Canonische data-analyse

Dr. A.W. Vogelesang, P.J.G. Verhoef & drs. S. Oppe

D-2000-3

Canonische data-analyse

Leiden verschillende procedures tot verschillende resultaten?

D-2000-3
Dr. A.W. Vogelesang, P.J.G. Verhoef & drs. S. Oppe
Leidschendam, 2000
Stichting Wetenschappelijk Onderzoek Verkeersveiligheid SWOV

Documentbeschrijving

Rapportnummer:	D-2000-3
Titel:	Canonische data-analyse
Ondertitel:	Leiden verschillende procedures tot verschillende resultaten?
Auteur(s):	Dr. A.W. Vogelesang, P.J.G. Verhoef & drs. S. Oppe
Onderzoeksmanager:	Ing. J.A.G. Mulder
Projectnummer SWOV:	74.207
Trefwoord(en):	Mathematical model, correlation (math, stat), data processing, program (computer), analysis (math).
Projectinhoud:	Bij de analyse van problemen in de verkeersveiligheid zijn de klassieke technieken voor multivariate analyse vaak niet toepasbaar. Voor een belangrijk deel heeft dit te maken met het niveau waarop de data zijn gemeten. Sommige kenmerken zijn echte metingen, maar soms is alleen de rangorde van gegevens bekend, of slechts een indeling in klassen zónder dat daar een ordening aan kan worden toegekend. Indien we relaties tussen niet-lineaire kenmerken willen onderzoeken, kunnen we gebruikmaken van niet-lineaire canonische analyse. Een canonische analyse kan volgens verschillende procedures plaatsvinden. In dit onderzoek wordt nagegaan of deze verschillende procedures ook dezelfde resultaten opleveren, in die zin dat de interpretatie van de analyseresultaten tot dezelfde conclusies leiden.
Aantal pagina's:	24 + 13 blz.
Prijs:	f 20,-
Uitgave:	SWOV, Leidschendam, 2000

Stichting Wetenschappelijk Onderzoek Verkeersveiligheid SWOV
Postbus 1090
2260 BB Leidschendam
Telefoon: 070-3209323
Telefax: 070-3201261

Inhoud

1. Inleiding

De klassieke technieken voor multivariate analyse, zoals onder andere variantieanalyse, factoranalyse, multipele regressie en canonische analyse, zijn vaak niet toepasbaar bij de analyse van problemen in de verkeers-veiligheid. Voor een belangrijk deel heeft dit te maken met het niveau waarop de data zijn gemeten. Sommige kenmerken zijn echte metingen, zoals snelheden van voertuigen of leeftijden van bestuurders, maar soms is alleen de rangorde van gegevens bekend, of slechts een indeling in klassen zónder dat daar een ordening aan kan worden toegekend. Een voorbeeld hiervan is het nominale kenmerk ‘wijze van verkeersdeelname’. Nog moeilijker wordt het als kenmerken gerelateerd moeten worden aan verschillende reeksen ongevallen tegelijk (zie Oppe, 1992).
Een bijkomend probleem is dat kenmerken vaak worden onderzocht in relatie tot het aantal ongevallen. In dat geval hebben we geen metingen, maar tellingen. Lineaire modellen zijn hier lang niet altijd van toepassing, meestal worden log-lineaire modellen gebruikt.

Indien we relaties tussen niet-lineaire kenmerken willen onderzoeken, kunnen we gebruikmaken van niet-lineaire canonische analyse.

2. Toelichting canonische analyse

Vooraleer tot het doel van het onderzoek te geraken willen we eerst een kleine toelichting geven op de betekenis van canonische analyse zoals deze geïmplementeerd is in CANALS (Gifi, 1981; Van der Burg & de Leeuw, 1983; De Leeuw, 1984; Van der Burg, 1988; Gifi, 1990) en OVERALS (Gifi, 1981, 1990; Verdegaal, 1986). Deze implementatie heeft plaatsgevonden in het kader van de multivariate technieken die ontwikkeld zijn bij de vakgroep Datatheorie van de Universiteit Leiden.

In de situatie waarin twee sets van variabelen moeten worden vergeleken, is het uitvoeren van een CANALS-analyse een goede manier om de overeenkomst tussen deze twee sets te onderzoeken. Bij canonische analyse wordt de correlatiecoëfficiënt tussen een gewogen som van variabelen uit een eerste set van variabelen en een gewogen som van variabelen uit een tweede set gemaximaliseerd. De lineair gewogen sommen worden de ‘canonical variates’, ‘canonische variabelen’ of ‘canonische assen’ genoemd. De correlatie tussen de canonische assen is de ‘canonische correlatie’. Lineaire transformatie van de oorspronkelijke variabelen leidt tot dezelfde canonische correlatiecoëfficiënt. Anders gezegd: de resultaten van CCA (Canonische CorrelatieAnalyse) zijn *invariant onder lineaire transformaties*.

2.1. Datatransformaties

Het principe van invariantie onder bepaalde transformaties is ook toegepast op andere schaalniveaus, namelijk op nominaal en ordinaal niveau. *Ordinale* data zijn invariant onder *monotone* transformaties, dat wil zeggen transformaties die de ordening van de data behouden. *Nominale* data zijn invariant onder transformaties die gelijken (‘ties’) gelijk houden en de categorieën gescheiden houden: de zogenaamde *multipel nominale transformaties* (de transformaties uit de correspondentieanalyse, zie Van der Burg, 1988; Benzécri, 1973).
Als er verscheidene (*k*) groepen/sets zijn, spreekt men van *k*-sets CCA of *generalised canonical analysis*. Deze methode van canonische analyse is gebaseerd op ‘optimale transformaties’ van de variabelen, in combinatie met een ‘alternating least squares’ schattingsalgoritme voor het vinden van een oplossing (Gifi, 1981; 1990). Deze methode is geschikt voor data van numeriek, ordinaal en nominaal niveau. ‘Gifi’ is *nom de plume* voor de groep auteurs aan de Rijksuniversiteit Leiden die de procedures ontwikkeld en beschreven heeft.

2.2. Canonische correlatieanalyse

De canonische correlatietechniek voor de analyse van continue variabelen gaat terug op Hotelling (1936). Canonische correlatie voor de analyse van kruistabellen werd voor het eerst toegepast door Sir Ronald Fisher (1940). De techniek werd pas echt populair in de jaren tachtig (zie Goodman, 1981, en de daarin vermelde referenties). Toepassing van optimale transformaties, om te komen tot optimale (= maximale) correlaties tussen de variabelen in een kruistabel, stamt ook uit de jaren veertig (Hirschfeld, 1935). Belangrijke publicaties zijn die van Hill (1974, ‘reciprocal averaging’),

van Guttman (1941; 1946; de ‘Guttman-schaal’), van Nishishato (1978, ‘optimal scaling’) en van Goodman (1981).

3. Doel van het onderzoek

Het doel van dit onderzoek is om na te gaan of de verschillende procedures waarmee een canonische analyse kan worden uitgevoerd ook dezelfde resultaten opleveren, in die zin dat de interpretatie van de analyseresultaten tot dezelfde conclusies leiden.

3.1. Programmatuur

Voor de uitvoering van canonische analyses zijn er diverse programma's ontwikkeld:

1. CANALS: ontwikkeld door de vakgroep Datatheorie van de Universiteit Leiden (Van der Burg & de Leeuw, 1983; Van der Burg, 1988; Van der Burg, De Leeuw & Verdegaal, 1988);
2. TRANSREG: de CANALS-procedure is in aangepaste vorm opgenomen in SAS, als optie in PROC TRANSREG (zie Kuhfeld, 1985; Kuhfeld, Sarle & Young, 1985; Kuhfeld, Young & Kent, 1987);
3. OVERALS: Leidse procedure voor canonische analyse (of homogeniteitsanalyse) van verschillende sets van variabelen (Van der Burg & de Leeuw, 1986; Verdegaal, 1986; Van der Burg, 1988). Voor twee sets variabelen moet het dezelfde resultaten opleveren als CANALS.
4. SPSS: OVERALS is opgenomen in SPSS (SPSS, 1990; Gifi, 1990).

3.2. Optimale schaling versus datatransformaties

Door de vakgroep Datatheorie aan de Universiteit Leiden zijn programma's ontwikkeld die wel kunnen worden toegepast in bovengenoemde situaties. Ook aantallen kunnen met deze programma's worden geanalyseerd. Het belangrijkste kenmerk van deze technieken is dat gezocht wordt naar een zodanige weergave van de klassen van een kenmerk op een metrische schaal, dat een optimale oplossing met behulp van klassieke technieken wordt bereikt - het meetniveau van de data in aanmerking genomen. Het meest gebruikte programma is HOMALS (homogeniteitsanalyse op basis van een Alternating Least Squares of 'ALS'-algoritme), Principale ComponentenAnalyse (PCA), gegeneraliseerd naar (multipel) nominaal, ordinaal en numeriek meetniveau. HOMALS wordt veelvuldig gebruikt bij de analyse van enquêtegegevens om samenhangen tussen antwoordcategorieën of antwoordprofielen op te sporen, of als multidimensionele schaaltechniek.

Naast HOMALS zijn onder andere CANALS en OVERALS ontwikkeld. CANALS is ontwikkeld voor canonische correlatieanalyse, multipele regressieanalyse en discriminantanalyse.
OVERALS is een programma voor canonische analyse van verschillende groepen/sets; voor andere toepassingen zie Van der Burg, (1988). Voor een uitgebreide behandeling van de Leidse programma's wordt verwezen naar Gifi (1981, 1990). OVERALS voor twee sets en met 'single restrictions' is hetzelfde als CANALS (zie § 3.1).

CANALS-LEIDEN, CANALS-TRANSREG, OVERALS-LEIDEN en OVERALS-SPSS zijn procedures voor *niet-lineaire canonische correlatie-*

analyse (zie Van der Burg, 1988). Canonische correlatieanalyse maximaliseert de correlatie tussen een lineaire combinatie van variabelen in de ene groep en een lineaire combinatie van variabelen in de andere groep. Er kunnen méér canonische correlaties gemaximaliseerd worden, onder de voorwaarde dat de afzonderlijke lineaire combinaties onderling onafhankelijk zijn. De lineaire combinaties zijn weer de canonische assen. Als de canonische analyse gecombineerd wordt met *optimale schaling*, spreken we van *niet-lineaire canonische correlatieanalyse*; de drie bovengenoemde programma's zijn alle voorbeelden hiervan. CANALS en CANALS-TRANSREG zijn bedoeld voor 2-sets analyses en OVERALS en OVERALS-SPSS voor *k*-sets analyses ($k \geq 2$). Verwacht kan worden dat er geen noemenswaardig verschil is tussen OVERALS-LEIDEN en OVERALS-SPSS (OVERALS is overgenomen in SPSS). Daarentegen is CANALS-TRANSREG opnieuw geprogrammeerd en opgenomen in SAS-PROC TRANSREG (Kuhfeld, 1985; Kuhfeld, Sarle & Young, 1985; Kuhfeld, Young & Kent, 1987).

3.3. OVERALS versus CANALS: multiple en single transformaties

OVERALS-SPSS is een procedure voor canonische analyse van verscheidene sets van variabelen, maar zal voor twee sets slechts onder speciale condities dezelfde resultaten geven als CANALS. Indien voor een variabele op elke dimensie dezelfde *optimale* (zie § 3.4) transformatie wordt gebruikt, spreekt men van '*k*-sets CCA met single transformations'. *Single transformations* betekent dat voor elke dimensie dezelfde transformatie wordt gebruikt; bij *multiple transformations* (zoals multipel nominaal) mag de transformatie verschillen per dimensie. Als voor elke dimensie dezelfde optimale transformatie wordt gebruikt, is de meetkundige representatie van de getransformeerde variabele een kegel. Uit deze kegel wordt één vector gekozen, diegene die de canonische correlatie maximaliseert. De transformatie die overeenkomt met de richting van deze vector is de zgn. 'optimale transformatie'. OVERALS voor twee sets en met 'single restrictions' is hetzelfde als CANALS. OVERALS zónder de restrictie van enkelvoudige transformaties is 'OVERALS-met-multipele-restricties' en is gelijk aan HOMALS: homogeniteitsanalyse, nagaan of verschillende sets op hetzelfde neerkomen (Gifi, 1990).

3.4. Optimale schaling

Optimale schaling is een methode om nominale, ordinale of numerieke transformaties aan de variabelen toe te kennen. De variabelen worden zó getransformeerd, dat de modelpassing optimaal wordt. Dit gebeurt tijdens een iteratieve procedure, waarbij de transformaties en de gewichten voor de lineaire combinaties alternerend worden aangepast. Op dezelfde manier als voor twee groepen kan optimale schaling toegevoegd worden aan canonische analyse voor *k* groepen. Dat is OVERALS.

OVERALS-SPSS is gebaseerd op OVERALS, zoals die ontwikkeld is bij de vakgroep Datatheorie van de Universiteit Leiden (Van der Burg & de Leeuw, 1986). Verwacht wordt dat *alleen* bij het vergelijken van *twee* datasets (onder *single* restricties), OVERALS-SPSS en CANALS-LEIDEN dezelfde resultaten zullen geven.

In het hiernavolgende wordt CANALS-LEIDEN afgekort tot ‘CANALS’ en wordt CANALS-TRANSREG aangeduid met ‘TRANSREG’. Tevens kan OVERALS-SPSS worden afgekort tot ‘OVERALS’ zonder dat dit tot verwarring leidt. Paragraaf 7.3 geeft de setup van een OVERALS-SPSS-analyse voor de voorbeelddata.

Niet-lineaire CCA (canonische correlatieanalyse) met *k* sets van variabelen is zowel een generalisatie van lineaire CCA met *k* sets van variabelen als ook een generalisatie van niet-lineaire CCA met twee sets (Van der Burg, 1988). De OVERALS-variant die de relaties tussen *k* variabelen analyseert door ze te stoppen in *k* sets met elk één variabele, is vergelijkbaar met HOMALS - maar dat doet hier eigenlijk niet terzake. Het gaat om OVERALS toegepast op twee sets met in de tweede set de afhankelijke variabele. Deze variant is vergelijkbaar met CANALS. In Van der Burg & de Leeuw (1983) wordt OVERALS beschouwd als een *vectorgeöriënteerde* techniek met een vector-interpretatie van de variabelen, een volstrekt andere interpretatie van de variabelen dan bij homogeniteitsanalyse. Alleen bij twee datasets kunnen beide technieken op hetzelfde neerkomen (zie boven). Van der Burg, de Leeuw & Verdegaal (1986) beschouwen OVERALS als *homogeniteitsanalyse* (‘multipele correspondentieanalyse’). (zie Van der Burg, 1988, p. 96-97).

De CANALS-optie in TRANSREG staat voor canonische correlatieanalyse, maar is iets anders geprogrammeerd dan CANALS in Gifi (1981; 1990), zoals geprogrammeerd door Van der Burg. Het oorspronkelijke CANALS-algoritme itereert tussen twee stappen. In de eerste plaats moet de ALS-verliesfunctie geminimaliseerd worden over de parameters uit de eerste set - waarbij de orthogonaliteitsrestricties constant gehouden worden over de andere set - en vice versa. De SAS-versie, PROC TRANSREG, voert een gewone canonische correlatieanalyse uit en gebruikt vervolgens de resultaten van deze analyse om de afhankelijke variabele(n) te trans-formeren. Daarna voert TRANSREG een nieuwe, ordinaire canonische correlatieanalyse uit en gebruikt dan deze resultaten om alle onafhankelijke variabelen te transformeren. Het itereert hierover.
Het TRANSREG-CANALS algoritme is door Warren F. Kuhfeld (Statistical Research and Development, SAS Institute Inc., Cary, NC) in TRANSREG geïmplementeerd in Leiden, en is uitvoerig bediscussieerd met prof. J. de Leeuw en dr. E. van der Burg en door hen goedgekeurd (E-mails van A.W. Vogelesang met W. Kuhfeld d.d. 16-08-'96 en 04-01-'94). In de woorden van Warren Kuhfeld: “The outer algorithm of the original CANALS is slightly different than TRANSREG's”. Als er een oplossing is, dan moeten beide procedures die vinden. Volgens Kuhfeld is TRANSREG sneller en minder gevoelig voor abnormale data of specificatiefouten.

4. Beschrijving van de procedures

Voor het uitvoeren van een CANALS-analyse staan de volgende procedures ter beschikking:
a. CANALS-programma van de Vakgroep Datatheorie (Universiteit Leiden);
b. TRANSREG-procedure in SAS;
c. OVERALS-procedure in SPSS-CATEGORIES;
[d. OVERALS-programma van de Vakgroep Datatheorie (Universiteit Leiden)].

Zoals is uitgelegd in hoofdstuk 3, zijn de alternatieven *c* en *d* uitwisselbaar. Verder geldt dat OVERALS-SPSS moderner en gebruiksvriendelijker is dan de Leidse versie. De beschikbaarheid van OVERALS is beperkt tot VAX-VMS.

4.1. Procedure CANALS (Leiden)

CANALS kent de volgende datatransformaties: nominaal, ordinaal en numeriek. De bij dit onderzoek gebruikte data hebben betrekking op variabelen die in klassen zijn geordend, daarom is hier gekozen voor een ordinale transformatie.

Naast de schaling van de klassen per variabele geeft CANALS de volgende gegevens:
1. canonische correlatie voor iedere dimensie;
2. variabele gewichten per dimensie;
3. correlaties tussen de optimaal getransformeerde variabelen en de canonische assen van de eerste set voor elke dimensie;
4. correlaties tussen de optimaal getransformeerde variabelen en de canonische assen van de tweede set voor elke dimensie.

4.2. Procedure TRANSREG (SAS)

TRANSREG kent de volgende optimale transformaties:
- linear ('lin'): geeft een lineaire transformatie voor elke variabele.
- monotoon ('mon'): geeft een monotone transformatie voor elke variabele, met de restrictie dat 'ties' gelijk blijven.
- opscore ('ops'): geeft een optimale scoring voor elke variabele.
- untie ('unt'): zelfde als monotoon, echter zonder restricties.

Naast deze transformaties bevat TRANSREG nog andere klassen van transformaties: variable expansions (class, epoint) en 'non-optimal' transformaties (log, logit, exp). Deze zijn hier niet relevant. TRANSREG geeft naast de schaling van de klassen per variabele de *n* canonische variabelen, waar *n* de waarde is van de NCAN = *n* optie. Bij CANALS kan ook het aantal canonische variabelen opgegeven worden, maar alleen met gelijke schaling van de variabelen voor alle canonische assen.

4.3. **Procedure OVERALS (Leiden en SPSS-CATEGORIES)**

SPSS-CATEGORIES bevat de Leidse programma's ANACOR, HOMALS, PRINCALS en OVERALS. OVERALS is een procedure voor niet-lineaire canonische correlatie. Ook hier wordt geen intervalniveau verondersteld voor de data en worden geen lineaire relaties verondersteld tussen de variabelen. OVERALS bepaalt de gelijkenis tussen de sets door gelijktijdig lineaire combinaties van de variabelen in elke set te vergelijken met een 'onbekende' set: de objectscores. De 'objecten' zijn de individuen (rijen), de variabelen de categorieën.

SPSS-CATEGORIES kent de volgende optimale transformaties:

1. nume	-->numerical	lineaire transformatie van variabelen;
2. ordi	-->ordinal	monotone transformatie van variabelen;
3. snom	-->single nominal	objecten met dezelfde waarde voor een variabele krijgen dezelfde scoring op dimensie 1, 2, 3, ... (zoals bij CANALS en PRINCALS).
4. mnom	-->multiple nominal	objecten met dezelfde waarde voor een variabele kunnen verschillende scores krijgen op dimensie 1, 2, 3, ... (als bij HOMALS).

Een *set* is een groep variabelen en OVERALS bepaalt de relaties tussen de sets, niet tussen de variabelen (tenzij elke set uit één variabele bestaat). De *centroïden* zijn de gemiddelden van alle objecten die tot dezelfde categorie behoren - onafhankelijk van de andere variabelen. De centroïden worden geprojecteerd op een lijn door de oorsprong. De gewichten zijn regressie-gewichten.

4.4. **Vergelijking van CANALS en TRANSREG**

In deze paragraaf zullen de technieken voor canonische correlatieanalyse in CANALS en TRANSREG vergeleken worden aan de hand van een testbestand, dat afkomstig is uit een onderzoek naar ongevalstypen bij vierarmige rotondes (Maycock & Hall, 1984; zie Oppe, 1992).

4.4.1. *Het testbestand*

Het testbestand bestaat uit 21 variabelen: 20 kenmerken van vierarmige rotondes en één afhankelijke variabele, A1 ('entering/circulating accidents').

In totaal zijn er vijf afhankelijke variabelen (A1 - A5). Alleen de eerste, A1: 'Entering/circulating accidents', is gebruikt bij de vergelijking van CANALS en TRANSREG. De 20 verklarende variabelen zijn gecodeerd als V01 t/m V20. Gezien het feit dat in databestanden over het algemeen missing values voorkomen, zijn ook in het testbestand missing values opgenomen.

V01: JP Junction Period in month (max 72)

V02: T6 Roundabout category (1 - 6)

	30/40 mph	50/70 mph
- Small	cat. 1	cat. 2
- Conventional	cat. 3	cat. 4
- Dual Carriageway	cat. 5	cat. 6

V03: QPE Pedal cycle flow (entering)
V04: QME Motor cycle flow (entering)
V05: QE Total vehicle flow (entering)
V06: QPC Pedal cycle flow (circulating)
V07: QMC Motor cycle flow (circulating)
V08: QC Total vehicle flow (circulating
V09: QX Total vehicle flow (exiting)
V10: QPED Pedestrian crossing flow (measured in 16 hour flows)
V11: CE Entry curvature
V12: E Entry width
V13: V Approach width
V14: ANG Angle between arms
V15: G Gradient category, from severe downhill (-3) to uphill (+3)
V16: VR Visibility to the right
V17: ICD Inscribed Circle Diameter
V18: CID Central Island Diameter
V19: AC Approach Curve (just before roundabout)
V20: NDA Nearside Deflection Angle

Accident types:
V21: A1 Entering/circulating accidents
[V22: A2 Approaching accidents]
[V23: A3 Single vehicle accidents]
[V24: A4 Other accidents]
[V25: A5 Pedestrian accidents]

4.4.2. *Vergelijking van de verschillende opties*

Om het effect van de verschillende opties na te gaan, zijn voor TRANSREG de volgende analyses uitgevoerd, waarbij A1 de afhankelijke of criteriumvariabele is en de variabelen V01-V20 de voorspellers zijn:

1. ops_mon --> model ops(A1) = mon (V01, ... ,V20) /method = can ncan=1 nomiss etc.
2. mon_mon --> model mon(A1) = mon(V01, ... ,V20)
3. ops_ops --> model ops(A1) = ops (V01, ... ,V20)
4. linear --> model lin(A1) = mon(V01, ... ,V20)
5. untie --> model unt(A1) = mon(V01, ... ,V20)

Ad 1.
De *afhankelijke* variabele, A1, wordt bij **ops_mon** op nominaal niveau gebruikt, dat wil zeggen dat elke klasse van A1 een reëel getal krijgt toebedeeld als optimale schaalwaarde. De reële getallen ter identificatie van de klassen 1 t/m *n* hoeven niet geordend te zijn. De twintig *onafhankelijke* variabelen V01 t/m V20 worden *monotoon* geschaald. Aan de klassen worden dus reële getallen toegekend in oplopende of aflopende volgorde.

Ad 2.
Zowel de afhankelijke variabele, A1, als de onafhankelijke variabelen, V01 - V20, worden bij **mon_mon** monotoon geschaald, dat wil zeggen dat alle variabelen de klassen reële getallen krijgen in oplopende of aflopende volgorde.

Ad 3.
Bij **ops_ops** worden alle variabelen op nominaal niveau gebruikt, per variabele krijgt elke klasse dus een eigen optimale schaalwaarde.

Ad 4.
Bij de optie **linear** wordt de afhankelijke variabele als een numerieke variabele beschouwd, dat wil zeggen dat aan de klassen reële getallen toegewezen worden in oplopende of aflopende volgorde èn dat de *verschillen* tussen de opeenvolgende getallen betekenis hebben. Voor de onafhankelijke variabele geldt een monotone transformatie: getallen in oplopende volgorde, maar de verschillen tussen deze getallen hebben geen enkele betekenis.

Ad 5.
Bij de optie **untie** hoeven ‘ties’ (gelijke scores) niet als ‘gelijk’ beschouwd te worden. Bij de afhankelijke variabele kunnen categorieën met eenzelfde score na transformatie toch een verschillende optimale schaalwaarde krijgen. De afhankelijke variabelen worden weer monotoon geschaald.

5. Vergelijking van de verschillende procedures

Bij TRANSREG worden met betrekking tot de schaling van de variabelen default de gemiddelden per klasse weergegeven, maar met een speciale optie kunnen in plaats van de gemiddelden de *z*-scores van de variabelen worden opgevraagd. Bij het oorspronkelijke CANALS worden alléén de *z*-scores gegeven. De vergelijking van beide procedures is dan ook op basis van de *z*-scores per klasse per variabele. Bij de CANALS-analyse zijn, tenzij anders vermeld, de variabelen steeds als ordinaal opgegeven.

5.1. CANALS (ordinaal) versus TRANSREG (mon-mon)

De eerste analyse is uitgevoerd over het totale bestand, *inclusief* missing values. Voor TRANSREG is bij deze eerste vergelijking gekozen voor de optie '*mon*otonic' voor de afhankelijke variabele A1 èn de verklarende variabelen, voor CANALS een ordinale transformatie:

model monotonic (A1) = monotonic (V01, ... , V20)
/ method = CANALS ncan=1 (één canonische dimensie)

Er blijken duidelijke verschillen op te treden tussen CANALS en TRANSREG. De resultaten van deze analyse staan in *Bijlage 1a*. De schaling van de klassen van A1 is duidelijk verschillend in CANALS en TRANSREG. Bij CANALS blijkt de rangorde van de klassen niet te zijn veranderd na optimale transformatie. Het aantal klassen is gereduceerd van 11 naar 6, waarbij samenvoeging heeft plaatsgevonden van de klassen 1-2-3, 5-6, 7-8 en 10-11. Bij TRANSREG daarentegen geeft de rangorde na transformatie van de klassen een duidelijk ander beeld te zien; zo heeft klasse 11 hier de hoogste score gekregen (8), klasse 10 heeft score 7. Ook de schaling van de onafhankelijke variabelen geeft verschillen te zien tussen de twee procedures. Echt afwijkend zijn de variabelen V02, V06, V09, V17 en V19; wel goed overeen komen de variabelen V01, V03, V11, V12 en V16. De afhankelijke variabele, A1, komt matig overeen: CANALS maakt meer 'ties'.

5.2. CANALS (ordinaal) versus TRANSREG (ops-mon)

De tweede analyse is uitgevoerd over hetzelfde bestand. Voor TRANSREG is echter bij de tweede vergelijking voor de afhankelijke variabele A1 de optie 'opscore' genomen, voor de rest blijft alles gelijk:

model opscore (A1) = monotonic (V01, ... , V20)
/ method = CANALS ncan=1

De resultaten van deze analyses staan in *Bijlage 1b*. Voor CANALS geldt dat de schaling van alle variabelen exact gelijk blijft. De schaling in TRANSREG is sterk veranderd: 'ties' worden gebroken en voor A1 ('optimaal geschaald') is de ordening is nu verre van monotoon. TRANSREG maakt onder de ops-mon-transformatie voor bijvoorbeeld variabele V01 één categorie meer dan onder de mon-mon-transformatie, in andere gevallen één minder. Het lijkt erop dat CANALS meer 'ties' maakt. Echt afwijkend zijn de variabelen V02, V06, V09, V17 en V19; wel goed overeen komen de variabelen V03, V11, V12 en V16. De afhankelijke variabele, A1, komt slecht overeen voor de beide procedures. TRANSREG

maakt in het algemeen meer categorieën en een wezenlijke andere rangorde, waar CANALS gelijken maakt.

Mogelijk zijn de afwijkingen veroorzaakt door een verschillende behandeling van missing values binnen de procedures. Derhalve zijn nieuwe analyses uitgevoerd, maar nu exclusief de missing values (zie *Bijlage 2*), voor CANALS ordinaal/nominaal versus TRANSREG met diverse opties (zie § 5.3).

5.3. CANALS versus TRANSREG zonder missings voor diverse opties

Er zijn zeven nieuwe analyses uitgevoerd over het testbestand, exclusief de missing data. Voor de procedure CANALS zijn dat de opties ordinaal en nominaal en voor TRANSREG de opties :

1. model opscore (A1) = monotonic (V01,..,V20)
2. model monotonic (A1) = monotonic (V01,...,V20)
3. model opscore (A1) = opscore (V01,...,V20)
4. model linear (A1) = monotonic (V01,...,V20)
5. model untie (A1) = monotonic (V01,...,V20)

Als de cases met missings worden geëlimineerd, via ‘listwise deletion’, worden de variabelen aanmerkelijk beter gescoord (zie *Bijlage 2,* kolom 3: de transformatie ops-mon). De betere resultaten zijn te vinden in *Bijlage 2* onder ‘TRANSREG-mon-mon’ en onder ‘SPSS’ en worden hierna uitgebreid besproken. Echter, het samenvoegen van klassen (tot ties) is niet gelijk voor de verschillende opties, ops-mon, mon-mon en SPSS; dit wordt hieronder uitgebreid besproken.

De resultaten van deze analyses en de schaling van de variabelen onder de verschillende opties, staan vermeld in *Bijlage 2*. Bekijken we de resultaten van deze analyses dan blijkt dat onder TRANSREG de optie ‘model mon (A1) = mon (V01, ... , V20)’ de grootste overeenkomst vertoont met CANALS. De schaling van de afhankelijke variabele A1 komt nagenoeg overeen in beide programma's; afwijkend zijn de klassen 9 en 10. Bij CANALS zijn deze twee klassen gescheiden, bij TRANSREG vallen ze samen. Van de verklarende variabelen zijn in lichte mate afwijkend de variabelen V07, V15 en V16, de variabele V06 geeft het grootste verschil te zien tussen beide procedures; de overige variabelen komen wat schaling betreft keurig overeen. Ook de optie ‘model ops(A1) = mon(V01, ... , V20)’ geeft overeenkomstige resultaten in CANALS en TRANSREG, maar geeft toch meer verschillen te zien dan de optie ‘model mon(A1) = mon(V01, ... , V20)’. In lichte mate betreft dat de variabelen V06, V09, V10, V13, V14 en V20; sterk afwijkend zijn de afhankelijke variabele A1 en de verklarende variabelen V08, V15, V19 en V20.

Conclusie 1: CANALS ordinaal ≈ TRANSREG mon-mon

In principe leveren CANALS (ordinaal) en TRANSREG met de optie ‘model mon(A1) = mon(V01, ... , V20)’ - exclusief de missings values - praktisch hetzelfde resultaat en geven geen verschil in interpretatie. Zoals te verwachten was, wijken de resultaten verkregen bij de overige transformaties (lin, ops, untie) in TRANSREG in veel sterkere mate af van CANALS, deze kunnen daarom niet als een goede benadering voor CANALS worden beschouwd.

5.4. **CANALS nominaal**

Voor CANALS is naast de optie ‘ordinaal’ ook een analyse uitgevoerd met de optie ‘nominaal’ en exclusief de missing values. Verondersteld wordt dat de optie ‘nominaal’ onder CANALS overeenkomt met de optie ‘model opscore (A1) = opscore (V01.....V20)’ onder TRANSREG. Deze vergelijking is uitgevoerd, de resultaten (zie *Bijlage 2*) geven echter voor een aantal variabelen een duidelijk verschil in schaling te zien, waaruit volgt dat de veronderstelling niet juist is. *In feite* is dit resultaat verrassend, aangezien verwacht zou kunnen worden dat optimale schaling de beste resultaten zou geven. Om na te gaan of er bij de verschillen tussen de procedures mogelijk sprake zou kunnen zijn van een constante factor, zijn - naast de rechtstreekse vergelijking van de schaling van de klassen per variabele - ook per variabele per klasse de quotiënten berekend van CANALS met de verschillende opties van TRANSREG. De resultaten hiervan zijn vermeld in *Bijlage 3*. Uit vergelijking van deze resultaten blijkt dat de hypothese van een mogelijke aanwezigheid van een constante factor moet worden verworpen.

Conclusie 2: CANALS nominaal ≠ TRANSREG ops-mon

In principe verwachten we dat CANALS (nominaal) en TRANSREG met de optie ‘model ops(A1) = ops(V01, ... , V20)’ praktisch hetzelfde resultaat zullen geven. Dit blijkt maar gedeeltelijk het geval te zijn (zie *Bijlage 2*). Dit is mogelijk door een grotere schalingsvrijheid voor deze optie en is een meer toevallig effect. De resultaten voor de overige transformaties (lin, ops, untie) in TRANSREG wijken nog sterker af van CANALS.

Zoals reeds opgemerkt geeft TRANSREG naast de schaling van de klassen per variabele de canonische coëfficiënten per variabele. Nagegaan is of er verband kon worden gevonden tussen de TRANSREG-coëfficiënten enerzijds met de overige gegevens die door CANALS worden meegegeven, te weten:

a. de ‘variable weights for each dimension’;
b. de ‘correlations between the optimally scaled variables’ en de ‘canonical variates of the first set for each dimension’;
c. de ‘correlations between the optimally scaled variables’ en de ‘canonical variates of the second set for each dimension’.

De TRANSREG-coëfficiënten blijken het meest overeen te komen met de ‘variable weights for each dimension’ uit CANALS. Een duidelijke relatie tussen deze gegevens is echter niet gevonden.

5.5. **CANALS versus OVERALS (SPSS)**

Naast de vergelijkingen tussen CANALS en TRANSREG is ook een CANALS-analyse uitgevoerd met OVERALS-SPSS, omdat verwacht kan worden dat dit hetzelfde oplevert. Vergelijking van de schaling onder OVERALS-SPSS met de schaling onder CANALS (ordinaal) laat naast een aantal goed overeenkomende variabelen ook duidelijke verschillen zien. Goed overeen komen de schalingen van de variabelen V01, V02, V04, V12, V18 en V19. De schaling van de afhankelijke variabele wijkt enigszins af, en ook de schalingen van de variabelen V03, V05, V08, V09, V10, V11, V13, V16, V17 en V20 wijken in lichte mate af. Sterk afwijkend zijn de

variabelen V06, V07 en V14. De procedure onder SPSS vertoont minder overeenkomst met CANALS dan de procedures mon_mon en ops_mon onder TRANSREG. Resultaten verkregen met SPSS zullen tot een iets andere interpretatie Leiden dan de resultaten verkregen onder CANALS.

Conclusie 3: CANALS ordinaal ≠ SPSS

De procedure onder SPSS vertoont iets minder overeenkomst met CANALS (ordinaal) dan de procedures mon_mon onder TRANSREG. De resultaten verkregen met SPSS zullen tot een iets andere interpretatie leiden dan de resultaten verkregen met CANALS. Evengoed is SPSS een goed alternatief voor TRANSREG (in ieder geval voor deze data).

6. Resultaten

Bij CANALS worden naast de schaling van de klassen per variabele nog de volgende gegevens geleverd (zie ook § 4.1):

1. de ‘canonical correlation for each dimension’;
2. de ‘variable weights for each dimension’;
3. de ‘correlations between the optimally scaled variables’ en de ‘canonical variates of the first set for each dimension’;
4. de ‘correlations between the optimally scaled variables’en de ‘canonical variates of the second set for each dimension’.

TRANSREG levert naast de schaling van de variabelen nog een extra coëfficiënt per variabele; wat deze coëfficiënt werkelijk betekent is op dit moment echter nog niet duidelijk. Ook het ontbreken van de canonische correlaties kan als een gemis worden gezien.

Missing values
Indien de dataset missing values bevat, en we willen deze niet meenemen in de analyse, dan moet bij CANALS een nieuw bestand worden aangemaakt en ook het programma (Leidcan.des) moet worden aangepast. Bij TRANSREG kan dit eenvoudig worden opgegeven met de optie ‘nomiss’.

Variabelenamen
Bij TRANSREG en OVERALS-SPSS kan gebruik gemaakt worden van bestaande variabelenamen; deze worden toegekend tijdens het maken van de SAS- of SPSS-dataset (zie § 7.2 en § 7.3). Bij CANALS daarentegen is dit niet mogelijk, dit werkt alleen met variabele-nummers, hetgeen vooral bij een wat groter aantal variabelen nogal lastig is.

7. Voorbeelden van de verschillende procedures

7.1. Setup CANALS-procedure

```
    1
CANALS-ANALYSE; Heli-Project  ===>  met york1 york2;
  125    1   15   76
    1   20
    8    0    0    1    0    1    0    0    1    0
   50   10    6    2   13    2    2    7    7   13   6    6    6    6    6   76
    1    0    0    0    1    0    0    1    1    1   1    1    1    1    1    1
(4X,I4,6I3,12X,4I3,3X,5I3)
```

regel [geen regelnummers in tekst meegeven; format 16F5.0; regel 2 is free format]

```
1: · · · · 1
2:CANALS - analyse
3  · · · 312   20    1   11
4: · · · · 2   20    0    0
5: · · · · 8    0    0    1   0   1   0   1   1   0
6: · · · · 6    6    6    8   7   8   8   9   7   7   7   5   5   5   7   5   7
   · · · · 8    7    6   11
7: · · · · 1    1    1    1   1   1   1   1   1   1   1   1   1   1   1   1   1
   · · · · 1    1    1    1
8: · · · · 0
9/10/11.......
(6X,21I3,12X)
```

Toelichting

1 1 = aantal analyses
2 titel voor de uitvoer-tabel
3 312 = aantal records (individuen)
20 = aantal onafhankelijke variabelen
1 = aantal afhankelijke variabelen
11 = hoogste aantal klassen (maximum van regel 6)
4 2 = aantal dimensies
20 = aantal iteraties (default: 20)
0 = minimum stress difference (default: 0,001)
0 = maximum weight increase (default: 0,0001)
5 8 = standaard voor VAX
0 = print de eerste 10 variabelen (1 = allemaal)
0 = geen output optimally scaled variables
1 = print output of stress (0=no print)
0 = no extra output
1 = print canonical scores (0=no print)
0 = no plot (1=one or more plots specified in line 8)
1 = plot canonical scores (0=no plot)
1 = print projections category quantifications (0=no print)
0 = no output quantifications missing data (1=print output)
6 aantal klassen per variabele
7 0 = nominaal, 1 = ordinaal, 2 = numeriek
8 0 = no plot, 1 = plot category quantifications for each variable
9 Fortran-input (moet 3 regels zijn, dus hier 2 blanco regels toevoegen (10,11)).

Het RUNnen van dit programma gaat als volgt:
prompt> @runals.com <enter>
er verschijnen nu de volgende vragen:
1. filename programma ---> CANALS;
2. filename Deck Setup ---> Leidcan.des (zelf te kiezen naam);
3. filename data matrix ---> test.dat (de ASCII-dataset);
4. filename van output ---> Leidcan.lis (uitvoer resultaat).

7.2. Setup TRANSREG-procedure

```
options     nocenter;
libname     owner base '[]';
proc        TRANSREG data = owner.test;
            model ops(A1) = mon (V01  V02 ..... V19  V20)
            / methode=can   ncan=1   tst=z   nomiss;
            output out=dump   coe;
proc        print ; where _name_ = 'CAN1';

data        ;
            set dump;
            if _type_  = 'SCORE';

proc means  mean; class A1  ; var tA1  ;
proc means  mean; class V01; var tV01  ;
        ,,,
        ,,,
proc means  mean; class V19; var tV19  ;
proc means  mean; class V20; var tV20 ;
proc print; where _type_ ne 'score';
```

Toelichting
Na opgave van het model kunnen onder TRANSREG de volgende opties worden opgegeven (na de slash):

De gewenste methode:
met(hode) = can(als)
met(hode) = mor(als)
met(hode) = red(udancy)
met(hode) = uni(variate)

ncan = 1: aantal canonische variabelen in de analyse

tst = z: Bepaalt hoe gemiddelde en variantie moeten worden getransformeerd in de output-dataset. Default worden de gemiddelden per klasse weergegeven.
Met de optie tst = z worden de z-scores weergegeven; dit is een transformatie naar gemiddelde = 0 en variantie = 1. (CANALS doet dit standaard, heeft geen andere mogelijkheid)

nomiss: laat observaties met missing values weg uit de analyse.

7.3. Setup OVERALS-(SPSS-)procedure

Hieronder wordt de OVERALS-SPSS-analyse gegeven voor de data uit de vorige paragraaf. Voor alle analyses (OVERALS-, CANALS- en SAS-analyses) zijn de variabelen gehercodeerd tot V01 - V21.

De ‘variable labels’ bevatten de oorspronkelijke namen.

```
DATA LIST FREE/          A1 V01 V02 V03 V04 V05 V06 V07 V08 V09 V10 V11
                         V12 V13 V14 V15 V16 V17 V18 V19 V20 .
VARIABLE LABELS          A1  ‘A1'
                         V01  ‘JP’
                         V02  ‘T6'
                         V03  ‘QPE’
                         V04  ‘QME’
                         V05  ‘QE’
                         V06  ‘QPC’
                         V07  ‘QMC’
                         V08  ‘QC’
                         V09  ‘QX’
                         V10  ‘QPED’
                         V11  ‘CE’
                         V12  ‘E’
                         V13  ‘V’
                         V14  ‘ANG’
                         V15  ‘G’
                         VR6 ‘VR’
                         V17 ‘ICD’
                         V18  ‘CID’
                         V19  ‘AC’
                         V20  ‘NDA’

OVERALS VARIABLES =      A1 (6) V01 (6) V02 (6) V03 (8) V04 (7) V05 (8)
                         V06 (8) V07 (9) V08(7) V09 (7) V10 (7)
                         V11 (5) V12 (5) V13 (5) V14 (7) V15 (5)
                         V16 (7) V17 (8) V18 (7) V19 (6) V20 (11)
                         / DATA = OVERALS.DAT
                         / ANALYSIS = A1 TO NDA (ORDI)
                         / SETS = 2(1,20)  / INITIAL = NUMERICAL
                         / PRINT= FREQ HIST QUANT CENTROID WEIGHTS
                         OBJECT FIT
                         / PLOT = OBJECT LOADINGS (4)  TRANS (4)
                         CENTROIDS (A1 TO NDA)  (4)
```

Literatuur

Benzécri, J.P. (1973). *L' Analyse des données*, Vol. 1, 2. Dunod, Paris.

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

Burg, E. van der (1983). *CANALS User's Guide*. Department of Datatheory, Faculty of Social Sciences, University of Leiden.

Burg, E. van der (1988). *Nonlinear canonical correlation and some related techniques*. Thesis. University of Leiden.

Burg, E. van der, Leeuw, J. de, & Verdegaal, R. (1988). *Homogeneity analysis with k sets of variables: An alternating least squares method with optimal scaling features*. Psychometrika, 53, pp. 177- 197.

Fisher, R.A. (1940). *The precision of discriminant functions.* Annals of Eugenics, 10, pp. 422-429.

Gifi, A. (1981). *Nonlinear multivariate analysis*. Department of Datatheory, Faculty of Social Sciences, University of Leiden.

Gifi, A. (1990). *Nonlinear multivariate analysis*. Wiley, New York.

Goodman, L.A. (1981). *Association models and canonical correlation in the analysis of cross-classifications having ordered categories*, JASA, 76, pp. 320-334.

Guttman, L. (1941). *The quantification of a class of attributes: A theory and method of scale construction.* In: Horst, P. et al. (eds.) (1941). *The prediction of personal adjustment*. The Social Science Research Council, Bulletin No. 48, pp. 319-348.

Guttman, L. (1946). *An approach for quantifying paired comparisons and rank order*. Ann. Math. Statist. 17, pp. 144-163.

Hill, M.O. (1974). *Correspondence analysis: a neglected multivariate method*. Journal of the Royal Statistical Society, Series C, 23, pp. 340-354.

Hirschfeld, H.O. (1935). *A connection between correlation and contingency*. Proceedings Cambridge Philosophical Society, 31, pp. 520-524.

Hotelling, H. (1936). *Relations between two sets of variates*. Biometrika, 28, pp. 321-377.

Kuhfeld, W.F. (1985). *Principal components of ordered categorical data*. Unpublished Doctoral Dissertation, University of Northern Carolina.

Kuhfeld, W. F., Sarle, W.S., & Young, F.W. (1985). *Methods in generating model estimates in the PRINQUAL macro.* In: SUGI- Proceedings, 10, pp. 962-971. SAS Institute, Cary, NC:.

Kuhfeld, W. F., Young, F.W., & Kent, D.P. (1987). *New developments in psychometric and market research procedures.* In: SUGI-Proceedings, 12, 1101-1106. SAS Institute, Cary, NC.

Leeuw, J. de (1984). *Canonical analysis of categorical data*. Thesis. University Leiden.

Nishishato, S. (1978). *Optimal scaling of paired comparison and rank order data: an alternative to Guttman's formulation*. Psychometrika, 43, pp. 263-271.

Oppe, S. (1992). *A comparison of some statistical techniques for road accident analysis*. Accident Analysis & Prevention, 24, pp. 397-423.

SPSS (1990). *Nonlinear canonical correlation analysis: Procedure OVERALS.* SPSS CATEGORIES, B80 - B94. SPSS Inc. ISBN 0-918469-93-7. Chicago.

Verdegaal, R. (1986). *OVERALS* (User's guide UG-86-01). Vakgroep Datatheorie, Faculteit Sociale Wetenschappen, Rijksuniversiteit Leiden.

Bijlage 1 t/m 4