m) = \frac{1}{K}\sum_{s=1}^{K} \text{tr}\,(X - G_jY_j - V_j^k)'(X - G_jY_j - V_j^k) \qquad (3)$$

waar $V_j^k = \sum_{\substack{\ell \in I_K \\ \ell \neq j}} G_\ell Y_\ell$, te differentiëren naar Y_j wordt de optimale $\hat{Y}_j$, gegeven door

$$\hat{Y}_j = D_j^{-1}G_j'(X - V_j^k) \qquad (4)$$

Waarbij $D_j = G_j'G_j$.

Deze meervoudige kategoriekwantifikaties zijn ongelijk aan de kategoriecentroïden, d.w.z. de gemiddelden van de objekten die tot dezelfde kategorie behoren.

$$C_j = D_j^{-1}G_j'X \qquad (5)$$

Ook hier is eenvoudig te zien dat een geval waar slechts sprake is van één variable per groep, en dus de $V_j^k = 0$, de optimale meervoudige kwantifikaties en de centroïden aan elkaar gelijk zijn.

2.3.3 Andere mogelijke restrikties.

In het hierboven beschreven geval waar sprake is van meervoudig nominale variabelen wordt dus voor iedere dimensie p een zo'n optimaal mogelijke kwantifikatie gezocht voor de kategorieën, waarbij optimaal betekent dat het verlies minimaal is, met als enige restriktie dat het gewogen gemiddelde van de kategoriekwantifikaties de oorsprong is. De gewichten zijn hier de marginale frekwenties van de kategorieën. Dit geldt voor meer sets analyse met meervoudig nominale variabelen, waar geen observaties ontbreken.

Naast deze restriktie om groepen variabelen te definiëren kunnen er ook andere restrikties aan de variabelen opgelegd worden. Zoals de rang-één en de kegel restrikties (De Leeuw, 1977, 1984a).

De rang-één restriktie houdt in dat er geeist wordt dat alle kategoriekwantifikaties van een variabele op een lijn moeten liggen die door de

oorsprong loopt in de p-dimensionale ruimte. Dus het opleggen van de rang-één restriktie aan de matriks Y_j van orde k_j x p, waarbij k_j het aantal kategorieën van variabele j voorstelt, betekent dat alle kolommen van Y_j proportioneel aan een vektor (y_j) moeten zijn. Dus in matriksnotatie

$$Y_j = y_j a_j'. \qquad (6)$$

De elementen van de vektor y_j (k_j x 1) worden de enkelvoudige kategoriekwantifikaties genoemd en de a_j (p x 1) zijn de gewichten. Elke variabele heeft p gewichten.
Door de kategoriekwantifikaties op een lijn te projekteren ontstaat er een bepaalde volgorde van de elementen van y_j op die lijn. Met behulp van de zogenoemde kegelrestrikties kan van te voren aangegeven worden aan welke eisen deze volgorde moet voldoen. Zo is het mogelijk de vektor y_j geheel te definiëren, d.w.z. dat zijn elementen gelijk moeten zijn aan een gegeven genormaliseerde vektor. Dit is het geval als de variabele aan het model "enkelvoudig numeriek" moet voldoen. De optimale kategoriekwantifikaties zijn dan lineaire transformaties van de originele kategorieskores. Voor variabelen kan ook geeist worden dat de enkelvoudige kategoriekwantifikaties van laag naar hoog (of omgekeerd) geordend zijn. Dit wordt de "enkelvoudig ordinale" eis genoemd. De "enkelvoudig nominale" variabelen hebben geen verdere restrikties op de vektor y_j. Hier zijn dus naast de rang-één restriktie alleen de normalisatie eisen $u'D_j y_j = 0$ en $y_j'D_j y_j = n$ van toepassing.
Wordt er niet gesommeerd binnen groepen, maar wel gebruik gemaakt van de rang-één en de kegelrestrikties dan wordt een niet-lineaire principale komponenten analyse uitgevoerd (PRINCALS: Gifi 1981, 1983; De Leeuw & Van Rijckevorsel, 1980)

2.3.4 Relatie met lineaire kanoniese korrelatie analyse.

Het is niet moeilijk om te laten zien dat niet-lineaire kanoniese korrelatie analyse (KKA) voor enkelvoudige variabelen in feite hetzelfde is als lineaire KKA, maar dan met toestaan van niet-lineaire transformaties. Door te schrijven

$$G_j Y_j = G_j y_j a_j' = q_j a_j'. \qquad (7)$$

waarbij $q_j = G_j y_j$, is q_j een transformatie van h_j waar alleen gehandhaafd is dat objekten die in de oorspronkelijke datamatriks H tot dezelfde kategorie behoren ook ná de transformatie dezelfde kwantifikatie hebben. Worden deze transformatie q_j en gewichten a_j vervolgens per set in de matriksen Q_k en A_k verzameld dan kan de meer sets verliesfunktie voor enkelvoudige variabelen ook opgeschreven worden als

$$\sigma(X,Q,A) = \frac{1}{K}\sum_{k=1}^{K} \mathrm{tr}(X - Q_k A_k)'(X - Q_k A_k). \qquad (8)$$

Door vervolgens de verliesfunktie van lineaire KKA op te schrijven (Gifi 1981)

$$\sigma(X,A) = \frac{1}{K}\sum_{k=1}^{K} \mathrm{tr}(X - H_k A_k)'(X - H_k A_k). \qquad (9)$$

met X genormaliseerd en H_k de datamatriksen met variabelen gegroepeerd naar set, dan kan daarna t_j als de niet-lineaire transformatie van de variabele j gedefinieerd worden.

$$q_j = t^j(h_j). \qquad (10)$$

Deze $q_j = h_j$, als de transformatie t^j oplegt dat de kategoriekwantifikaties gelijk moeten zijn aan een genormaliseerde vektor. Dus bij de enkelvoudig numerieke restriktie.
Worden de q_j per set k gegroepeerd, dan kan vergelijking (9) als vergelijking (8) geschreven worden.

2.3.5 Behandeling van de niet geobseerveerde gegevens (missing data).

Mocht er voor een individu/objekt op een variabele een observatie ontbreken dan draagt de groep waarvan deze variabele deel uit maakt niet bij aan het verlies. Er wordt per set een binaire diagonale matriks M_k (n x n) gekonstrueerd, die aangeeft welke observaties er binnen een set ontbreken. Als observatie i ontbreekt voor een van de variabelen in set k dan is het diagonale element i van M_k gelijk aan nul. In de gevallen dat er binnen groep k geen ontbrekende gegevens voor i zijn is het betreffende element gelijk aan één.
In de komende theorie wordt deze behandeling van de ontbrekende gegevens opgenomen.

2.3.6 Bepalen van de optimale kategorie kwantifikaties.

Oplossen van het algemene niet-lineaire meer sets probleem vereist een minimalisatie van de verliesfunktie binnen de sets. Dit zou mogelijk zijn door de indikatormatriksen G_j en de matriksen voor de kategoriekwantifikatie Y_j van variabele j binnen sets samen te voegen tot respektievelijk een indikatorsuper matriks G^k en een supermatriks Y^k voor set k en het probleem vervolgens in deze termen op te lossen door het verlies

$$\sigma(X,Y) = \frac{1}{K}\sum_{k=1}^{K} \text{tr } (X - G^kY^k)'M_k(X - G^kY^k) \qquad (11)$$

te minimaliseren voor Y^k. De berekening van de optimale $\hat{Y}^k$ waarde

$$\hat{Y}^k = (D^k)^+(G^k)'M_kX \qquad (12)$$

waarbij $D^k = (G^k)'M_kG^k$ en het symbool '+' aangeeft dat de Moore-Penrose inverse gebruikt wordt, is niet eenvoudig omdat de de matriks D^k in het algemeen niet diagonaal is.
Een algoritme waarbij de Y_j afzonderlijk geschat worden ligt daarom meer voor de hand. Bovendien worden de rang-één restrikties en de kegelrestrikties ook in termen van de Y_j gedefinieerd en niet in termen van Y^k.
Ontbindt daarom de som over de getransformeerde variabelen binnen de groepen $\sum_{j\in I_k} G_jY_j$ in de bijdrage van de variable j en de bijdrage van de andere variabelen in de set V_j^k. Het verlies kan dan opgeschreven worden als:

$$\sigma(X,Y) = \frac{1}{K}\sum_{k=1}^{K} \text{tr } (X - V_j^k - G_jY_j)'M_k(X - V_j^k - G_jY_j) \qquad (13)$$

met $V_j^k = \sum_{\substack{\ell\in I_k \\ \ell\neq j}} G_\ell Y_\ell$. Minimaliseren van deze funktie over Y_j met X en de andere Y_ℓ vast geeft optimale kategorie kwantifikaties voor Y_j:

$$\hat{Y}_j = D_j^{-1}G_j'M_k(X - V_j^k) \qquad (14)$$

met $D_j = G_j'M_kG_j$, een diagonale matriks die de marginale frekwenties van de verschillende kategorieën van de variabele j bevat, gekorrigeerd voor ontbrekende gegevens binnen set k.

2.3.7 Optimaliseren van de objektskores.

Binnen elke iteratiestap wordt eerst een optimale basis voor gegeven waarden van de transformaties berekend, daarna worden nieuwe waarden voor de optimale transformaties berekend voor de gegeven basis die berekend is in de eerste stap. Dit algoritme steunt op het principe ven de alternerende kleinste kwadraten (vergelijk o.a. Young 1981). Het afwisselen van deze substappen geeft een steeds kleiner wordende waarde van de verliesfunktie. Om elke iteratiestap een nieuwe optimale basis X te vinden uitgaande van gegeven Y_j moet het verlies

$$\sigma(X,Y) = \frac{1}{K}\sum_{k=1}^{K}(X - U_k)'M_k(X - U_k) \qquad (15)$$

waarbij $U_k = \sum_{j\in I_k} G_j Y_j$, geminimaliseerd worden voor X. Rekening houdend met de konditiês $u'M_*X = 0$ en $X'M_*X = nI$ levert dit vergelijking (16) op.

$$M_*X\Psi = \left(I - \frac{M_*uu'}{u'M_*u}\right)\sum_{k=1}^{K} M_kU_k \qquad (16)$$

met Ψ een symmetriese matriks van Lagrange multipliers en M_* het gemiddelde aantal niet ontbrekende observaties.

$$M_* = \frac{1}{K}\sum_{k=1}^{K} M_k. \qquad (17)$$

De matriks $\left(I - \frac{M_*uu'}{u'M_*u}\right)$ die J genoemd zal worden zet de matriks $W = \sum_{k=1}^{K} M_kU_k$ in afwijking van zijn kolomgemiddelden. En $\hat{W} = JW$.

Met behulp van singuliere waarde dekompositie (SVD) kan vergelijking (16) opgelost worden.

$$M_*^{-\frac{1}{2}}X\Psi = M_*^{-\frac{1}{2}}\hat{W} = F\Lambda L' \qquad (18)$$

met $F'F=I$, $L'L=I$ en Λ diagonaal. Dus

$$\Psi'X'M_*X\Psi = n\Psi^2 = L\Lambda^2L' = \hat{W}'M_*^{-1}\hat{W} \qquad (19)$$

En zo wordt gevonden dat de nieuwe optimale X gelijk is aan

$$X = \sqrt{n}\, M_*^{-1}\hat{W}L\Lambda^{-1}L' \qquad (20)$$

Dus elke iteratiestap wordt een nieuwe basis X gevonden door de matriks $J\frac{1}{K}\Sigma_{k=1}^{K}\Sigma_{j\in I_k} G_jY_j$ te orthogonaliseren.

Substitueer de in de vorige paregraaf gevonden optimale $\hat{Y}^k$ (vergelijking 8) in vergelijking (7) en schrijf $\sigma(X,*)$ voor het minimum van $\sigma(X,Y)$ naar

Y^k op als volgt:

$$\sigma(X,*) = \frac{1}{K}\sum_{k=1}^{K} \text{tr}\ (X - P_k X)'M_k(X - P_k X) \tag{21}$$

met $P_k = M_k G^k((G^k)'M_k G^k)^+(G^k)'M_k$ (22)

Door te differentiëren naar X kan het minimum van $\sigma(X,*)$ bepaald worden. Dit levert op

$$M_* X\Psi = JPX. \tag{23}$$

Waar $P=\Sigma_{k=1}^{K} P_k$ en Ψ een symmetriese matriks van Lagrange multipliers. Omdat Ψ ook geschreven mag worden als

$$\Psi = L\Lambda^2 L' \tag{24}$$

waarbij $L'L=I$ en Λ is diagonaal, kan vergelijking (23) ook geschreven worden als

$$XL\Lambda^2 = M_*^{-1}JPXL \tag{25}$$

zodat duidelijk wordt dat X een rotatie van de eigenvektoren van $M_*^{-1}JP$ is. Na konvergentie zijn dit ook de eigenvektoren van $\tilde{W}M_*^{-1}\tilde{W}$. De gevonden eigenwaarden Λ^2 geven aan hoe goed de fit per dimensie (oplossing) is. Daarom wordt X na konvergentie geroteerd, omdat het dan echt de eigenvektoren zijn.

Verder kunnen we nog laten zien dat het minimale verlies proportioneel is met de p grootste eigenwaarden van de matriks M_*JP. Substitueer hiervoor de minimale X-waarden te berekenen uit vergelijking (25) in vergelijking (21). Dit levert dan het minimale verlies $\sigma(*,*)$ op, hetgeen gelijk is aan

$$\sigma(*,*) = n\ K\ p\ (1 - \frac{1}{p}\sum_{i=1}^{p} \lambda_{ii}). \tag{26}$$

Is er sprake van enkelvoudig geschaalde variabelen, dan wordt het probleem iets ingewikkelder. Door te schrijven

$$G_j Y_j = G_j y_j a_j' = q_j a_j' \tag{27}$$

met $q_j=G_j y_j$ en deze transformaties q_j en gewichten a_j' vervolgens te verzamelen in matriksen Q_k en A_k' kan de verliesfunktie voor enkelvoudige variabelen ook opgeschreven worden als

$$\sigma(X,Q,A) = \frac{1}{K}\sum_{k=1}^{K} \text{tr}\ (X - Q_k A_k)'M_k(X - Q_k A_k) \tag{28}$$

Bepalen van het minimum van deze funktie naar X, Q en A geeft een minimaal verlies, waarbij de eigenwaarden een funktie zijn van de

kwantifikaties.

$$\sigma(*,*,*) = n\ K\ p\ (1 - \frac{1}{p}\sum_{i=1}^{p}\lambda_{ii}(Q)). \qquad (29)$$

2.3.8 De extra stap bij enkelvoudige variabelen.

Als aan variabele j de rang-één restriktie opgelegd is dan moet een additioneel probleem opgelost worden. Bij een optimale meervoudige kwantifikatie $\hat{Y}_j = D_j^{-1}G_j'(X - V_j^k)$ voor variabele j kan het verlies voor deze variabele uit set k ook geschreven worden als

$$\sigma_j(X,Y) = tr(X - V_j^k - G_j\hat{Y}_j)'M_k(X - V_j^k - G_j\hat{Y}_j) + tr(\hat{Y}_j - y_ja_j')'D_j(\hat{Y}_j - y_ja_j'). \qquad (30)$$

met $Y_j = y_ja_j'$. Verder staan de optimale kwantifikaties niet noodzakelijkerwijs in afwijking van het gemiddelde, namelijk $u'D_j\hat{Y}_j = u'G_j'M_k(X - V_j^k) = u'M_k(X - V_j^k)$ hoeft niet nul te zijn. Dit geldt ook voor de enkelvoudige kwantifikaties. $u'D_jy_j = u'D_j\hat{Y}_ja_j = u'M_k(X - V_j^k)a_j$ is ook ongelijk aan nul in het algemeen. (Als $M_k \neq I$.)
Wel wordt er vereist dat $y_j'D_jy_j = n$. Dus de tweede komponent van de verliesfunktie uit vergelijking (30) is minimaal als $a = \frac{1}{n}\hat{Y}_j'D_jy_j$.
Een optimale enkelvoudige kwantifikatie $\hat{y}_j$ wordt vervolgens verkregen door $\hat{y}_j = \hat{Y}_ja_j$ te berekenen. Dit is voldoende voor enkelvoudig nominale variabelen. Voor enkelvoudige ordinale variabelen wordt eerst een monotone regressie uitgevoerd op $\hat{y}_j$ en daarna de lengte genormaliseerd op n, $\hat{y}_j'D_j\hat{y}_j=n$. En in het enkelvoudig numerieke geval moet de enkelvoudige kwantifikatie voldoen aan een vergelijking van de vorm $y = \alpha h + \beta u$, waar aangenomen wordt dat de ruwe data h genormaliseerd zijn $h'Dh=n$ en $u'Dh=0$. De optimale y ziet er dan als volgt uit

$$\hat{y} = \{(h'D\hat{y})h + (u'D\hat{y})u\}/\{(h'D\hat{y})^2 + (u'D\hat{y})^2\}^{\frac{1}{2}} \qquad (31)$$

Voor de duidelijkheid zijn de indeksen hier weggelaten.

2.3.9 Partitionering van het verlies.

In de vorige paragraaf is een splitsing van de verliesfunktie al ter sprake gekomen. Hier wordt een verdere partitionering van het verlies behandeld. Binnen een set k kan op m_j manieren het verlies van die set

opgesplitst worden. m_j is het aantal variabelen in set k. Gebruik dan $Y = \hat{Y} + (Y - \hat{Y})$ en schrijf het verlies van set k in matriksvorm op als

$$\Sigma_k(X,Y) = (X - V_j^k)'M_k(X - V_j^k)/n - \hat{Y}_j'D_j\hat{Y}_j/n + (\hat{Y}_j - Y_j)'D_j(\hat{Y}_j - Y_j)/n \qquad (32)$$

Dan wordt de diskriminatiematriks van variabele j gedefinieerd als

$$\Delta_M^j = \hat{Y}_j'D_j\hat{Y}_j/n = (X - V_j^k)'P_j(X - V_j^k)/n \qquad (33)$$

met $P_j = M_kG_jD_j^{-1}G_j'M_k$ de projektor van variabele j.
De diagonaal van deze matriks Δ_M^j worden de diskriminatiewaarden genoemd. Als er geen ontbrekende waarden zijn, zijn dit de varianties van de meervoudige kwantifikaties.
Een andere benaming voor deze diskriminatie matriks is meervoudige fit. Voor enkelvoudige variabelen ziet de diskriminatiematriks er uit als

$$\Delta_S^j = \hat{Y}_j'D_j\hat{Y}_j/n = \hat{a}_j\hat{y}_j'D_j\hat{y}_j\hat{a}_j'/n = \hat{a}_j\hat{a}_j' \qquad (34)$$

En deze matriks wordt ook wel de enkelvoudige fit genoemd. Naast deze diskriminatie matriks van een variabele wordt ook een 'totale dipersie' matriks Ξ^j gedefinieerd

$$\Xi^j = (X - V_j^k)'M_k(X - V_j^k)/n \qquad (35)$$

In woorden kan de matriks Ξ^j beschreven worden als de bijdrage van de andere variabelen ℓ aan het verlies van set k (dus $\ell \neq j$).
Dus als de variabele j een meervoudig geschaalde variabele is dan is het verlies van set k

$$\Sigma_k = \Xi^j - \Delta_M^j \qquad (36)$$

op deze manier op te splitsen. Voor enkelvoudige variabelen met restriktie $Y_j = \hat{y}_j\hat{a}_j'$ vinden we dat

$$(\hat{Y}_j - \hat{y}_j\hat{a}_j')'D_j(\hat{Y}_j - \hat{y}_j\hat{a}_j') = \Delta_M^j - \Delta_S^j \qquad (37)$$

dus

$$\Sigma_k = \Xi^j - \Delta_M^j + (\Delta_M^j - \Delta_S^j) \qquad (38)$$

De komponent $(\Delta_M^j - \Delta_S^j)$ wordt het 'enkelvoudige verlies' Σ_S^j van variabele j genoemd en $(\Xi^j - \Delta_M^j)$ wordt 'meervoudig verlies' Σ_M^j genoemd.
dus

$$\Sigma_k = \Sigma_M^j + \Sigma_S^j \qquad (39)$$

Voor meervoudige variabelen bestaat dit totale verlies alleen uit het meervoudige verlies Σ_M^j.

Tot slot nog een vergelijking met homogeniteitsanalyse (HOMALS) en principale komponenten analyse (PRINCALS). Als set k slechts één variabele bevat, dus $V_j^k = 0$, dan is het optellen over Δ_M^j en Δ_S^j even zinvol als het optellen van Σ_M^j en Σ_S^j. Zijn er bovendien geen ontbrekende gegevens in de analyse, dus geldt $M_* = \frac{1}{K}\Sigma_{k=1}^K M_k = I$, dan wordt gevonden $\Xi^j = I$. In dit 'één variabele per set' geval is het gemiddelde van de sommatie van de diagonaalwaarden van de diskriminatie matriksen over alle variabelen gelijk aan de totale fit van de analyse. Ook per dimensie kan een gemiddelde berekend worden, wat de eigenwaarde van die dimensie oplevert.
Voor meerdere variabelen per set is alleen het optellen van de meervoudige en enkelvoudige verliezen zinvol. In het algemeen geldt voor iedere analyse dat de som over alle verliezen van de afzonderlijke sets ($\Sigma_{k=1}^K \sigma_k$) proportioneel is met het totale verlies van de gehele analyse.

2.3.10 Gewichten en komponentladingen.

In het voorgaande is al ter sprake gekomen hoe de gewichten voor enkelvoudige variabelen bepaald kunnen worden. Zij zijn het resultaat van de regressie die voor deze variabelen uitgevoerd wordt. De beste schatting van gewicht a voor variabele j wordt gegeven door

$$\hat{a}_j = (X - V_j^k)'G_j\hat{y}_j/n \qquad (40)$$

De $\hat{a}_j$ zijn dus geen korrelaties van de gekwantificeerde variabelen met de kanoniese variabelen. Dit zou wel het geval zijn als j de enige variabele behorend tot set k is.
Daarom worden de korrelatie met de kanoniese variabelen apart berekend. Voor meervoudige variabelen zijn de korrelaties van de gekwantificeerde variabelen G_jY_j met de kanoniese variabelen X

$$\begin{aligned}\Gamma^j &= \operatorname{diag}(X'X)^{-\frac{1}{2}}X'G_j\hat{Y}_j\operatorname{diag}(\hat{Y}_j'G_j'G_j\hat{Y}_j)^{-\frac{1}{2}} \\ &= X'G_j\hat{Y}_j\Omega_j^{-\frac{1}{2}}/n = X'P_j(X - V_j^k)\Omega_j^{-\frac{1}{2}}/n \qquad (41)\end{aligned}$$

met $X'X=nI$, $\Omega_j=\operatorname{diag}(\Delta^j)/n$ en $P_j=G_jD_j^{-1}G_j'$. Δ^j is de diskriminatiematriks van variabele j. Voor variabelen waar de rang-één restriktie aan opge-

legd werd $Y_j=y_j a_j'$, met diskriminatiematriks $\Delta^j=\hat{a}_j\hat{a}_j'$, ziet de komponentladingenmatriks Γ^j (p x p) er als volgt uit

$$\Gamma^j = X'G_j\hat{y}_j\hat{a}_j'\Omega^{-\frac{1}{2}}/n \qquad (42)$$

Nu kunnen voor enkelvoudige variabelen de $b_j = X'G_j\hat{y}_j/n$ de komponentladingen genoemd worden. De komponentladingenmatriks $\Gamma^j = b_j\hat{a}_j'\Omega_j^{-\frac{1}{2}} = b_j s_j'$, met s_j = teken($\hat{a}_j$) verschilt per kolom dus alleen van teken.
Merk verder op dat als $P_j V_j^k=0$, dat dan $\hat{a}_j=b_j$.
Als er ontbrekende gegevens zijn dan zijn de komponentladingen geen echte korrelaties meer met de kanoniese variabelen, want de X skores worden niet gestandaardiseerd per set. Dit kan leiden tot komponentladingen,die groter zijn dan 1.000.
Voor meervoudige variabelen kunnen ook gewichten berekend worden, maar dezen zijn niet uniek gedefinieerd. Zie voor enkele mogelijke keuzen; de Leeuw (1983).

Hoofdstuk 3 : Voorbeelden

3.1 Beoordeling van verkeerssituaties.

In het komende voorbeeld worden PRINCALS, CANALS en OVERALS analyses naast elkaar gelegd om zo de voordelen die een OVERALS analyse kan hebben te kunnen laten zien. Dit voorbeeld betreft konfliktsituaties tussen weggebruikers op drie kruispunten in de middelgrote stad Malmö in het zuiden van Zweden. Op deze kruispunten hebben twaalf beoordelaarsteams uit acht verschillende landen een aantal uren per dag de verkeerssituaties beoordeeld op "de ernst van een konfliktsituatie tussen twee weggebruikers". Ieder land hanteerde daarbij een eigen systeem van konfliktbeoordeling. Omdat er nogal wat variatie te konstateren viel tussen een definitie van 'konflikt' en in gehanteerde procedures bij de beoordeling van een konfliktsituatie was een vergelijking van de konfliktbeoordelingstechnieken nodig. Daarom hebben alle twaalf teams dezelfde 973 verkeerssituaties beoordeeld. In de volgende analyses worden alleen de meer ernstige konfliktsituaties bekeken, namelijk die situaties waar minstens vier teams een ernstig konflikt meenden waar te nemen en enkele konflikten, die door minstens één team als zeer ernstig werden beoordeeld. Twee konflikten werden niet geanalyseerd omdat dit 'echte' ongelukken waren. Een uitgebreide beschrijving en rapportering van deze data is terug te vinden in "The Malmö Study" (SWOV, 1984) en bij Oppe (1983).
Zodoende bleven er 116 verkeerssituaties over, welke beoordeeld zijn door achtereenvolgens Oostenrijk (Oos), Canada (Can), Duitsland (Dui), Frankrijk team 1 (Fr1), Frankrijk team 2 (Fr2), Engeland (Eng), Zweden team 1 t/m 4 (Zw1, Zw2, Zw3, Zw4), Finland (Fin) en het U.S.A. team (USA). Al deze teams hebben als eerste skoringskategorie de "niet geobserveerd konflikt" kategorie, welke inhoudt dat een verkeerssituatie geen konflikt genoemd wordt. De kategorieën 2, 3 en hoger wijzen op een als steeds ernstiger beoordeeld konflikt. Ook zijn er enkele kategorieën met zeer lage marginale frekwenties in een nabij liggende klasse opgenomen, omdat dezen in een ook uitgevoerde drie-dimensionale analyse tot instabiele resultaten leidden. Dit betrof één 5-skore van Frankrijk1, welke na rekodering tot klasse 4 behoorde. Verder werden een 4-skore voor Duitsland en een voor de USA beide 3-skores. (Vergelijk Oppe, 1983)

Naast deze beoordelingen van elke verkeerssituatie, zijn deze situaties ook op videobanden vastgelegd. Van deze video opnamen zijn later objektieve variabelen afgeleid. Deze afgeleide variabelen betreffende het gedrag van de weggebruikers zijn : Type konflikt (TY), Manoeuvre type 4 (M4), snelheid weggebruiker 1 (V1), snelheid weggebruiker 2 (V2), versnelling weggebruiker 1 (A1), versnelling weggebruiker 2 (A2), de afstand tussen de weggebruikers (MD), de tijd die nodig is om tot een botsing te komen als geen van beide weggebruikers ingrijpt (TTC), de afstand op het meest kritieke punt (DTTC), en tot slot de tijd om het punt te bereiken waar de andere weggebruiker is binnengedrongen (PET). Tussen de haakjes staat de afkorting aangegeven, die men herhaaldelijk in de tekst en vooral in de figuren zal aantreffen. Voor een uitgebreidere beschrijving van wat de variabelen precies betekenen en een indeling van hun kategorieën kan men in bijlage 1 terecht.

Ook zijn het tweede team van Frankrijk en het tweede, derde en vierde team van Zweden niet in de komende analyses opgenomen , omdat direkt al bleek dat deze teams op respektievelijk precies dezelfde manier als Frankrijk 1 en Zweden 1 de beoordelingen maakten en daarom de oplossing volledig domineerden.

Daarom volgen nu de resultaten van een PRINCALS analyse op de overgebleven acht teams. De teams zijn één-dimensionaal en enkelvoudig nominaal geanalyseerd. Dit gaf de volgende resultaten :

TABEL 1 : Resultaten één-dimensionale PRINCALS; 8 teams.

Teams	komponentladingen	volgorde kategorieën na herschaling			
Oostenrijk	.570	3	1	2	
Canada	-.567	2	3	1	4
Duitsland	.579	3	1	2	
Frankrijk1	-.627	1	2	3	4
Engeland	-.719	2	3	1	4
Zweden1	-.502	1	2	3	4
Finland	-.730	2	1	3	4
USA	.650	3	1	2	

Deze analyse lijkt te wijzen op een gemeenschappelijke ernstigheidsdimensie voor alle teams, waar alle teams een redelijk hoge bijdrage aan hebben. Als alle kategorieën van laag naar hoog gerangschikt worden dan is duidelijk dat alle teams in dezelfde richting skoren. Alleen de "niet geobserveerd konflikt" kategorie (1), houdt zich niet mooi aan de volgorde. Het lijkt er dus op dat de variatie tussen de teams erin bestaat dat het moeilijk is om van een verkeerssituatie te konstateren dat er sprake is van een konflikt, maar als een situatie eenmaal als konflikt wordt aangemerkt er grote overeenstemming tussen de teams bestaat over de mate van de ernst van het konflikt.

Een analyse in drie dimensies gaf geen aanleiding een meer gekompliceerde ernstigheidskoring van de teams te veronderstellen. De drie principale komponenten van deze analyse verklaarden respektievelijk 32,5%, 18,3% en 16,5% van de totale variantie, tegenover 38,8% verklaarde variantie van de één-dimensionale analyse. De eerste principale komponent is duidelijk het meest belangrijk.

Om uit te vinden of de variatie in skores, die niet door de gemeenschappelijke dimensie verklaard kunnen worden systematies of random is, werden deze skores vergeleken met de objektieve variabelen, afgeleid uit de videoopnamen. Omdat voor deze vergelijking CANALS werd gebruikt, om zo type konflikt en manoeuvre type op enkelvoudig nominaal nivo en de overige objektieve variabelen op enkelvoudig ordinaal nivo te kunnen analyseren, was het nodig de PRINCALS skores in kategorieën in te delen. Zodoende zijn deze objektskores tot 6 klassen van skores teruggebracht. Deze nieuwe variabele (OB8) werd daarna in een CANALS analyse aan de achtergrondsvariabelen van het konflikt gerelateerd. Twee nadelen hiervan zijn dat we ten eerste te maken hebben met een erg beperkte klassifikatie van de konflikten en ten tweede dat we alleen de gemeenschappelijkheid van de 8 beoordelaarsteams relateren aan de objektieve variabelen en niet voor elk team apart een eigen bijdrage t.o.v. de objektieve variabelen kunnen bekijken.

Met het doen van een OVERALS analyse worden deze beide nadelen opgeheven. Daarom wordt nu eerst kort een overzicht van de CANALS resultaten gegeven en wordt daarna de OVERALS analyse wat uitgebreider behandeld.

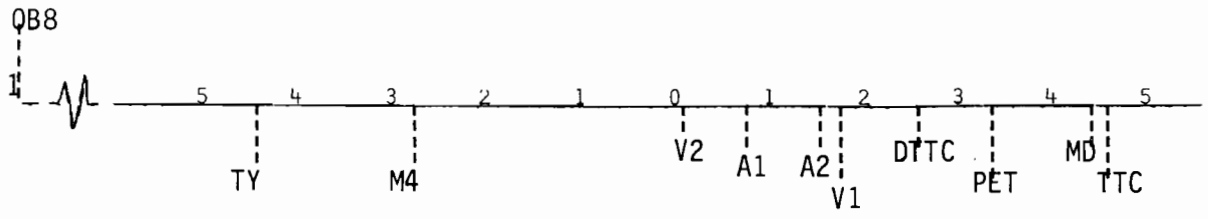
Het bleek voor 14 konflikten niet mogelijk te zijn een TTC skore, en daarmee ook een DTTC waarde te berekenen, omdat de weggebruikers in deze situaties niet in konflikterende richtingen reden. Bovendien waren er 43 ver-

keerssituaties, die geen PET-waarde opleverden, daar de betrokken weggebruikers volledig stopten. Deze niet beschikbare waarden zijn in de CANALS analyse als extra kategorieën van de betreffende variabelen behandeld.
De resultaten van deze analyse zijn te vinden in tabel 2.

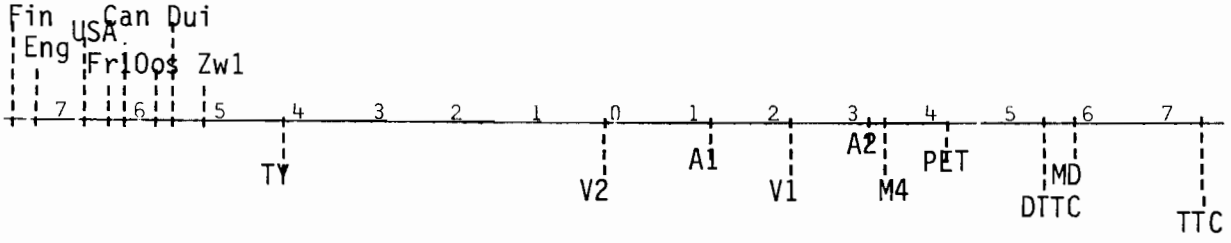
TABEL 2 : CANALS; korrelaties met de kanoniese variaat van de tweede set.

Variabelen		komponentladingen	volgorde kategorieën
Konflikt type	TY	-.428	1 3 4 2 5
Manoeuvre type	M4	-.295	6 3 5 7 9 1 2
Snelheid weggebr.1	V1	.181	1 2 (3 4) 5
Snelheid weggebr.2	V2	.004	1 (2 3) (4 5)
Versnelling weggebr.1	A1	.091	1 2 3 4 5 6
Versnelling weggebr.2	A2	.157	(1 2) (3 4) (5 6)
Minimale afstand	MD	.440	1 (2 3 4) 5 6
Time To Collision	TTC	.462	1 2 (3 4)
Distance at TTC	DTTC	.255	(1 2) 3 4 5 6
Post Encroachment Time	PET	.332	1 (2 3 4 5)
PRINCALS skore	OB8	-1.000	1 2 3 4 (5 6)

Deze resultaten zijn tevens overzichtelijk in figuur 1 in beeld gebracht. Alvorens deze resultaten te bespreken worden eerst de resultaten van de OVERALS analyse bekeken. Deze OVERALS analyse kan min of meer opgevat worden als een kombinatie van de hier besproken PRINCALS en CANALS analyse. Alle teams zijn namelijk in acht aparte sets op enkelvoudig nominaal nivo geanalyseerd met als negende set de tien achtergrondsvariabelen met dezelfde meetnivo's als in de CANALS analyse. Overigens is de behandeling van de ontbrekende gegevens in de OVERALS analyse anders. Namelijk per set wordt er gekeken naar welke objekten binnen die set een ontbrekend gegeven hebben, en deze objekten worden dan vervolgens niet in de berekening betrokken. Daarom is de berekening van de kwantifikaties in de negende set op slechts 59 objekten gebaseerd. Bovendien berekent OVERALS een gemiddelde kanoniese ruimte en wordt er geen ruimte voor elke set berekend, zoals dit in CANALS het geval is.
De OVERALS resultaten zijn terug te vinden in tabel 3 en figuur 2.



Figuur 1 : CANALS resultaten.



Figuur 2 : OVERALS resultaten gekorrigeerd voor richting.

Tabel 3 : OVERALS, komponentladingen van de variabelen

Variabelen	komponentladingen	volgorde kategorieën
Oostenrijk	-.595	2 1 3
Canada	-.615	2 3 1 4
Duitsland	-.574	2 1 3
Frankrijk1	-.635	1 2 3 4
Engeland	.717	4 1 3 2
Zweden1	.529	4 3 2 1
Finland	.741	4 3 1 2
USA	.672	3 1 2
Type konflikt	-.402	3 1 6 4 8 2 5 7
Manoeuvre 4	.330	2 1 7 5 9 6 3
Snelheid 1	.214	(1 2 3) 4 5
Snelheid 2	-.028	1 2 (3 4) 5
Versnelling 1	.136	1 2 3 4 5 6
Versnelling 2	.304	(1 2) (3 4 5 6)
Min. Distance	.598	1 (2 3) (4 5) 6
TTC	.752	1 2 3
DTTC	.540	1 (2 3) (4 5)
PET	.411	1 2 (3 4)

Als punten niet mooi in de goede volgorde geschaald kunnen worden is dit in tabel 2 & 3 aangegeven met haakjes om de betreffende kategorieën, die op één punt worden afgebeeld. Bij veel van de objektieve variabelen kan dit gekonstateerd worden. Toch wordt dit ordinale nivo gehandhaafd, daar de inhoud van de variabelen dit rechtvaardigt.
Alvorens nu aandacht wordt besteed aan de interpretatie van de analyse-resultaten, worden eerst de overeenkomsten van de OVERALS en de CANALS analyse bekeken. Figuur 1 en 2.

Links in de figuren kunnen de beoordelingen van de teams teruggevonden worden, in het ene geval gerepresenteerd door bijdragen van alle teams apart, en in het andere geval door de variabele OB8, welke de gemeenschappelijkheid in de konfliktbeoordeling van de acht teams verkregen door de eerder besproken één-dimensionale PRINCALS analyse representeert.
Voor alle variabelen in beide figuren geldt dat hun afbeelding wijst in de richting van de hoogste kategorie. Dat houdt dus in dat de meer ernstig beoordeelde verkeerssituaties links in de figuren afgebeeld kunnen worden, terwijl de minder ernstige konflikten aan de rechterkant van de dimensie terug te vinden moeten zijn. Ook hier moet opgemerkt worden evenals bij de PRINCALS analyse, dat de kategorie "niet geobserveerd konflikt" niet voor alle teams op dezelfde manier op deze dimensie te plaatsen is. Deze kategorie gedraagt zich exact hetzelfde als al in de PRINCALS analyse naar voren kwam. Hierop hebben de objektieve variabelen kennelijk niet veel invloed.
Ook de objektieve variabelen wijzen in de richting van de hoogste kategorie. Voor de nominale variabelen is het nodig de volgorde van gekwantificeerde kategorieën wat beter te bezien. Zo lijkt de variabele M4 op het eerste gezicht omgeklapt naar de andere kant van de dimensie, maar een blik op d volgorde van de kategorieën toont dat ook dezen min of meer dezelfde volgorde hebben behouden als in de CANALS analyse. Voor de variabele TY wordt de vergelijking iets minder omdat hier enkele kategorieën voor de analyse zijn samengenomen. Maar de overige kategorieën representeeen gelijke beeld voor OVERALS en CANALS.
De enige verschuivingen die te konstateren zijn, zijn alleen die van het belang van enkele variabelen in de analyse. Zo worden de variabele A2 en DTTC belangrijker en wordt de PET onbelangrijker in de OVERALS analyse.

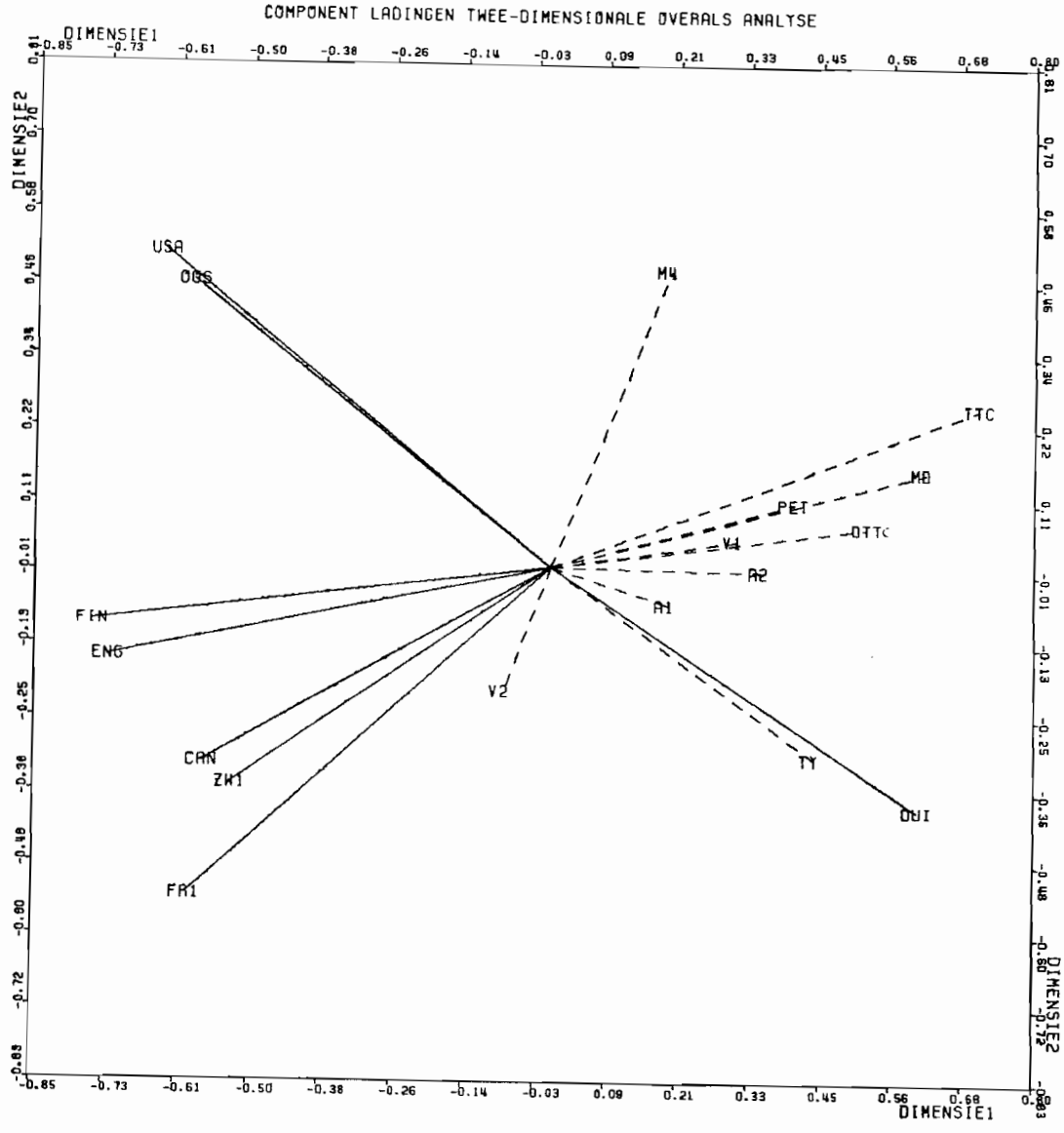
Dit zou te wijten kunnen zijn aan het verschil tussen de behandeling van de ontbrekende gegevens in beide analyses. In ieder geval lijken de resultaten dermate op elkaar, dat voor de interpretatie geen onderscheid gemaakt hoeft te worden. Daarom volgt nu een algemene interpretatie van de gegevens.
Voor de teams in de OVERALS analyse geldt hetzelfde als wat al gezegd is voor de PRINCALS analyse. Dus kan er volstaan worden naar de relatie te kijken , die de objektieve variabelen met de ernst van de beoordeling hebben.
Uit figuur 1 en 2 valt duidelijk op te maken dat de niet ernstig beoordeelde konflikten samen hangen met een hoge skore op de variabele TTC. Dus als er veel tijd is voordat er een botsing zal plaatsvinden, dan wordt het betreffende konflikt als niet-ernstig beoordeeld. Hetzelfde geldt voor de variabele MD. Een grote afstand tussen de weggebruikers hangt samen met een niet ernstig konflikt en omgekeerd hangt een kleine afstand samen met een wel ernstige konfliktsituatie. De DTTC en PET variabele spelen ook een belangrijke rol bij de beoordeling van de mate van ernst door de teams. Een kleine afstand op het meest kritieke punt (DTTC) wordt gezien als een meer ernstig konflikt en een kleine PET-waarde (dus de tijd om het punt te bereiken waar de andere weggebruiker het parkoers van degene die voorrang heeft is binnengedrongen) relateert ook aan de ernst van de situatie. Voor de variabelen TY en M4 moet eerst naar de kategorie kwantifikaties gekeken worden. M.b.v. de volgorde van de kategorie kwantifikaties uit tabel 2 en 3 kan afgeleid worden dat voor TY de kategorieën 5 (voetganger - fiets) en 2 (auto - voetganger) samenhangen met ernstige konfliktsituaties en de kategorieën 3 (auto - fiets) en 1 (auto - auto) weinig ernstig beoordeelde situaties zijn. Voor de OVERALS analyse hebben we ook nog de ernstig beoordeelde kategorie 7 (fiets - fiets), die in de CANALS analyse met kategorie 5 werd samengenomen. Op dezelfde manier kan zo ook naar variabele M4 gekeken worden. Hiervan hangen de kategorieën 2 en 1 (zie bijlage 1 voor de inhoud van deze kategorieën) samen met ernst van de situatie.
Verder duidt hard remmen van weggebruiker 2 (A2) op een niet-ernstig konflikt. In mindere mate geldt dit voor weggebruiker 1 (A1). En tot slot heeft V1 nog een kleine bijdrage. Namelijk een lage initiële snelheid van weggebruiker 1 hangt samen met ernst. Voor V2 kan eigenlijk niet meer van een bijdrage gesproken worden.

Het kan natuurlijk zijn dat een wat gespecificeerdere relatie van de teams t.o.v. de objektieve variabelen gewenst is. Hieruit vloeit opnieuw een voordeel van het doen van een OVERALS analyse voort, namelijk dat de gegevens simpelweg in twee of meer i.p.v. in één dimensie geanalyseerd kunnen worden. De resultaten van zo'n twee-dimensionale analyse zijn te vinden in tabel 4.

TABEL 4 : OVERALS; resultaten twee-dimensionale analyse.

Variabelen	komponentladingen		volgorde kategorieën
Oos	-.592	.464	2 1 3
Can	-.574	-.316	1 2 3 4
Dui	.607	-.392	3 1 2
Frl	-.600	-.533	1 2 3 4
Eng	-.728	-.147	2 3 1 4
Zwl	-.527	-.351	1 2 3 4
Fin	-.755	-.090	1 2 3 4
USA	-.639	.512	2 1 3
TY	.436	-.309	7 2 8 5 6 4 3 1
M4	.197	.483	1 2 7 5 3 6 9
V1	.307	.045	1 2 (3 4) 5
V2	-.077	-.202	(1 2) 3 (4 5)
A1	.196	-.061	(1 2 3 4) 5 6
A2	.351	-.005	1 2 (3 4) (5 6)
MD	.620	.157	1 (2 3) 4 5 6
TTC	.705	.259	1 2 3
DTTC	.521	.067	1 (2 3) 4 5
PET	.399	.105	1 2 (3 4)

De eigenwaarden van deze analyse .392 en .198 voor respektievelijk de eerste en de tweede dimensie. Er is dus een stijging van de fit waar te nemen, maar de eerste dimensie draagt meer dan twee maal zoveel bij aan de oplossing. Toch is het leuk deze twee dimensionale resultaten te bekijken, omdat de manier waarop ernst van een konflikt door een bepaald land wordt geskoord, dan wat duidelijker wordt. Kijk hiervoor naar figuur 3.

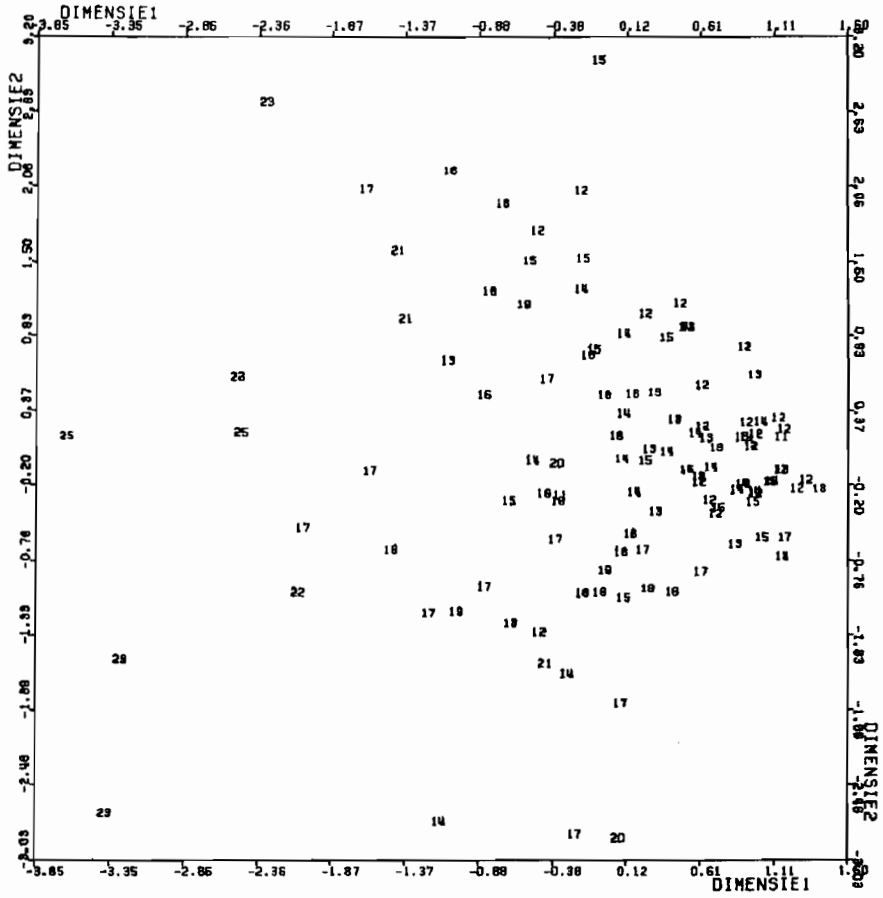


Figuur 3 : Korrelaties van de variabelen met de kanoniese variaten

De belangrijkste relaties tussen de beoordelingen van de teams en de objektieve variabelen, die in deze figuur makkelijk te overzien zijn, zullen besproken worden. Te zien is dat de relatie tussen de teams erg belangrijk blijft, alhoewel er wel een splitsing te konstateren valt.

Het oordeel van Engeland, Finland, Canada Zweden en Frankrijk hangt vooral samen met een lage TTC en MD-waarde en in iets mindere mate met een lage DTTC en PET skore. Voor het franse team wordt ook nog een samenhang met het type manoeuvre (M4) gevonden. De manoeuvres, die tot de eerste en tweede kategorie behoren worden als meer ernstige konflikten beoordeeld. Naast de overeenstemming tussen deze vijf teams wordt er erg grote overeenstemming gevonden tussen de teams van Oostenrijk, Duitsland en de USA. Voor

deze teams lijkt het konflikttype (TY) een belangrijke achtergrondsvariabele voor de beoordeling te zijn. Hiervan worden voornamelijk de situaties die tot de zevende, tweede en achtste kategorie van deze variabele behoren door deze teams als ernstig ervaren.
Verder spelen ook in deze analyse de variabelen V1, V2, A1 en A2 een minder belangrijke rol.
Tot slot nog iets over de objekten uit de analyses. In geen van de analyses zijn echte 'uitbijters' aan het licht gekomen. Dit zijn objekten, die in de kanoniese ruimte een waarde hebben gekregen, die ver van de andere objekten verwijderd ligt. Op deze manier kunnen enkele objekten de oplossing domineren. Voor de laatste analyse zijn deze objektskores te zien in figuur 4. De waarden waarmee deze objektskores gelabeld zijn, zijn verkregen door de skore, waarin elk team de mate van ernst voor ieder konflikt heeft uitgedrukt over alle teams bij elkaar op te tellen. De minst ernstige konflikten verkregen zo een totaalskore 11 en de meest ernstige kwamen op een skore 25.

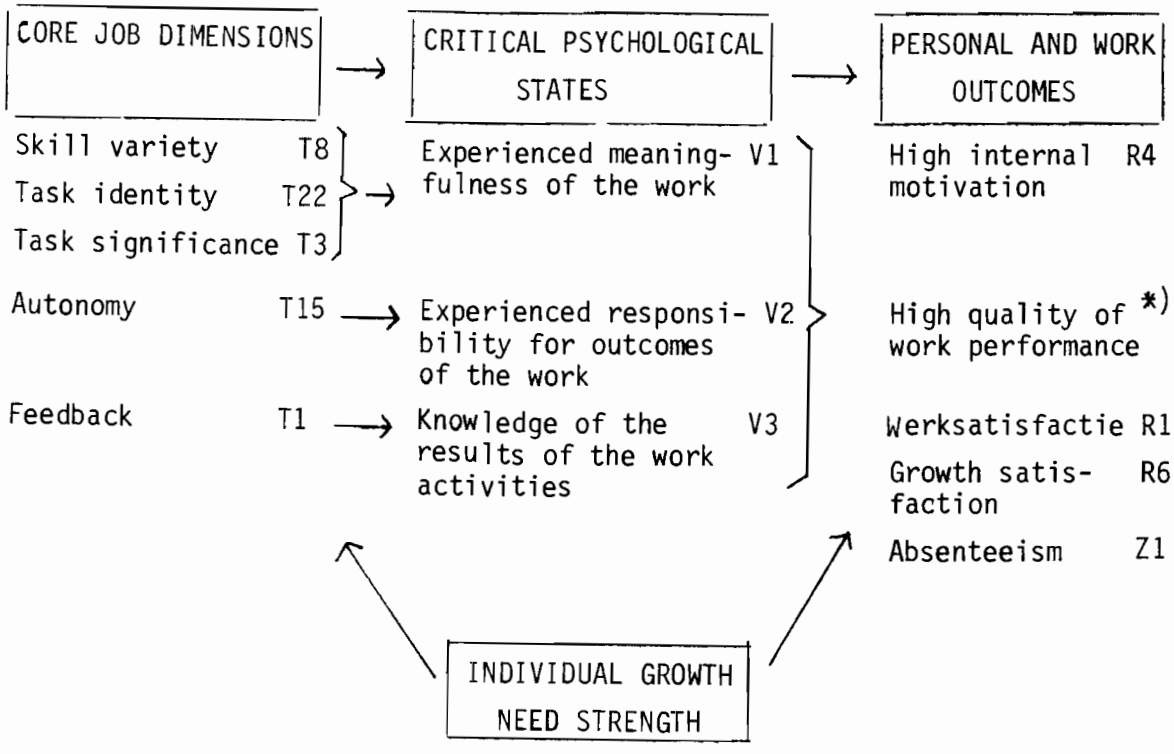


Figuur 4 : Objektskores gelabeld naar somskore van de teams.

Duidelijk wordt dan ook gelijk dat de meest ernstige konflikten meer links liggen, wat al bekend was daar de pijlen, die de komponentladingen van de teams in figuur 3 weerspiegelen ook allemaal, behalve voor het duitse team in de richting van de meest ernstige beoordeling wijzen.

3.2 Taakkenmerken.

Op basis van een literatuuronderzoek heeft Algera (1980) vierentwintig taakkenmerken oftewel werkaspekten geselekteerd, die verondersteld worden effekt te hebben op gedrag en attitude van degenen die de taak uitvoeren. Daarnaast heeft hij achttien afhankelijke variabelen gekozen, die de reakties van de taakuitvoerenden op de kenmerken van hun werk aangeven. Om tussen bovengenoemde variabelen diverse relaties te onderzoeken heeft hij 61 verschillende funkties bij Hoogovens IJmuiden BV onderzocht. Elk van deze funkties is door drie taakuitvoerenden én door drie 'deskundigen' (bijv. chefs of personeelsfunktionarissen) die de funktie voldoende kenden beoordeeld.
Naast deze beoordelingen van de taakkenmerken heeft Algera ook skores verzameld met betrekking tot een aantal "afhankelijke variabelen", die opgevat kunnen worden als een specifikatie van de reakties van de taakuitvoerders op hun werksituatie.
Met behulp van deze gegevens wordt in door Algera onder andere het 'Job Characteristics Model' voor werkmotivatie van Hackman and Oldham (1976) getoetst. De kern van dit 'Job Characteristics Model' (JCM) is dat positieve resultaten, zoals grote motivatie, hoge satisfactie, hoge prestaties en laag verzuim verkregen worden wanneer een drietal 'Critical Psychological States' aanwezig zijn bij de taakuitvoerder. Het model is gebaseerd op de redenering dat het uitvoeren van een taak intrinsiek belonend kan werken en aanzet tot verdere inspanning van een individu, zolang het uitvoeren van de taak de taakuitvoerder een positieve gewaarwording geeft.
De drie 'Critical Psychological States' vormen hierbij een klasse van intermediërende variabelen tussen enerzijds de taakkenmerken en anderzijds de hierboven genoemde resultaten.
Verder stelt de theorie dat de 'Critical Psychological States' veroorzaakt worden door de aanwezigheid van een vijftal taakkenmerken. En omdat niet elk individu op dezelfde wijze zal reageren op een funktie met een hoog 'motivating potential' wordt in het model rekening gehouden met de modererende variabele 'individual growth need strength'. Deze variabele fungeert als moderator tussen de relatie 'job-dimensions' - 'psychological states' en tussen de relatie 'pshychological states' - 'outcomes'.



Figuur 5 : Schematiese weergave van het 'Job Characteristics Model'.

Dit model is schematies weergegeven in figuur 5.
Achter de door Hackman and Oldham gegeven formulering van de modelaspekten, staat de afkorting van de variabele aangegeven waarmee de aspekten in dit onderzoek gemeten zijn.
Een meer volledig overzicht van de inhoud van de variabelen wordt gegeven in bijlage 2. Deze beschrijving komt niet helemaal overeen met Algera. De vragen die betrekking hebben op de schaling van taakkenmerken, zijn in Algera gemeten m.b.v. een grafiese schaal. Op deze schaal, gepresenteerd als vertikale lijn met algemeen bekende beroepen of werksituaties als schaalankers, moest de taakuitvoerder aangeven waar zijn/haar beroep thuishoorde. Later is in milimeters langs deze lijn opgemeten (met een range van -100 mm tot +100 mm) welke skore het taakkenmerk van een taak-uitvoerder had gekregen.
Voor de OVERALS analyse, welke uitgaat van kategoriese gegevens zijn de skores tot vijf klassen teruggebracht. De skores -100 tot -60 behoren

*) De variabele 'high quality of work' is om bepaalde redenen niet in het onderzoek van Algera gemeten.

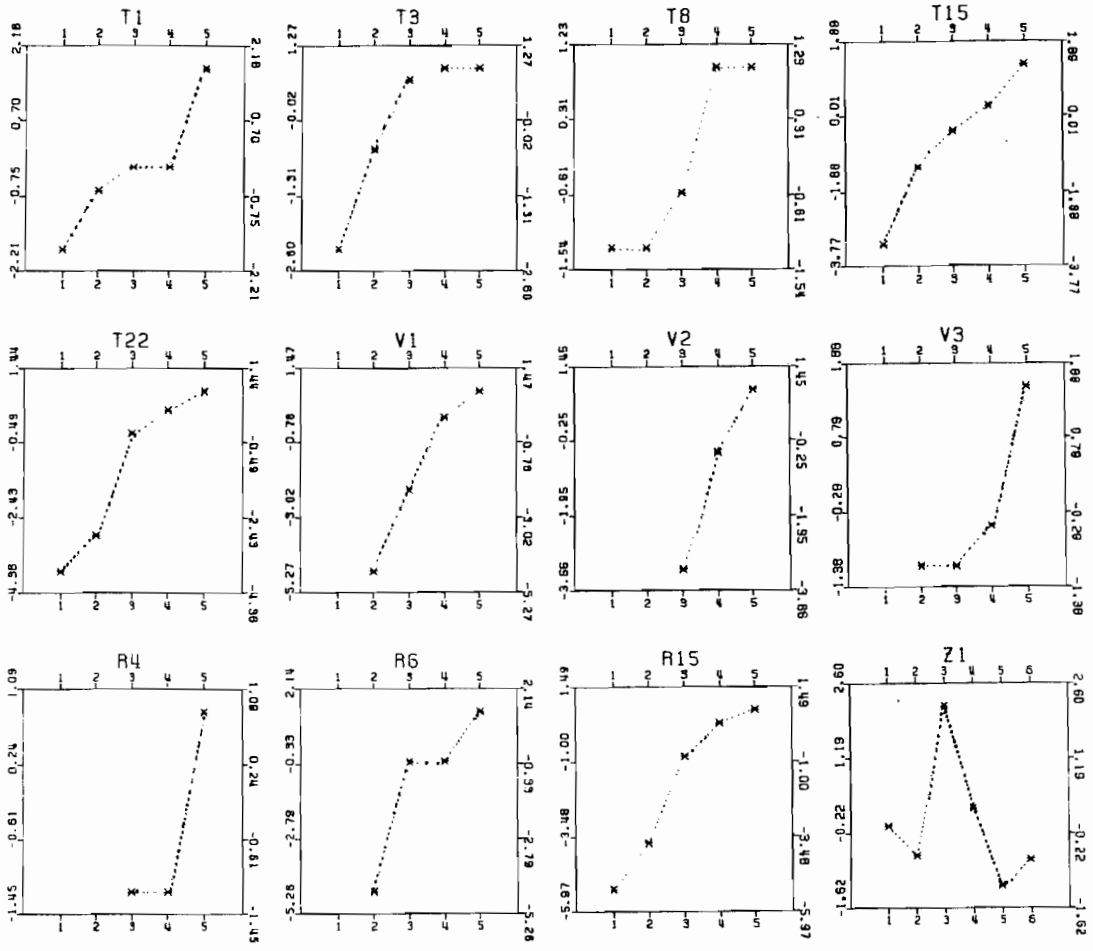
tot klasse 1, -60 tot -20 tot klasse 2, -20 tot +20 tot klasse 3 en zo verder tot klasse 5. De andere variabelen in het model zijn ieder afzonderlijk tot stand gekomen door 4 à 8 vragen te laten beantwoorden op vijfpuntsschalen. Per variabele zijn de betreffende vragen vervolgens opgeteld en opnieuw tot 5 klassen teruggebracht.

In dit voorbeeld zal worden getracht de relaties tussen de variabelen in het JCM m.b.v. een OVERALS analyse zichtbaar te maken, waarbij alleen gebruik gemaakt zal worden van de gegevens die betrekking hebben op de taakuitvoerders. Hiervoor werd gekozen omdat volgens de theoretiese opvattingen van Hackman and Lawler (1971), voorlopers op het JCM, de reakties van de taakuitvoerder eerder bepaald worden door de perceptie van de taak dan door de objektieve taakkenmerken.

Dat voor niet-lineaire meer sets analyse gekozen wordt is aannemelijk omdat hier duidelijk onderscheid gemaakt kan worden tussen drie verschillende groepen variabelen. Verder is een ordinaal verband tussen de kategorieën van de afhankelijke variabelen aannemelijker dan het numerieke verband , welke door andere technieken aangenomen wordt.

Een bezwaar zou kunnen zijn dat OVERALS geen toets verschaft of de empiriese gegevens overeenstemmen met het van te voren gespecificeerde model. Maar het doel van dit voorbeeld is ook niet om de causale relaties die het JCM postuleert te toetsen maar om de kategorieën van de variabelen onder bepaalde restrikties zo te schalen dat dit resulteert in een optimale samenhang tussen de groepen.

De eerste OVERALS analyse ziet er daarom als volgt uit. De eerste set bestaat uit de taakkenmerken T1, T3, T8, T15 en T22. De variabelen V1, V2 en V3 die vorm geven aan de 'critical psychological states' behoren tot de tweede set en de resultaatvariabelen R4, R6, R15 en Z1 tot de derde set. Al deze variabelen kregen in de analyse de enkelvoudig ordinale restriktie opgelegd. Dit betekent dat de kategorieën van de variabele op één lijn moeten liggen (enkelvoudig) en bovendien dat de volgorde van de nieuwe kategorie waarden dezelfde moet zijn als de volgorde van de oorspronkelijke kategorie volgorde (ordinaal). Deze ordinaliteitskeuze bleek na de eerste analyse een erg slechte keus geweest te zijn voor de variabele 'ziekteverzuim'. De eerste twee én de laatste vier kategorieën vielen samen. Daarom is besloten de variabele Z1 enkelvoudig nominaal te analy-



Figuur 6 : Verloop van de enkelvoudige kategoriekwantifikaties

seren. Van deze laatste analyse zijn de schalingen in figuur 6 weergegeven. Op de horizontale as staan de oorspronkelijke kategorienummers en op de vertikale as de kwantifikaties na herschaling. Duidelijk is in deze figuur te zien dat de variabele Z1 (ziekteverzuim) beslist niet aan een monotoon stijgende funktie voldoet. Had deze variabele wel deze ordinaliteitsrestriktie meegekregen dan zouden de meeste kategoriepunten dezelfde schaling hebben gekregen. Dit is bijvoorbeeld wel gebeurd bij kategorie 3 en kategorie 4 van variabele T1.

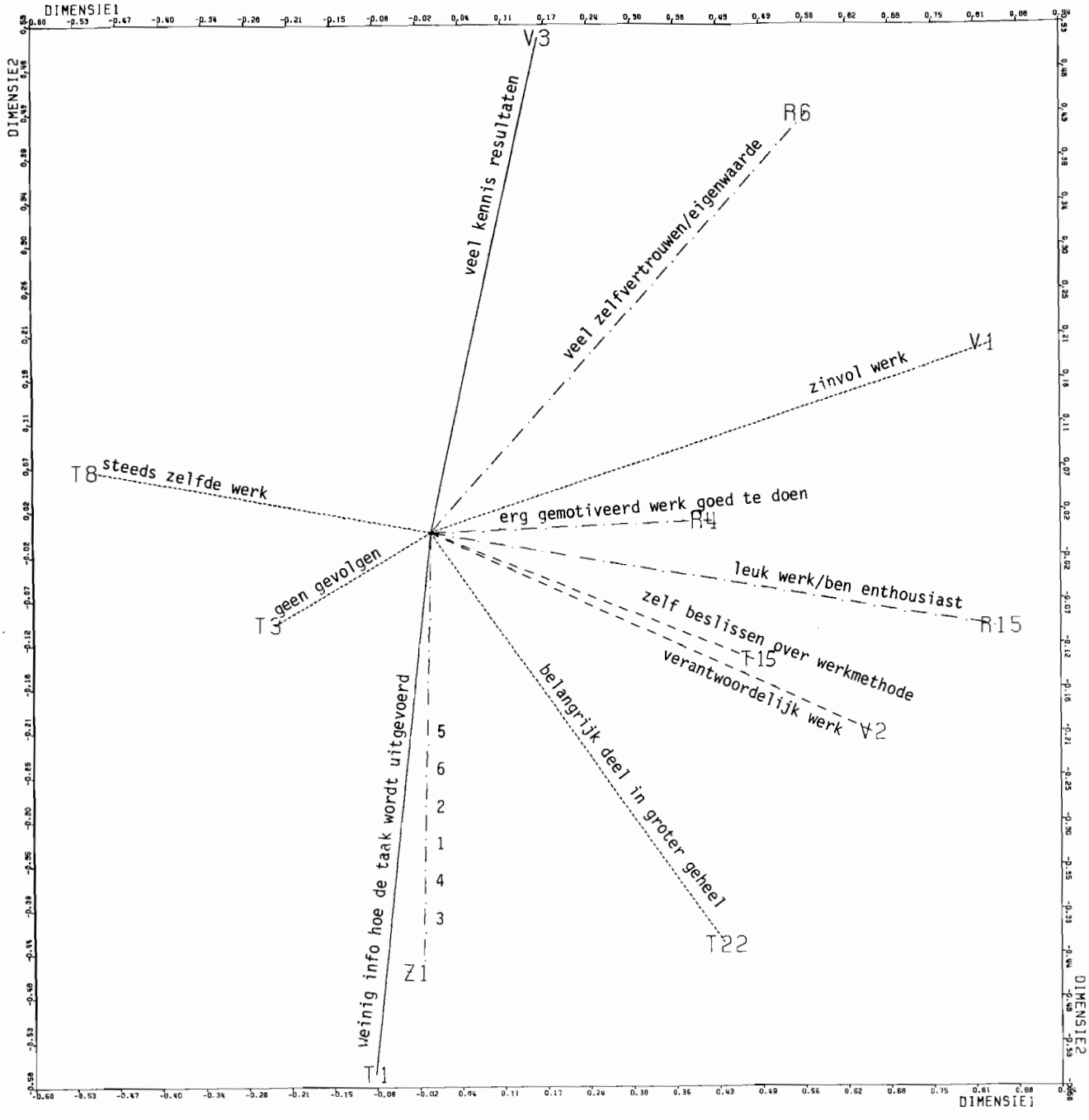
De komponentladingen en de gewichten die het resultaat zijn van deze OVERALS analyse in twee dimensies worden gegeven in tabel 5. De gewichten zijn rechtstreeks het resultaat van de 6 meervoudige regressies die het programma uitvoert, namelijk binnen elk van de drie groepen worden er twee kanoniese variabelen berekend.

TABEL 5 : Gewichten en komponentladingen bij een twee-dimensionale OVERALS analyse van het JCM.

Variabelen	Gewichten		Komponentladingen	
T1	-.216	-.593	-.086	-.562
T3	-.148	-.204	-.233	-.096
T8	-.382	-.100	-.506	.063
T15	.311	.023	.480	-.133
T22	.285	-.487	.437	-.429
V1	.685	.295	.836	.194
V2	.389	-.498	.658	-.209
V3	-.070	.597	.163	.513
R4	.218	-.021	.419	.011
R6	.145	.711	.562	.434
R15	.730	-.424	.850	-.099
Z1	-.095	-.497	-.014	-.457

De komponentladingen zijn de korrelaties van de variabelen met de uit de regressies voorspelde variaten.
Deze komponentladingen zijn direkt vergelijkbaar met de kanoniese skores (ook objektskores genoemd). Dit geldt niet voor de gewichten. De grootte van de gewichten is namelijk ook afhankelijk van de bijdrage van de andere variabelen binnen de set. Bij de interpretatie van de analyse resultaten wordt daarom in eerste instantie naar de korrelaties van de variabelen met de kanoniese variaten gekeken. Deze korrelaties kunnen als punten in de kanoniese ruimte (dit is de ruimte opgespannen door de kanoniese variaten) getekend worden. En omdat we te maken hebben met enkelvoudig gerestrikteerde variabelen, wat betekent dat de kategorieën geprojekteerd zijn op een lijn door de oorsprong, kunnen we deze punten met de oorsprong verbinden waardoor een vektordiagram ontstaat. Omdat in dit twee-dimensionale voorbeeld de kanoniese ruimte een plat vlak is, is in figuur 7 op eenvoudige wijze te zien op welke manier de variabelen aan de kanoniese variaten én aan elkaar gerelateerd zijn. De lengte van de vektor geeft tevens aan hoe belangrijk een variabele is. Voor de enkelvoudig ordinaal geanalyseerde variabelen geldt dat de vektor in de richting van de hoogste kategorie wijst. Wat deze hoogste kategorie voor elke variabele inhoudt is er voor iedere variabele in de figuur bijgeschreven. Dit mooie volgorde verloop geldt niet voor de enkelvoudig nominale variabele Z1. Voor deze

variabele staat aangegeven in welke volgorde de kategoriekwantifikaties op de lijn door de oorprong en het variabelepunt in de ruimte liggen. Voor een interpretatie van de OVERALS analyse kijken we eerst naar de fit. De fit bedraagt 1.2429, wat betrekkelijk hoog is omdat de fit een maksimale waarde heeft die gelijk is aan het aantal dimensies. In dit geval van twee dimensies had de fit een waarde 2.000 kunnen hebben als de data perfekt door twee dimensies gerepresenteerd hadden kunnen worden.



Figuur 7 : Komponent ladingen JCM-model in twee dimensies

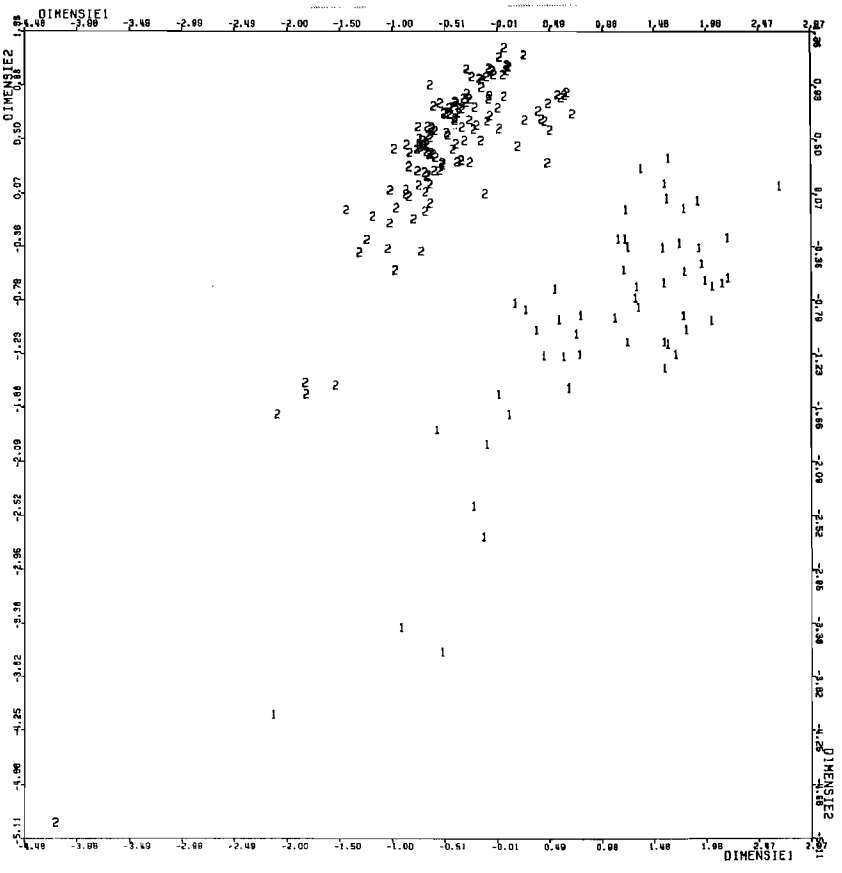
Deze twee-dimensionale analyse verklaart dus 62,1% van de totale variantie. Hiervan komt 35,5% voor rekening van de eerste dimensie (eigenwaarde .711) en 26,6% voor de tweede dimensie (eigenwaarde .532).
In figuur 5 is met verschillende stippellijnen aangegeven welke variabelen volgens het Job Characteristics model met elkaar samenhangen. In vertikale richting is duidelijk samenhang te konstateren tussen V3 (kennis resultaten) en T1 (informatie m.b.t. taakuitvoering). Dit laat zien dat het weten van jezelf of je het werk goed doet of juist niet,sterk samenhangt met het krijgen van informatie over de geleverde prestaties en het effekt van handelingen. Deze beide variabelen zijn ook aan de variabele "ziekteverzuim" (Z1) gerelateerd. Werksituaties waar beschikking is over veel informatie betreffende de geleverde prestaties en veel feedback, die hangen samen met een langdurig/veel ziekteverzuim. Een gemiddeld aantal verzuimdagen (1 tot 3 weken) wordt bij taken gevonden waar weinig informatie en weinig resultatenkennis ter beschikking staat van de taakuitvoerder. En mensen die erg weinig verzuimd hebben (0 tot 4 dagen) liggen dichtbij de oorsprong . Dit betekent dat veel/weinig informatie of kennis m.b.t. de resultaten hier niet van belang is.
Werk waar het maken van fouten niet tot ernstige gevolgen leidt (T3) gaat samen met het vinden dat men nutteloos werk verricht (V1). Natuurlijk geldt omgekeerd ook dat zinvol werk samengaat met fouten, die tot ernstige gevolgen leiden. Het hebben van zinvol werk staat ook in verband met de mate waarin werk een gevoel van zelfvertrouwen en eigenwaarde (R6) geeft. Bovendien vertoont deze variabele R6 verband met V3, namelijk dat veel zelfvertrouwen samengaat met een job waarvan men vindt dat veel kennis over de resulatataten zichtbaar zijn en ook voornamelijk met T3 (werk dat veel zelfvertrouwen/eigenwaarde geeft is tegelijkertijd werk waar fouten tot ernstige gevolgen leiden).
Zinvol werk (V1) hangt ook samen met 'veel motivatie om het werk goed te doen (R4) en met een hoge werksatisfaktie (R15). R15 heeft ook verband met V15, dus de mate waarin de taakuitvoerder zelf kan beslissen over zijn methode van werken, en met T8 (het niet steeds terugkeren van dezelfde werkzaamheden). Deze laatste twee kenmerken tonen vooral samenhang met elkaar en ook nog met 'vinden dat de eigen verantwoordelijkheid t.o.v. het werk hoog is' (V2).
Ten slotte nog T22. Hiervan is in de figuur te zien dat een job die een

belangrijk onderdeel van een groter geheel vertegenwoordigd voornamelijk verantwoordelijk werk is (V2), dat men zelf kan beslissen aangaande de methode van werken (T15) en dat men niet vaak te maken heeft met steeds dezelfde werkzaamheden (T8).
Deze relaties komen redelijk overeen met de in het JCM-model aangegeven relaties. Alleen voor T8 en T22 is er volgens de analyse eerder samenhang met V2, dan dat er samnehang is met V1, zoals het JCM-model postuleert.
Naast deze analyse is er een tweede analyse uitgevoerd. In het model van Hackman and Oldham is er namelijk ook nog sprake van een modererende variabele. Deze variabele zou de relatie tussen de vijf taakkenmerken en de 'outcome'-variabelen beïnvloeden. Hackman and Oldham toetsen deze veronderstelling door korrelaties tussen 'job dimensions' en 'critical psychological states' en tussen 'critical psychological states' en 'outcome'- variabelen te berekenen apart voor twee subgroepen respondenten, namelijk de respondenten die behoren tot de bovenste en de onderste 25% van de skoreverdeling op de variabele 'growth need strength'. De voorspelling dat de korrelaties voor de subgroep van respondenten in de bovenste 25% van de skoreverdeling hoger zijn dan voor de respondenten in de onderste 25% wordt bevestigd. Verder merken Hackman and Oldham op dat een (zeer) lage skore op de modererende variabele niet betekent dat individuen met dergelijke skores negatief zouden reageren op een komplekse funktie.
In de studie van Algera is er ook de beschikking over een variabele die inhoudelijk overeenstemt met de variabele 'growth need strength'. Dit is de variabele M18 welke 'preferentie voor autonomie' meet.
Om dit modererende effekt met behulp van OVERALS te laten zien is gekozen om de onafhankelijke variabelen interaktief te maken. Dit is niet de meest aangewezen manier, maar het geeft wel een mooie illustratie van de techniek. Dit interaktief maken van de variabelen wordt gedaan door bij de vijf taakkenmerken en de drie 'psychological states' binnen de kategorieën een tweedeling te maken tussen personen die weinig behoefte aan autonomie hebben en personen die hoge autonomie prefereren. Zodoende hebben deze variabelen nu 10 kategorieën i.p.v. de oorspronkelijke 5. In volgend schema is te zien hoe de nieuwe kategorieën tot stand zijn gekomen.

		autonomie	
		laag	hoog
oude kategorie nummers	1	1	2
	2	3	4
	3	5	6
	4	7	8
	5	9	10

In het kader staan de nieuwe kategorienummers.

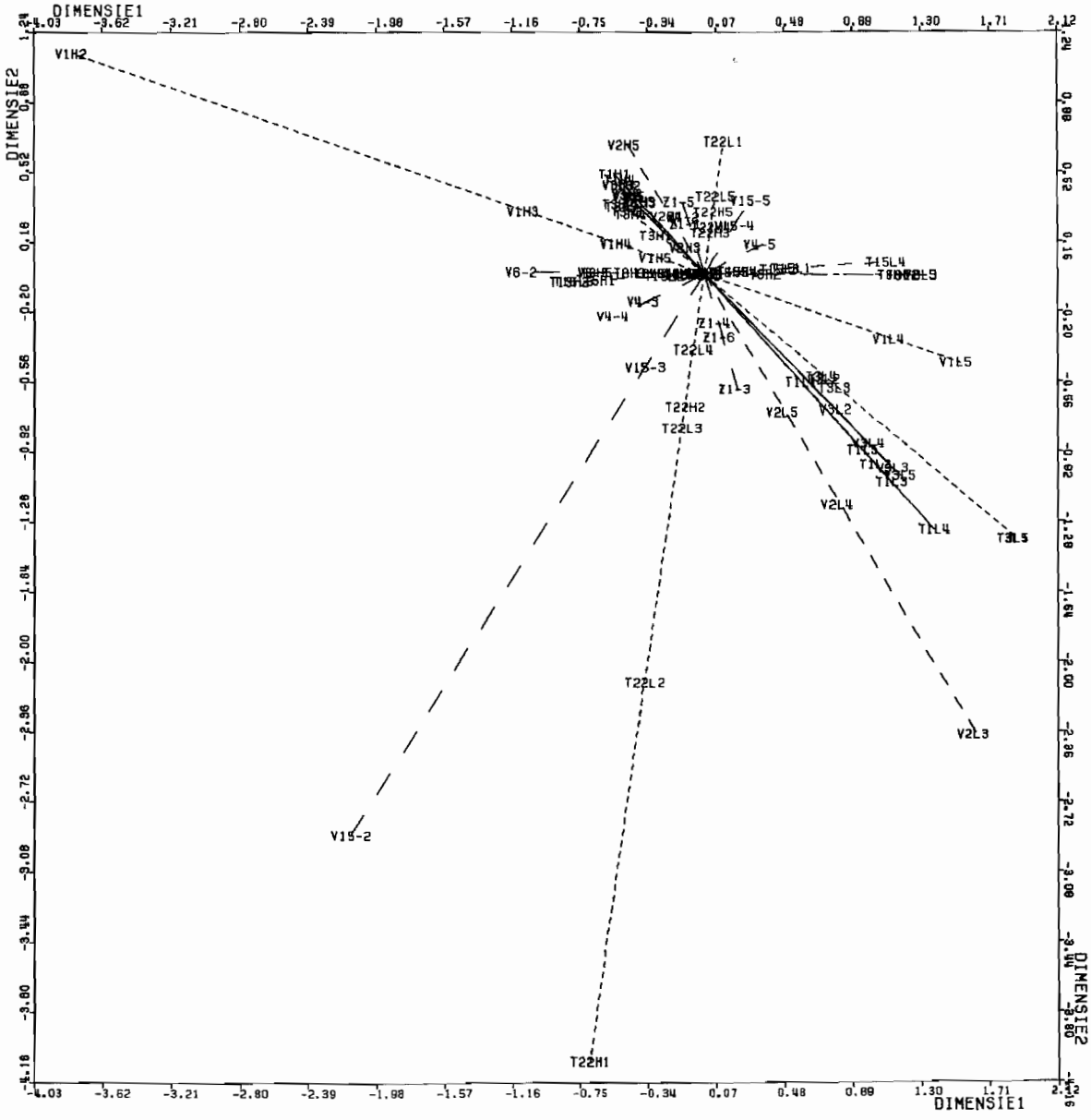
In de daarna uitgevoerde analyse kregen alle variabelen het enkelvoudig nominale meetnivo opgelegd. Bovendien werd er een persoon (nr. 153) buiten de analyse gelaten, omdat deze de oplossing erg domineerde. Deze persoon was dus een "uitbijter", dat wil zeggen dat hij/zij in de kanoniese ruimte ver van de andere personen vandaan ligt. In mindere mate geldt dit ook voor persoon nr 136. Deze ligt helemaal links onderaan in figuur 8. Uit deze figuur wordt ook het verschil duidelijk tussen mensen met een lage autonomie (gelabeld met een '1') preferentie hebben en mensen die een hoge autonome positie preferen. Deze laatste zijn met een label '2' aangegeven. Dat dit gevonden wordt is natuurlijk niet ver-



Figuur 8 : Objektskores 'Job Characteristics Model' met interaktieve variabelen

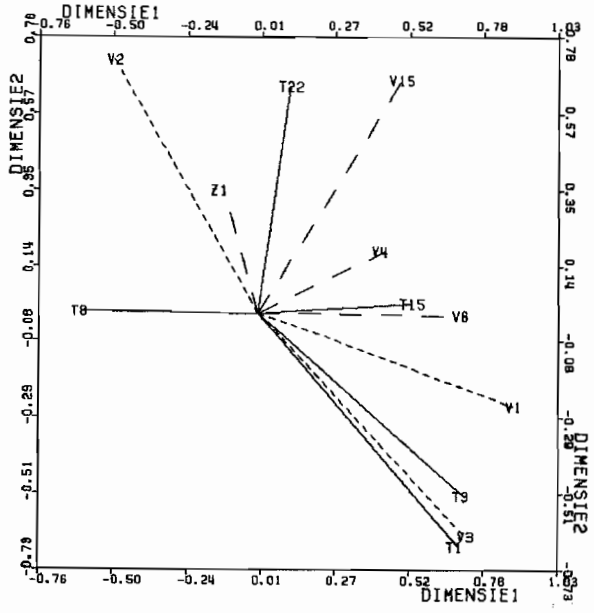
wonderlijk, maar rechtsstreeks een gevolg van het feit dat twee sets interaktief gemaakt zijn met de modererende variabele. Maar naast sit effekt blijft het interessant om te kijken hoe de andere verbanden tussen de variabelen liggen.

In figuur 8 is te zien waar de kategorieën van al deze variabelen zich in de kanoniese ruimte bevinden. (In bijlage 4 wordt een overzicht van de enkelvoudige kategoriekwantifikaties gegeven). De koördinaten van deze punten worden gegeven door de "average rank-one" kwantifikaties, die terug gevonden kunnen worden in de OVERALS output. De labels van de kategorieën in deze figuur hebben het oorspronkelijke kategorienummer behouden (laatste cijfer) en met een L of een H is aangegeven of het respon-



Figuur 9 : Positie van de kategorieën in de kanoniese ruimte

denten betreft die tot de laag autonome subgroep behoren (L) of juist tot de hoog autonome groep (H). De eerste twee letters geven aan met welke variabele we te maken hebben. Wat is er in deze figuur te zien :
Variabele V1 onderscheidt hoog of laag autonoom volledig van elkaar. De L-kategorieën, dus kategorieën die laag autonome personen bevatten liggen rechts van de oorsprong. Dit in tegenstelling tot hoog autonome kategorieën die links van de oorsprong liggen. Binnen beide groepen is er nog steeds een mooi onderscheid tussen mensen die het werk als nutteloos beschouwen of juist vinden zinvol werk te doen. Voor variabele V2 en in mindere mate voor variabele V3 wordt ook deze tweedeling teruggevonden. Ook hier wordt binnen de twee groepen min of meer een ordinaal verband tussen de kategorieën gevonden. Dus binnen de groepen wordt onderscheiden of wel/niet een grote eigen verantwoordelijkheid voor het werk wordt ervaren (V2) of dat men wel/niet goed weet of het werk goed of slecht is uitgevoerd (V3). Voor de taakkenmerken wordt niet zo'n mooie volgorde van de kategorieën gevonden. Wel wordt voor T1 (mate waarin blijkt dat de taak goed/slecht is uitgevoerd) en voor T3 (ernst van mogelijke fouten of vergissingen) de tweedeling in hoog/laag autonoom gevonden. Voor de andere drie taakkenmerken (T8, T15 en T22) geldt dit onderscheid niet. Bij deze variabelen is wel enigzins de oorpronkelijke ordinale kategorie volgorde gehandhaafd. Dit gaat nog het minst op voor T15 (mate waarin zelf beslist kan worden over de werkmethode). Maar voor T8 en T22 kunnen we wel konstateren dat de lage kategorieën (1 en 2) aan de ene zijde en de hoge kategorieën (4 en 5) aan de andere zijde liggen in de figuur. (Zie ook bijlage 3 als het uit de figuur niet duidelijk wordt). Deze variabelen dragen niet veel bij aan de hoog/laag autonome subgroepsindeling. T8 onderscheidt afwisselend werk van steeds terugkerende werkzaamheden en T22 maakt onderscheid tussen wel of niet een belangrijk onderdeel binnen een groter geheel. Bij variabele T15 (zelf beslissen) kan beter naar de individuele kategorie gekeken worden, want hier is geen echte volgorde te konstateren. We hebben nu alleen nog maar bekeken hoe wel/niet autonoom zich in deze analyse gedraagt, maar nog niet of het Job Characteristics Model nog uit deze resultaten te voorschijn komt. In figuur 9 (of misschien duidelijker in figuur 10) is te zien dat T1 en V3 zeer dicht bij elkaar liggen, wat konform het JCM model is. Ook T3 en V2 hangen bijzonder nauw samen met deze twee variabelen. Dit is natuurlijk voornamelijk te wijten aan de hoog/laag autonoom indeling. De enige afhankelijke variabele die



Figuur 10: Komponentladingen JCM model met interaktieve variabelen

hiermee samenhangt is "ziekteverzuim" (Z1). Mensen die erg weinig verzuimen vallen over het algemeen in de hoog autonome subgroep, doen minder verantwoordelijk werk (V2), krijgen veel feedback (T1) en ervaren ook dat ze goed weten of ze het goed of slecht doen (V3). Mensen die 5 tot 8 dagen verzuimen verschillen alleen van de hierboven genoemde groep (0 tot 4 dagen ziekteverzuim) omdat ze een lage autonomiepreferentie hebben i.p.v. dat ze een hoge autonomie prefereren. Ook bij de kategorieën die een hoog ziekteverzuim representeren vinden we hoog en laag autonome personen. Het hoogste ziekteverzuim (meer dan 30 dagen) wordt gevonden voor laag autonome respondenten, die redelijk verantwoordelijk werk hebben, over weinig informatie beschikken m.b.t. hoe de taak is uitgevoerd en dit ook zo ervaren. Voor respondenten, die vinden dat ze zeer verantwoordelijk werk hebben waarvan ook niet duidelijk is of de resultaten goed of slecht zijn, en waarvan ook m.b.v. taakkenmerk T1 gemeten is dat die taak weinig informatie hierover geeft vinden we personen die 9 tot 14 dagen verzuimen min of meer terug in de laag autonome subgroep en personen die binnen de kategorie 15 tot en met 29 dagen vallen in de hoog autonome groep. Ook is hier een verband met T3 en V1 te konstateren, waarvan de laatste (V1) in iets mindere mate het onderscheid tussen de subgroepen weerspiegeld. Voor variabele T3 wordt binnen beide subgroepen een wel - niet richting van de ernst van mogelijke fouten gevonden.

En voor de variabele V1 vinden we een nutteloos - zinvol richting binnen de subgroepen. Dit gebeurt op zo'n manier dat verantwoordelijk werk samengaand met weinig informatie en feedback tegelijkertijd samengaat met ernstige gevolgen van fouten of vergissingen (T3) en met het gevoel dat men zinvol werk verricht (V1). En voor de tegengestelde kategorieën gaat dit verband ook op. Verder vertoont deze V1 nog een samenhang met taakkenmerk T8 (terugkeren van steeds dezelfde werkzaamheden) en taakkenmerk T15 (zelf beslissen over werkmethode). Ook de afhankelijke variabele R6 en R4 (maar minder) houden hier verband mee. In woorden : Taak met afwisselende werkzaamheden houdt verband met gemotiveerde werknemers (R4) en veel zelfvertrouwen (R6). Hoe het verband met T15 is minder makkelijk af te leiden omdat de volgorde van de kategorieën geen logiese ordening heeft. Tot slot vertoont T22 nog enigzins samenhang met R15. Werk wat niet zo'n belangrijk onderdeel van een groter geheel is (T22) wordt over het algemeen als leuk ervaren (R15). Deze variabelen staan niet in eerste instantie in verband met 'psychologiese toestanden'. Dit komt dus niet met JCM model overeen. Verder zou T15 moeten samenhangen met V2 i.p.v. met V1 zoals uit deze analyse naar voren komt. En voor T22 had een verband moeten gelden met V1. Dit geldt ook voor T3, die in onze analyse voornamelijk verband heeft met V3.

Deze analyse wordt dus voor een groot deel gedomineerd door het grote effekt van de interaktieve variabele. Daarom wordt hier minder de andere verbanden uit het JCM model duidelijk. Een afzonderlijke analyse van de hoog/laag autonome respondenten zou een meer aangewezen manier zijn, maar omdat dit veel 'uitbijters' opleverde is deze manier verder achterwege gelaten. Bovendien had dit een minder leuke OVERALS illustratie opgeleverd.

LITERATUUR

Algera, J.A. (1980). Kenmerken van werk. Meppel, Krips repro

Benzécri, J.P., e.a. (1973). Analyse des donées (2 vols.) Paris, Dunod

Burg, E. van der (1983). CANALS user's guide. Department of Datatheory, University of Leiden

Burg, E. van der, and Leeuw, J. de (1983). Non-linear canonical correlation. British Journal of Mathematical and Statistical Psychology, 36, 54-80

Burg E. van der, Leeuw, J. de & Verdegaal, R.(1984). Non-linear canonical carrelation with M sets of variables. Submitted for publication.

Burt, C. (1950). The factorial analysis of qualitative data. British Journal of Statistical Psychology, 3, 166-185

Carroll, J.D. (1968). A generalisation of canonical correlation analysis to three or more sets of variables. Proceedings 76th Convention American Pshychological Association, 227-228

Dauxois, J. et Pousse, A. (1976). Les analyses factorielles en calcul des probabilités et en statistique : essai d'étude synthétique. Dissertation. Université Paul Sabatier, Toulouse

Eckart , C., and Young G. (1936). The approximation of one matrix by another of lower rank. Psychometrika, 1, 211-218

Edgerton, H.A., and Kolbe, L.E. (1936). The method of minimum variation for the combination of criteria. Pscychometrika, 1, 183-187

Fisher, R.A. (1940). The precision of discriminant functions. Ann. Eug., 10, 422-429

Geer, J.P. van de (1980). Introduction to multivariate lineair data analysis. Part V : Relations between K sets of data. Department of Datatheory, University of Leiden

Geer, J.P. van de (1984). Linear relations between k sets of variables. Pshychometrica, 49, 79-94

Gifi, A. (1981). Non-linear multivariate analysis. Department of Datatheory, University of Leiden

Gifi, A. (1981b). HOMALS user's guide. Department of Datatheory, University of Leiden

Gifi, A. (1983). PRINCALS user's guide. Department of Datatheory, University of Leiden

Guttman, L. (1941). The quantifications of a class of attributes: a theory and method of scale construction. In : P. Horst (ed.), the prediction of personal adjustment. New York, Social Science Research Council, 251-364

Hackman, J.R. and Oldham, G.R. (1976). Motivation through the design of work: test of a theory. Organizational Behavior and Human Performance, 16, 250-279

Hayashi, C. (1950). On the quantification of qualitative data from the mathematico-statistical point of view. Annals of the Institute of Statistical Mathematics, 2, 35-47

Hayashi, C. (1956). Theory and examples of quantification II. Proceedings Institute of Statistical Mathematics, 4, 19-30

Horst, P. (1936). Obtaini-g a composite measure from a number of different measures of the same attribute. Psychometrika, 1, 53-60

Horst, P. (1961a). Relations among m sets of variables. Psychometrika, 26, 129-149

Horst, P. (1961b). Generalised canonical correlations and their applications to experimental data. Journal of Clinical Psychology (Monograph Supplement), 14, 331-347

Hotelling, H. (1935). The most predictable criterion. Journal of Educational Psychology, 26, 139-142

Hotelling, H. (1936). Relations between two sets of variables. Biometrica, 28, 321-377

Kettenring, J.R. (1971). Canonical analysis of several sets of variables. Biometrica, 58, 433-460

Kruskal, J.B. and Shepard, R.N. (1974). A nonmetric variety of linear factor analysis. Psychometrika, 39, 123-157

Leeuw, J. de (1973). Canonical analysis of categorical data. Dissertation. University of Leiden. DSWO Press, Leiden, 1984

Leeuw, J. de (1977). A normalised cone regression approach to alternating least squares algorithms. Department of Datatheory, University of Leiden

Leeuw, J. de (1983). Notes: Multiple mysteries; More on multiple; Partitioning loss in OVERALS; Correlation matrices and eigenvalues in OVERALS; OVERALS plots; OVERALS sans larmes; The missing mystery. Department of Datatheory. University of Leiden.

Leeuw, J. de (1984). The Gifi system of nonlinear multivariate analysis. In: Data Analysis and Informatics. E. Diday et al. (eds.), North-Holland Publishing Company, Amsterdam

Leeuw, J. de (1984a). Beyond homogeneity analysis. Department of Datatheory. University of Leiden

Leeuw, J. de & Rijckevorsel J. van (1980). HOMALS and PRINCALS, some generalisations of principal component analysis. In : Data Analysis and Informatics. E. Diday et al. (eds.), North-Holland Publishing Company, Amsterdam

McDonald, W.R. (1967). Nonlinear factor analysis. Pshychometric monographs, No. 15

McKeon, J.J. Canonical analysis: some relations between canonical correlation, factor analysis, discriminant function analysis, and scaling theory. Pshychometric Monographs, No. 13

Nishisato, S. (1980). Analysis of categorical data: dual scaling and its applications. Toronto, University of Toronto Press

Oppe, S. (1983). An analysis of the similarities and dissimilarities of the international observation teams in Malmö. Leidschendam, Institute for Road Safety Research SWOV, The Netherlands

Pearson, K. (1901). One lines and planes of closest fit to points in space. Phil. Magazine, 2, 559-572

Pearson, K. (1904). Mathematical controbutions to the theory of evolution XIII. On the theory of contingency and its relation to association and normal correlation. Drapers Company Research Memoirs, Biometric Series, No. 1

Pearson, K. (1906). On certain points connected with scale order in the case of a correlation of two characters which for some arrangement give a linear regression line. Biometrica, 5, 176-178

Steel, R.G.D. (1951). Minimum generalised variance for a set of linear functions. Annals of Mathematical Statistics, 22, 456-460

SWOV (1984). The Malmö Study: A calibration of traffic conflict techniques. Leidschendam, Institute for Road Safety Research SWOV, The Netherlands

Takane, Y., Young, F. W. & Leeuw, J. de (1980). An individual differences additive model. An alternating least squares method with optimal scaling features. Psychometrica, 45, 183-209

Wilks, S.S. (1936). Weighting systems for linear functions of correlated variables when there is no independent variable. Psychometrica, 3, 23-40

Young, F.W., Leeuw, J. de & Takane, Y. (1976). Regression with qualitative and quantitative variables. An alternating least squares method with optimal scaling features. Psychometrika, 41, 505-529

Young, F.W. (1981). Quantitative analysis of qualitative data. Psychometrika, 46, 347-388

Bijlage 1

Betekenis van de kategorieën van de variabelen

Team 1 t/m 8
1 = niet als konflikt geskoord
2 t/m 4 = minder tot meer ernstig konflikt

Type konflikt
1 = C - C
2 = C - P
3 = C - B
4 = L - L, C - L
5 = P - B
6 = P - L
7 = B - B
8 = B - L
9 = anders

C = auto, taxi
L = bus, vrachtwagen
B = (brom)fiets
P = voetganger

Manoeuvre type
zie volgende bladzijde

V1 initiele snelheid van weggebruiker 1 (degene, die voorrang heeft) in m/s
1 = 0.4 tot 1.9
2 = 2.0 tot 5.9
3 = 6.0 tot 10.9
4 = 11.0 tot 13.9
5 = 14.0 tot 19.7

V2 initiele snelheid van weggebruiker 2 in m/s
1 = 0.1 tot 1.9
2 = 2.0 tot 2.9
3 = 3.0 tot 4.9
4 = 5.0 tot 9.9
5 = 10.0 tot 17.7

A1 maksimale versnelling weggebruiker 1 in m/s^2
1 = -7.7 tot -4.0
2 = -3.9 tot -3.0
3 = -2.9 tot -2.0
4 = -1.9 tot -1.0
5 = -0.9 tot +0.9
6 = 1.0 tot 3.3

A2 maksimale versnelling weggebruiker 2 in m/s^2
idem A1

MD minimale afstand tussen de weggebruikers in meters
1 = 0.0 tot 0.4
2 = 0.5 tot 0.9
3 = 1.0 tot 2.0
4 = 2.1 tot 3.4
5 = 3.5 tot 11.4

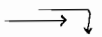
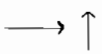
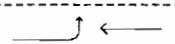
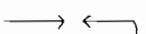
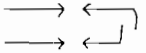
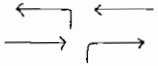
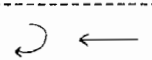
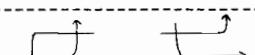
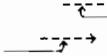
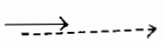
TTC minimale TTC waarde in sekonden
TTC (Time To Collision) wordt gedefinieerd als de tijd die nog over is voor er een botsing plaatsvindt, als de snelheden en richtingen van de weggebruikers niet veranderen.
1 = 0.0 tot 0.9
2 = 1.0 tot 1.5
3 = 1.6 tot 4.5
4 = missing

DTTC afstand tussen weggebruikers bij minimale TTC in meters
1 = 0.0 tot 1.0
2 = 1.1 tot 3.0
3 = 3.1 tot 6.0
4 = 6.1 tot 12.0
5 =12.1 tot 47.0
6 = missing

PET tijd om punt te bereiken waar andere weggebruiker is binnengedrongen in sekonden
1 = 0.0 tot 0.5
2 = 0.51 tot 0.9
3 = 1.0 tot 1.4
4 = 1.5 tot 2.0
5 = missing

OB8 PRINCALS skore
1 = $\geq$ 1.0
2 = 0.5 tot 1.0
3 = 0.0 tot 0.5
4 = -1.0 tot 0.0
5 = -2.0 tot -1.0
6 = $\leq$ -2.0

Manoeuvre (M4)

Code	Manoeuvre	
1	Rear end	
	Rear end with left turn	
	Rear end with right turn	
2	Weave or merge	
	Right angle (cut-in)	
3	Right angle	
4	Head-on	
5	Left turn	
	Head-on with left turn	
	Head-on with right turn	
6	Right angle with left turn	
	Right angle with right turn	
7	U-turn	
8	Double left turn	
	Left turn with opposing right turn	
9	Pedestrian with vehicle on straight path	
	Pedestrian with right turn	
	Pedestrian with left turn	
	Pedestrian crossing at angle	
10	Other	

Bijlage 2

kode

T1 De mate waarin bij het uitvoeren van een taak blijkt in hoeverre de taakuitvoerder het goed of slecht doet.
1= krijgt veel info over prestaties en het effect van handelingen.
5= " weinig " " " " " " " " " .

T3 De ernst van mogelijke fouten of vergissingen.
1= heeft zeer ernstige gevolgen.
5= " geen " " .

T8 De mate waarin steeds dezelfde werkzaamheden terugkomen.
1= niet vaak komen zelfde werkzaamheden terug.
5= steeds komen zelfde werkzaamheden terug.

T15 de mate waarin de taakuitvoerder zelf kan beslissen over zijn methode van werken.
1= kan niet zelf beslissen.
5= " wel " " .

T22 De mate waarin de taak een herkenbare bijdrage levert aan een groter geheel; de belangrijkheid van een taak in groter verband.
1= minder belangrijk onderdeel in een groter geheel.
5= een zeer " " " " " " .

V1 Experienced meaningfulness.
1= ik vind het erg nutteloos werk
5= ik vind het erg zinvol werk

V2 Experienced responsibility.
1= ik vind het volstrekt geen verantwoordelijk werk
5= ik vind het werk zeer verantwoordelijk

V3 Knowledge of results.
1= het is moeilijk te achterhalen of ik het werk goed doe of niet
5= zeker weten dat je het werk goed doet of juist slecht

R4 Internal work motivation.
1= ik ben weinig gemotiveerd om het werk goed te doen
5= ik ben erg gemotiveerd om het werk goed te doen

R6 Growth satisfaction.
1= het werk geeft mij weinig zelfvertrouwen/eigenwaarde
5= het werk geeft mij veel zelfvetrouwen/eigenwaarde

R15 Werksatisfactie
1= ik vind het werk eentonig/niet de moeite waard
5= ik vind het werk erg leuk/ben heel enthousiast

Z1 Ziekteverzuim.
1= 0 dagen
2= 1 t/m 4 dagen
3= 5 t/m 8 dagen
4= 9 t/m 14 dagen
5=15 t/m 29 dagen
6=meer dan 30 dagen

M18 (Preferentie voor) autonomie
1= ik vind een werksituatie prettig, waarin precies wordt gezegd wat ik wel of niet moet doen
5= ik vind werk aantrekkelijk, waarin ik mijn eigen gang kan gaan

Bijlage 3.

VAR.	KAT.	ENKELVOUDIGE KATEGORIEKWANTIFIKATIES
T1	1	0.841
T1	2	-0.778
T1	3	1.487
T1	4	-0.687
T1	5	1.626
T1	6	-0.553
T1	7	1.994
T1	8	-0.732
T1	9	1.376
T1	10	-0.559
T3	1	2.637
T3	3	-0.373
T3	3	1.045
T3	4	-0.688
T3	5	1.136
T3	6	-0.613
T3	7	1.022
T3	8	-0.586
T3	9	2.636
T3	10	-0.662
T8	1	0.0
T8	2	-1.930
T8	3	-2.023
T8	4	-0.647
T8	5	-2.203
T8	6	0.704
T8	7	0.207
T8	8	0.201
T8	9	0.504
T8	10	1.039
T15	1	0.988
T15	2	-1.238
T15	3	-0.458
T15	4	-1.544
T15	5	0.858
T15	6	-0.231
T15	7	2.069
T15	8	0.377
T15	9	-1.491
T15	10	0.273
T22	1	1.055
T22	2	-6.271
T22	3	-3.253
T22	4	-1.058
T22	5	-1.229
T22	6	0.323
T22	7	-0.602
T22	8	0.369
T22	9	0.616
T22	10	0.490

VAR.	KAT.	ENKELVOUDIGE KATEGORIEKWANTIFIKATIES
V1	1	0.0
V1	2	0.0
V1	3	0.0
V1	4	-4.307
V1	5	0.031
V1	6	-1.215
V1	7	1.290
V1	8	-0.577
V1	9	1.747
V1	10	-0.306
V2	1	0.0
V2	2	0.0
V2	3	0.0
V2	4	0.0
V2	5	-3.299
V2	6	0.185
V2	7	-1.657
V2	8	0.408
V2	9	-0.991
V2	10	0.922
V3	1	0.0
V3	2	0.0
V3	3	1.103
V3	4	0.0
V3	5	1.584
V3	6	-0.721
V3	7	1.380
V3	8	-0.645
V3	9	1.644
V3	10	-0.631
R4	1	0.0
R4	2	0.0
R4	3	-0.800
R4	4	-1.218
R4	5	0.827
R6	1	0.0
R6	2	-1.521
R6	3	-0.906
R6	4	-0.298
R6	5	1.867
R15	1	0.0
R15	2	-4.350
R15	3	-0.722
R15	4	0.370
R15	5	0.567
Z1	1	0.731
Z1	2	0.832
Z1	3	-1.689
Z1	4	-0.717
Z1	5	1.059
Z1	6	-0.925