ANALYSE VAN SAMENHANGEN TUSSEN KWALITATIEVE VERKEERSVEILIGHEIDS-KENMERKEN

Artikel Verkeerskunde 31 (1980) 7: 364 t/m 368
Artikel Verkeerskunde 31 (1980) 12: 629 t/m 631

R-80-20
Drs. S. Oppe
Voorburg, 1980
Stichting Wetenschappelijk Onderzoek Verkeersveiligheid SWOV

SAMENVATTING

De auteur merkt op dat er een toegenomen interesse is in modellen voor kwalitatieve gegevens.
Deze toename in interesse is er ook m.b.t. verkeersveiligheids-onderzoek. Er zijn twee belangrijke ontwikkelingen te noemen.
Ten eerste is er een doorbraak te constateren t.a.v. de bijna klassieke manier van het analyseren van kwalitatieve gegevens, t.w. de analyse van kruistabellen. De belangrijkste reden voor de ontwikkeling ligt in de presentatie van de gegevens d.m.v. het zgn. log-lineaire model. Deze presentatie heeft tot gevolg dat de analyse van kruistabellen analoog is aan het toepassen van "variantieanalyse" op geclassificeerde kwantitatieve gegevens.
Ten tweede vinden we generaliseringen van technieken welke van oorsprong alleen toepasbaar zijn op kwantitatieve gegevens zoals principale componentanalyse (factoranalyse) en canonische analyse, die deze technieken ook geschikt maken voor analyse van kwalitatieve gegevens.
In het eerste van twee artikelen beschrijft de auteur deze gegeneraliseerde technieken en geeft hij een voorbeeld van de toepassing op verkeersveiligheidsgegevens.
In een tweede artikel volgt een beschrijving van de belangrijkste aspecten van het log-lineaire model en de toepasbaarheid ervan op verkeersveiligheidsgegevens.

SUMMARY

Analysis of relations between qualitative traffic safety characteristics

The author notices an increased interest in models for qualitative data. This increase in interest is also found with regard to the analysis of traffic safety data. There are two major developments that can be mentioned.
First of all, a break through is found in the (almost) classicial

way of analysing qualitative data, i.e., the analysis of contingency tables. The main reason for this development is the representation of the data by means of the so-called log-linear model. This representation results in an analysis of contingency tables similar to the "analysis of variance" for classified quantitative data.
Secondly, we find generalisations of techniques like principal component analysis (factor analysis) and canonical analysis, that make these techniques applicable to qualitative data.
In this first of two articles, the author describes these generalised techniques and gives an example of the application to traffic safety data. In a second article he will give a description of the essential aspects of the log-linear model and its applicability to traffic safety data.

VOORWOORD

Aanleiding tot het schrijven van dit artikel is de brede belangstelling die de beschreven analysemethoden beginnen te krijgen, ook op het gebied van de verkeersveiligheid. Deze belangstelling blijkt ondermeer uit de bijeenkomst van het NATO-Advanced Study Institute gehouden in Urbino, Italie, over: "Contingency table analysis of road safety research", waarvan de Proceedings - waarin een bijdrage van drs. S. Oppe - binnenkort door Sijthoff en Noordhoff International worden gepubliceerd, en de in april 1980 gehouden studieweek van de afdeling Datatheorie van de R.U.-Leiden, onder de titel: "Niet-lineaire multivariate analyse".
De auteur houdt zich op dit moment bezig met het onderzoek naar de invloed van kleine steekproefaantallen op de resultaten van log-liniaire analyses. Voorts wordt door de afdeling Datatheorie in samenwerking met de SWOV gezocht naar uitbreidingen van de methode welke voor het verkeersveiligheidsonderzoek van belang zijn.
In een vervolg op dit artikel zal worden ingegaan op de analyse van kruistabellen.

1. INLEIDING

Bij de analyse van onderzoekgegevens is de keuze van het type analysemodel behalve van de theorievorming op het gebied van het onderzoek zelf, voornamelijk afhankelijk van de aard van de verzamelde gegevens en de daarmee samenhangende mogelijkheden tot verwerking.
We onderscheiden kwalitatieve gegevens, ook wel categorische gegevens genoemd en kwantitatieve gegevens, ook wel numerieke gegevens genoemd.
Voor elk van de twee typen gegevens gebruiken we doorgaans geheel verschillende methoden voor analyse. Dit verschil in keuze voor een analysemethode is terug te voeren op het verschil in 'meetniveau'.
Onderzoeken we de samenhang tussen kwalitatieve kenmerken dan kiezen we haast automatisch voor een kruistabelanalyse, resulterend in een Chi-kwadraattoets voor samenhang tussen twee kenmerken. We hanteren hierbij een geheel ander type verklaringsmodel dan bij kwantitatieve gegevens. Willen we de samenhang van twee kwantitatieve kenmerken vaststellen, dan gebeurt dit doorgaans met behulp van een lineaire regressie-analyse of het berekenen van de correlatiecoëfficient.
De situatie wordt gecompliceerder als van meer dan twee kenmerken sprake is.
Bij kwantitatieve gegevens zijn er diverse technieken die kunnen worden gebruikt om samenhangen tussen meerdere kenmerken of variabelen tegelijk te onderzoeken. Voorbeelden hiervan zijn: multiple lineaire regressie-analyse, variantie-analyse, factoranalyse en canonische analyse.
Als we te maken hebben met meerdere kwalitatieve kenmerken, dan bestaat de analyse van de samenhang tussen deze gegevens in de meeste gevallen uit het analyseren van kruistabellen, waarin telkens twee kenmerken op hun samenhang worden onderzocht. Wordt invloed van een ander kenmerk op de samenhang verondersteld, dan wordt de samenhang bestudeerd onder de verschillende condities van dat kenmerk. Dit alles resulteert in soms gigantische hoeveelheden kruistabellen en bijbehorende analyses, waarvan de onderlinge samenhang haast niet is aan te geven, zodat een fragmentarisch beeld van de onderzoekgegevens resulteert.
Dit is een van de redenen waarom aan de analyse van kwalitatieve ge-

gevens lang een betrekkelijke waarde is toegekend.
De laatste tijd hebben er echter een aantal ontwikkelingen plaatsgevonden, die de analyse van kwalitatieve gegevens sterk in de belangstelling hebben gebracht.
Allereerst heeft de analyse van kruistabellen meer aandacht gekregen. De oorzaak hiervan is vooral gelegen in een andere presentatie van het achterliggende model. Door in het model de logaritme van de celaantallen te beschrijven in plaats van de celaantallen zelf, ontstaat er een modelbeschrijving die geheel analoog is aan de modelbeschrijving die bijvoorbeeld bij variantie-analyse of lineaire regressie wordt gebruikt. Het hierbij gebruikte lineaire model heeft een aantal eigenschappen die de statistische beschrijving van de onderzoekgegevens vergemakkelijkt. Hierbij is het mogelijk in de analyse van een aantal kenmerken gezamenlijk een duidelijke scheiding te maken tussen de invloed van de diverse kenmerken en hun onderlinge samenhangen of interacties.
Eenzelfde beschrijving is nu door de toepassing van het lineaire model op de logaritmen van de celaantallen ook mogelijk voor kruistabellen. Het hier gebruikte model, dat vanwege de noodzakelijke logaritme-transformatie van de celwaarden wel het 'log-lineaire model' wordt genoemd, beschrijft de logaritme van de celwaarden als een som van parameters die corresponderen met de kenmerken en de onderlinge interacties tussen de kenmerken. Dit maakt het direct mogelijk de analyse ook toe te passen op kruistabellen uitgesplitst naar meer dan twee kenmerken. Daarnaast is het mogelijk ook uitspraken te doen over de bijdragen die specifieke klassen van kenmerken of combinaties ervan leveren aan geconstateerde samenhangen tussen kenmerken.
Naast deze ontwikkeling in de analyse van kruistabellen, waarop we later verder zullen ingaan, is er een ontwikkeling te constateren vanuit de andere tak van analyses, nl. de analyses die worden toegepast bij kwantitatieve gegevens.
Oorspronkelijk waren analyses als factoranalyse en canonische analyses uitsluitend toepasbaar op kwantitatieve gegevens. Nu zijn er generaliseringen van deze technieken beschikbaar gekomen waardoor de technieken ook toepasbaar zijn op kwalitatieve gegevens, met andere woorden op gegevens met een laag 'meetniveau'.

Voor we beginnen met een beschrijving hiervan, zullen we eerst een toelichting geven op het begrip 'meetniveau' en daarna wat nader ingaan op het factoranalysemodel en de canonische analyses.

2. HET MEETNIVEAU

Objecten, die worden onderzocht op een aantal kenmerken die voor het onderzoek van belang zijn, worden voor elk kenmerk geclassificeerd. Als we de term 'meten' in een ruime zin opvatten, kunnen we elke vorm van classificatie opvatten als meten. Meten van een object naar een bepaalde kenmerk is dan het indelen van dat object in een klasse van dat kenmerk, volgens een bepaald voorschrift. Tot zover is er geen verschil tussen kwalitatieve en kwantitatieve gegevens. Dit verschil komt naar voren als de onderlinge relaties tussen de diverse klassen van een kenmerk nader worden onderzocht. Wanneer het een kenmerk als 'wijze van verkeersdeelname' betreft, met de klassen: vrachtauto - personenauto - bromfiets - fiets - voetganger, dan zal de volgorde waarin deze klassen worden genoemd meestal arbitrair worden gevonden. Bij de beschrijving van de meetgegevens zal er dan van worden uitgegaan dat de gegevens op het laagste meetniveau zijn gemeten. De klassevolgorde is arbitrair en de klassewaarde, het getal dat aan elk van de klassen wordt toegekend, kan slechts als naam voor de klasse gelden. Vandaar dat wel wordt gesproken van een 'nominaal meetniveau'. Vaak is de klassewaarde meer dan alleen maar een naam. Ze geeft bijvoorbeeld iets weer van de mate waarin een object een bepaald kenmerk heeft. Hoe groter de klassewaarde, hoe meer de objecten in die klasse een bepaald kenmerk bezitten. In zo'n geval, waarin de klassen geordend zijn naar de mate waarin een kenmerk aanwezig is of ontbreekt, spreken we van een 'ordinaal meetniveau'. De volgorde van de klassen van het bovengenoemde kenmerk 'wijze van verkeersdeelname', kan worden opgevat als lopend van meer naar minder 'agressief' in het verkeer. Als dit aspect voor een bepaalde vraagstelling van belang is, zullen we uitgaan van een ordinaal meetniveau. Geldt een nog sterkere relatie tussen de klassen, zodanig dat gelijke verschillen tussen klassewaarden corresponderen met gelijke verschillen in de mate waarin een kenmerk aanwezig is, dan spreken we van meten op 'intervalniveau'. Geldt dit laatste niet alleen voor verschillen maar ook voor verhoudingen, dan spreekt men van meten op 'rationiveau'. Voor deze laatste vorm van meten is een absoluut

nulpunt noodzakelijk. Dit is bijvoorbeeld het geval bij het meten van lengte: nul meter is gelijk aan nul yards en correspondeert met het niet bezitten van lengte. Bij het meten van temperatuur in graden Celsius of Fahrenheit is dit niet het geval: nul graden Celsius is ongelijk aan nul graden Fahrenheit. Bij het nulpunt behoort een arbitrair gekozen temperatuur. Hier hebben we te maken met meten op intervalniveau of op een intervalschaal.
Als we spreken over kwalitatieve gegevens dan bedoelen we gegevens die op nominaal niveau gemeten zijn.
Bij analysetechnieken als de klassieke factoranalyse wordt minimaal verondersteld dat de gegevens zijn gemeten op intervalniveau. De techniek kan dus niet zomaar worden toegepast op ordinale en nominale gegevens. Zijn we echter in staat de ordinale gegevens zodanig te 'herschalen' dat de nieuwe klassewaarden voldoen aan de eisen van een intervalschaal, dan kunnen ook deze kenmerken worden opgenomen in de analyse. Voor nominale gegevens lijkt een dergelijke uitweg niet mogelijk.

3. FACTORANALYSE

Laten we eerst nagaan in welke situaties factoranalyse kan worden toegepast.
De probleemstelling die geleid heeft tot de ontwikkeling van factoranalyse is hierbij illustratief.
Uitgaande van de hypothese dat mensen het kenmerk 'algemene intelligentie' in meerdere of mindere mate bezitten, komen we tot de vraag hoe dit kenmerk moet worden gemeten. Direct meten lijkt niet mogelijk. Wel kan met behulp van diverse tests, een aantal vaardigheden en kundigheden worden gemeten die alle een veronderstelde samenhang hebben met de 'algemene intelligentie'. Elke test meet dus naast een specifieke vaardigheid of kundigheid iets van de algemene intelligentie. Deze algemene intelligentie wordt door de diverse tests gemeenschappelijk gemeten. De hypothese is nu: als we dit gemeenschappelijke aspect uit de diverse tests weten te scheiden van de specifieke effecten, dan hebben we te maken met de factor 'algemene intelligentie'. De methode van factoranalyse wordt nu gebruikt om deze algemene factor op te sporen en tevens de objecten (hier dus mensen) te meten (classificeren) naar hun algemene intelligentie. Soms zoeken we naar meerdere algemene factoren (bijv. Verbaal IQ, Performaal IQ etc.)
In het algemeen kunnen we stellen dat bij factoranalyse met behulp van diverse specifieke kenmerken wordt gezocht naar eraan ten grondslag liggende basiskenmerken. Nog algemener geldt dat bij factoranalyse wordt gezocht naar een beschrijving van de voor het onderzoek relevante informatie, die wordt geleverd door de classificatie van de onderzoekobjecten naar veel kenmerken door middel van een classificatie naar weinig basiskenmerken, welke basiskenmerken op zich niet te meten zijn.
Men spreekt in dergelijke gevallen vaak van factoren of ook wel van latente trekken. Latente trekken onderscheiden zich van manifeste variabelen doordat ze niet direct gemeten kunnen worden. De score op zo'n factor is te beschrijven als een lineaire combinatie van de scores op de gemeten kenmerken of, anders gezegd, als een gewogen som van deze. Het gewicht dat aan elk kenmerk wordt toegekend, geeft aan in hoeverre dit kenmerk samenhangt met de betreffende factor.

4. CANONISCHE ANALYSES

Als we te maken hebben met twee groepen van objecten, dan zouden we kunnen nagaan wat de overeenkomst tussen de groepen is en dus via factoranalyse kunnen zoeken naar factoren die aangeven hoe homogeen de kenmerken zijn. Vaak ook echter zijn we geinteresseerd in verschillen. We zoeken dan een factor die zo goed mogelijk discrimineert tussen beide groepen. Antwoord op deze vraag geeft de discriminantanalyse. Hierbij wordt gezocht naar de factor waarvoor geldt dat de scores van de objecten op die factor zo goed mogelijk corresponderen met de groepsindeling. Vanuit de score op deze factor kunnen we dan voorspellen tot welke groep een object behoort. Zo kunnen we ons bijvoorbeeld afvragen welke factoren de nachtongevallen van de dagongevallen onderscheiden, om zodoende bijvoorbeeld locaties waar het 's nachts potentieel gevaarlijk is op te sporen. Zijn er meer dan twee groepen dan spreken we van canonische-discriminantanalyse.
De groepsindeling kunnen we zien als een extra kenmerk waarop de objecten worden gecategoriseerd. Een kenmerk echter dat op nominaal niveau is gemeten. Zouden we de objecten niet in groepen indelen, maar bijvoorbeeld op intervalniveau meten op het betreffende kenmerk, dan kunnen we de vraagstelling veranderen in: welke factoren geven een zo goed mogelijke 'voorspelling' van het kenmerk waarin we geintereseerd zijn. In dat geval spreken we van multiple regressieanalyse. Er wordt dan gezocht naar die lineaire combinatie van (metrische) kenmerken, ook wel onafhankelijke variabelen genoemd, die een maximale voorspelling geeft van een bepaalde criteriumvariabele, ook wel afhankelijke variabele genoemd.
We kunnen (canonische) discriminantanalyse dus opvatten als een bijzonder geval van multiple lineaire regressie (MLR). Namelijk: MLR met één nominale criteriumvariabele. Als we bij MLR niet met één enkele afhankelijke variabele te maken hebben, maar met meerdere metrische criteriumvariabelen, dan spreken we van canonische correlatie-analyse. We zoeken dan naar die factor (of factoren) van de groep van verklarende variabelen die een maximale voorspelling geeft van (of maximale correlatie heeft met) de groep van criteriumvariabelen. Het doel is dus het vinden van een beschrijving van de samen-

hang tussen twee groepen van kenmerken door middel van die factoren van elke groep die deze samenhang zo goed mogelijk weergeven.
Het gaat er bij alle bovengenoemde methoden om lineaire combinaties te vinden van metrische variabelen zodanig, dat bepaalde relaties optimaal beschreven worden.
Bij factoranalyse gaat het dan om de relaties binnen een groep van variabelen, bij MLR en (canonische) discriminantanalyse om de relaties binnen een groep ten aanzien van een bepaald kenmerk of een groepsindeling van objecten, bij canonische analyse om de relaties binnen een groep van kenmerken ten opzichte van die van een andere groep van kenmerken. Met variantie-analyse behoren deze analysemodellen tot de groep van lineaire-analysemodellen. De aantrekkelijke statistische eigenschappen van het lineaire-analysemodel hebben geleid tot het formuleren van de genoemde varianten en de bijbehorende statistische toetsen.

5. GENERALISERINGEN VAN HET LINEAIRE MODEL

Een van de eisen voor toepassing van het lineaire model is dat de kenmerken tenminste gemeten zijn op intervalniveau. In feite betekent dit dat de scores op een kenmerk vergelijkbaar kunnen worden gemaakt met de scores op een ander kenmerk door aanpassing van het nulpunt van de meetschaal en de meeteenheid. We passen deze procedure bijvoorbeeld toe bij het vergelijken van graden Celsius en graden Fahrenheit: het aantal graden Celsius gaat na de lineaire transformatie F = 9/5 C + 32 over in graden Fahrenheit. Het zoeken van een lineaire combinatie van kenmerken kunnen we ook opvatten als het gewogen optellen van metingen die tot dezelfde schaal zijn teruggevoerd. Van ordinaal gemeten variabelen is al opgemerkt dat het soms mogelijk is het meetniveau te verhogen tot intervalniveau door een transformatie op de meetwaarden toe te passen. Voor nominale variabelen lijkt dit principieel onmogelijk.
In de oplossing die voor dit probleem is gekozen, gaan we ervan uit dat elke klasse van een kenmerk wordt opgevat als een apart kenmerk. Dit nieuwe klassekenmerk heeft dan twee klassen: een object valt in de ermee corresponderende klasse van het oorsponkelijke kenmerk of niet. Op deze klassekenmerken wordt het lineaire-analysemodel toegepast.
Wanneer we te maken hebben met variabelen met twee klassen, dan spreken we van binaire variabelen. Omdat een binaire variabele slechts twee klassen bezit, voldoet deze automatisch aan de voorwaarden voor het meten op intervalniveau: elke willekeurige keuze van waarden voor de twee klassen kan met behulp van een lineaire transformatie worden veranderd in elke andere keuze. In het algemeen geldt dan ook dat binaire variabelen in een lineair analysemodel kunnen worden opgenomen. Dit geldt ook voor de klassekenmerken die zijn geconstrueerd volgens bovenstaande procedure. De geconstrueerde binaire klassekenmerken verschillen natuurlijk wel van normale binaire variabelen. Zouden we bijvoorbeeld uit een normale binaire variabele voor elk van de twee klassen een binair klassekenmerk construeren, dan zijn de scores op beide klassekenmerken tegengesteld: een object dat behoort tot het eerste klassekenmerk, behoort niet tot het tweede klas-

sekenmerk en omgekeerd. Met deze onderlinge relaties tussen klassekenmerken behorend bij de klasse van eenzelfde kenmerk dient rekening te worden gehouden bij het oplossen van het analyseprobleem.
Als nominale variabelen op bovenstaande wijze in de analyse worden opgenomen, dan moeten aan de oplossing bepaalde restricties worden opgelegd. De gewichten die in een analyse aan de klassekenmerken van elk kenmerk worden toegekend zijn van elkaar afhankelijk. Als we in een enkelvoudige lineaire-regressieanalyse een binaire variabele met klassewaarden +1 en -1 opnemen en een gewicht a vinden voor de binaire variabele, dan geldt voor de objecten in de eerste klasse dat zij score a krijgen en voor de objecten in de tweede klasse dat zij score -a krijgen.
Indien klassekenmerken waren gevormd met klassewaarden 1 en 0, dan zou het eerste klassekenmerk het gewicht a moeten krijgen en het tweede klassekenmerk het gewicht -a om de objecten uiteindelijk dezelfde scores te laten krijgen. De som van de scores op de klassekenmerken levert dus weer de score op de oorsponkelijke binaire variabele op: $a.1 + (-a).0 = 1$ en $a.0 + (-a).1 = -a$. De gewichten die voor de klassekenmerken worden gevonden zijn dus evenredig met de oorsponkelijke klassewaarden. Indien we u nu voor een nominaal kenmerk klassekenmerken vormen, dat kunnen naar analogie van het bovenstaande voorbeeld, de gewichten die worden gevonden voor de klassekenmerken worden opgevat als scores van de objecten op het oorsponkelijke kenmerk. Kennen we de gevonden gewichten bijvoorbeeld als nieuwe klassewaarden toe aan de klassen van het oorsponkelijke kenmerk, dan levert een analyse waarbij dit kenmerk als metrische variabele wordt opgenomen hetzelfde resultaat op.
De toegepaste truc heeft geleid tot een soort herschaling van de klassewaarden zo dat deze zo goed mogelijk past in de gewenste beschrijving.
In Tabel 1 is een MLR-voorbeeld gegeven. Zes objecten zijn geclassificeerd op drie kenmerken. De klasse-indeling op het derde kenmerk (Y) wordt voorspeld met behulp van de eerste twee kenmerken (X_1 en X_2) met resp. drie en twee klassen die echter eerst geschaald worden.
Uiteindelijk geldt voor $\hat{Y}$, de voorspelling van Y en X_1 en X_2:

$$\hat{Y} = 2 \times (1 \times X_{11} + 0 \times X_{12} + 1 \times X_{13}) + 1 \times (1 \times X_{21} + 2 \times X_{22})$$

We zien dat de klassen 1 en 3 van het eerste kenmerk dezelfde schaal-

waarde (=1) krijgen en klasse 2 een afwijkende waarde (=0). Zouden we de schaalwaarden twee maal zo groot kiezen en het gewicht voor de geschaalde variabele (=2) halveren, dan zou de oplossing gelijk blijven.

objecten	kenmerken			klassekenmerken					voorspelling	geschaalde kenmerken	
	X_1	X_2	Y	X_{11}	X_{12}	X_{13}	X_{21}	X_{22}	$\hat{Y}$	X_1^*	X_2^*
1	3	1	3	0	0	1	1	0	3	1	1
2	1	2	4	1	0	0	0	1	4	1	2
3	2	1	1	0	1	0	1	0	1	0	1
4	2	2	2	0	1	0	0	1	2	0	2
5	3	1	3	0	0	1	1	0	3	1	1
6	1	1	3	1	0	0	1	0	3	1	1

Tabel 1.

Men kent volgens deze procedure dus aan nominale klassen metrische klassewaarden toe die leiden tot een zo eenvoudig mogelijke modelbeschrijving. Als kritiek op deze methode wordt wel genoemd dat we in nog sterkere mate dan bijv. bij MLR of factoranalyse het geval is, toewerken naar een zo goed mogelijke voorspelling. Men werkt dus toe naar het gewenste resultaat. Vandaar dat met name de belangstelling groot is voor de stabiliteit van de oplossing voor steekproeffluctuaties.

Het moge uit het bovenstaande duidelijk zijn dat de volgorde van de klassen binnen een kenmerk niet van belang is: de oplossing is onafhankelijk van de volgorde van de kenmerken waarvoor gezamenlijk een oplossing wordt gezocht, dus ook onafhankelijk van de volgorde van de klassekenmerken.

Als we de methode toepassen op de klassen van een ordinale variabele dan kunnen we dus nagaan of de volgorde van de gewichten in overeenstemming is met de verwachting. In dat geval moeten de gewichten van de klassekenmerken in stijgende of dalende volgorde staan. Daarnaast

is het echter mogelijk om extra restricties toe te passen op de gewichten van de klassekenmerken. We kunnen bijvoorbeeld aan de oplossing de restrictie toevoegen dat de gewichten voor opeenvolgende klassekenmerken in toenemende grootte behoren te worden toegekend aan de klassekenmerken. De dan resulterende gewichten kan men interpreteren als een herschaling van de ordinale variabele op intervalniveau. De transformatie van de ordinale variabele naar intervalniveau behoeft dus niet vooraf te worden meegegeven, maar kan achteraf worden geconstateerd. De herschaalde variabelen kunnen vervolgens worden opgenomen in metrische analyses. Op dezelfde wijze kan worden nagegaan of de schalingen van variabelen waarvan we veronderstellen dat zij op intervalniveau zijn gemeten, inderdaad aan deze voorwaarden voldoen.

Bij de afdeling Datatheorie van de R.U.-Leiden zijn een aantal programma's ontwikkeld die het mogelijk maken factoranalyse en canonische analyses ook toe te passen op nominale en ordinale gegevens. Het generaliseerde factoranalyseprogramma heet HOMALS en het canonische-analyseprogramma CANALS. De uitgang 'ALS' is een afkorting van 'Alternating Least Squares'. Hiermee wordt aangegeven dat voor het vinden van een optimale oplossing gebruik wordt gemaakt van de kleinste-kwadratenmethode waarbij een iteratief proces om en om wordt toegepast om de herschaling van de klassen te verbeteren en de gewichten van de herschaalde variabelen opnieuw vast te stellen, totdat de beste oplossing wordt gevonden.

Uit de beschrijving van de verschillen tussen MLR, (canonische) discriminant-analyse en canonische analyse zal duidelijk zijn dat in het gegeneraliseerde canonische-analyseprogramma CANALS de andere genoemde analyses als bijzonder geval kunnen worden gedefinieerd. Is er bijvoorbeeld in de tweede groep variabelen sprake van slechts één metrische variabele dan hebben we te maken met MLR. Zijn de variabelen in de eerste groep niet alle metrisch, dan spreken we van niet-metrische MLR. Is de enkele variabele in de tweede groep nominaal dan spreken we van discriminantanalyse of canonische-discriminant-analyse. Worden aan alle variabelen intervalrestricties opgelegd, dan volgt een normale (metrische) canonische analyse. Alle mengvormen

van metrische en niet-metrische kenmerken kunnen worden geanalyseerd. CANALS is dus een zeer algemeen programma.

In een gezamenlijk project van de afdeling Datatheorie R.U.-Leiden en SWOV wordt gezocht naar verdere uitbreiding voor situaties met meer dan twee groepen van kenmerken. Hierdoor is het mogelijk stapsgewijze verklaringen te geven voor gecompliceerde relaties binnen de verkeersveiligheid, zoals relaties tussen kenmerken van ongevallen, de weg en het verkeer.
De schalingen van de variabelen die resulteren uit HOMALS zijn anders te interpreteren dan de schalingen van CANALS. Binnen een HOMALS-analyse, gericht bijvoorbeeld op de beschrijving van de homogeniteit van wegkenmerken, zullen we bijvoorbeeld een ordinale schaal kunnen vinden voor het kenmerk wegbreedte. Als het gaat om een voorspelling van de onveiligheid als functie van o.a. de wegbreedte met CANALS, dan zou een U-vormige schaling wel een grotere verklaringskracht kunnen hebben. Dit zou dan suggereren dat smalle en brede wegen relatief veiliger zijn dan de middencategorie.
Dergelijke problemen bij de verklaring van bepaalde verschijnselen die bijvoorbeeld bij MLR leiden tot schijnbaar afwezige samenhangen, kunnen met CANALS aan het licht worden gebracht.

6. BESCHRIJVEN EN TOETSEN

De generalisering van technieken die oorspronkelijk zijn bedoeld voor metrische variabelen naar situaties waarbij ook ordinale of nominale kenmerken een rol spelen, heeft vooral voor de beschrijving van relaties tussen veel nominale kenmerken een grote betekenis. Problemen rijzen echter bij het statistisch toetsen van gevonden relaties. De grote mate van vrijheid die er is bij het zoeken van een best-passende oplossing in de boven omschreven technieken brengt de vraag naar de stabiliteit en de betrouwbaarheid van de oplossing met zich mee. Het ontbreken van een gefundeerde statistische theorie hierover heeft geleid tot een zekere reserve bij statistici tegenover deze generaliseringen. De behoefte, vooral bij onderzoekers binnen de sociale wetenschappen, aan op zijn minst eenvoudige beschrijvingen van de samenhangen tussen grote hoeveelheden onderzoekgegevens heeft echter geleid tot een grote belangstelling voor genoemde generaliseringen.
Men moet zich wel realiseren dat het daarbij in de eerste plaats gaat om theorievorming en niet om verificatie van theorieën of toetsing van hypothesen. De laatste tijd echter, vooral door de ontwikkeling in computerverwerking, worden ook de mogelijkheden om betrouwbaarheidsuitspraken te doen omtrent de gegeven beschrijvingen groter. Hoewel het nog steeds zeer moeilijk is langs theoretische weg bruikbare statistische grootheden te formuleren, is het wel mogelijk langs empirische weg na te gaan hoe betrouwbaar of stabiel bepaalde oplossingen zijn. We kunnen dit doen door uit te gaan van theoretische verdelingen waaruit we steekproeven trekken die we vervolgens analyseren. De analyseresultaten kunnen worden vergeleken met de kenmerken van de populaties waaruit de steekproeven afkomstig zijn. Bij grote hoeveelheden echte onderzoekgegevens kan dit gebeuren door steekproeftrekking uit de tot populatie verheven oorsponkelijke steekproef. Uit de fluctuaties in de schatters van de vele steekproeven kan een indicatie van de betrouwbaarheid van de uiteindelijke oplossing worden verkregen. Dergelijke procedures, vaak beschreven onder namen als 'monte-carlo-studies', 'jackknife-studies', 'bootstrap-studies', worden nu met behulp van de computer toegepast om theoretische complicaties te omzeilen.

7. VOORBEELD VAN EEN ANALYSE MET BEHULP VAN CANALS

Het volgende voorbeeld betreft een canonische analyse van nominale gegevens afkomstig uit het SWOV-onderzoek aan autogordels, uitgevoerd met behulp van het programma CANALS.
Van een aantal kenmerken van bij ongevallen betrokken bestuurders, hun voertuig en de omstandigheden waaronder het ongeval plaatsvond, is nagegaan welke relatie deze hebben met kenmerken waarin de ernst en de schade van het ongeval is vastgelegd.
De onafhankelijke variabelen die in deze analyse worden gebruikt, zijn: bebouwing (plaats van het ongeval: binnen of buiten de bebouwde kom), botspartner (of botsobject), soort (of plaats) van de schade, snelheid, gewicht, lengte, bouwjaar en type voertuig, aantal portieren, plaats van de motor, merk voertuig, geslacht bestuurder, leeftijd, gordelgebruik, uitslingeren uit voertuig en de inzittendenconfiguratie.
De afhankelijke variabelen zijn: algemene ernst van het letsel bij de bestuurder, plaats van het ernstigste letsel, soort letsel en ernst van de schade aan het voertuig.
Bij de analyse zijn dus zestien onafhankelijke variabelen en vier afhankelijke variabelen gebruikt. Alle zijn als nominale variabelen in de analyse opgenomen. Het onderzoek betrof 7748 ongevallen met gewonde bestuurders.
Bij deze analyse is slechts naar één enkele factor gezocht, of liever naar een 'canonische as'.
De canonische correlatie, dit is de correlatie tussen de gevonden lineaire combinatie van geschaalde onafhankelijke variabelen met die van de geschaalde afhankelijke variabelen, bedraagt 0.62. Een optimale beschrijving wordt gevonden bij een maximale waarde voor deze canonische correlatie.
Een beoordeling van de hoogte van de gevonden correlatie is moeilijk te geven. Deze hangt ondermeer af van het aantal variabelen en hun categorieën. Op grond van ervaring met andere analyses kan de correlatie als redelijk hoog worden aangeduid.
De gevonden gewichten voor de variabelen staan vermeld in Tabel 2.

Onafhankelijke variabelen:	Gewichten geschaalde variabelen	Correlaties met canon. scores tweede groep	Correlaties met canon. scores eerste groep
Bebouwing	-.12	-.02	-.03
Botspartner	-.23	-.22	-.36
Soort schade	.77	.56	.90
Snelheid vtg.	-.21	-.39	-.63
Gewicht vtg.	-.06	.02	.03
Lengte vtg.	-.09	-.05	-.08
Bouwjaar vtg.	.06	.06	.10
Type vtg.	-.07	-.07	-.11
Portieren (aantal)	-.03	-.03	-.04
Plaats motor	-.04	-.00	-.00
Merk vtg.	.09	.05	.08
Geslacht best.	.07	.05	.08
Leeftijd best.	.04	.06	.10
Gordelgebruik	-.08	-.10	-.16
Uitslingeren	-.14	-.16	-.27
Inzittendenconf.	-.05	-.03	-.04
Afhankelijke variabelen:			
Algemene ernst	-.23	-.40	-.24
Plaats letsel	.78	.85	.52
Soort letsel	.04	-.26	-.16
Ernst schade	-.44	-.60	-.37

Tabel 2.

Over het teken van deze gewichten kan worden opgemerkt dat dit mag worden veranderd, als we tegelijk alle tekens van de klassescores die zijn gevonden ook veranderen. Vinden we bijvoorbeeld zoals hier een negatief gewicht voor 'snelheid' (-.21) en oplopende klassescores van lage naar hog snelheden, dan wordt, als de klassescores van teken worden verwisseld, dus dan aflopen met de snelheid, een positief gewicht gevonden. De combinatie van tekens is natuurlijk wel eenduidig. We zouden kunnen zeggen: als toename in snelheid negatief moet worden beoordeeld, dan moet toename in traagheid positief worden beoordeeld. Hetzelfde verschijnsel wordt alleen anders geformuleerd.
De absolute grootte van het gewicht is ondermeer afhankelijk van de

categoriescores. Vermenigvuldigen we de categoriescores alle met een constante en delen we het gewicht van het kenmerk door diezelfde constante, dan behouden we hetzelfde resultaat. Een zekere normering van de categoriescores is dan ook gewenst om de gewichten vergelijkbaar te maken. Voor elke variabele geldt dat de gemiddelde score gelijk is aan nul en de kwadratensom van de scores gelijk is aan het totale aantal observaties.

Als we kijken naar de gewichten van de vier afhankelijke variabelen, dan zien we dat de plaats van het letsel de variabele is met het hoogste gewicht (.78). Als we niet letten op het teken, dan volgen ernst van de schade (-.44) en algemene ernst van het letsel (-.23). Het soort letsel krijgt een klein gewicht (.04).
In Afbeelding 1 t/m 4 zijn de categoriescores voor de vier kenmerken weergegeven.
Men ziet daaruit dat bij de interpretatie van de canonische as bij de plaats van het letsel (Afb. 2), vooral het onderscheid tussen nekletsel (categorie 3) en andersoortig letsel een rol speelt.
Met betrekking tot de ernst van de schade (Afb. 4) en de algemene ernst van het letsel (Afb. 1) zien we een toename van de toegekende categoriescore met de toename in schade en letselernst.
Uit het negatieve gewicht van beide kenmerken in combinatie met deze schaling, kan worden afgeleid dat de ernst van nekletsel relatief laag is en bij de relatief minder ernstige ongevallen plaatsvindt: een negatief gewicht voor algemene ernst (-.23) en een lage categoriescore voor licht + matig letsel (-.40) hangt bijvoorbeeld samen met een positief gewicht voor plaats van het letsel en hoge categoriescore voor nekletsel. Voor zover het soort letsel hierbij een rol speelt, is de schaling consistent met bovenstaand beeld: bij nekletsel gaat het niet om wonden of breuken, maar om inwendige letsels.

Bij de verklaring voor het hierboven beschreven letsel zien we dat met name de onafhankelijke variabele: soort schade, met een gewicht van .77, van belang is, daarnaast kunnen variabelen: botspartner (-.23), snelheid (-.21), uitslingeren (-.14) en bebouwing (-.12) worden genoemd. Het gewicht voor het dragen van de gordel is klein (-.08).

In Afbeelding 5 t/m 10 staan de categoriescores voor de onafhankelijke variabelen afgebeeld.
Met betrekking tot de soort schade (Afb. 7) blijkt uit het gewicht in combinatie met de categoriescores dat schade achter aan het voertuig (categorie 7, 8 en 9) met het nekletsel samenhangt. De botspartner (Afb. 6) is bij nekletsel niet een vrachtauto (categorie 3, 12, 13), boom of paal (categorie 7, 9), maar het betreft hier eerder gecombineerde botsingen zonder vrachtauto (categorie 4) en slipongevallen (categorie 1). De snelheid (Afb. 8) bij ongevallen waarbij nekletsel optreedt is laag, minder vaak is de bestuurder uit de auto geslingerd (Afb. 10: categorie 2), vaker vindt het ongeval binnen de bebouwde kom plaats (Afb. 5: categorie 1). Er wordt vaker een gordel gedragen (Afb. 9: categorie 1).

In Tabel 2 vinden we tevens de correlaties tussen de scores op de geschaalde onafhankelijk variabelen met de canonische scores van de tweede groep.
Bekijken we deze correlaties, waarin naast de invloed van de gewichten van de variabelen ook het effect van de schaling is verwerkt, dan zien we dat voor bebouwing (plaats van het ongeval binnen of buiten bebouwde kom) de correlatie (-.02) relatief lager is dan het gewicht, en voor snelheid (-.39), gordelgebruik (-.10) en uitslingeren (-.16) hoger.
Uit de correlaties van de variabelen van de tweede groep met de canonische scores van groep twee, blijkt dat ook het soort letsel hiermee redelijk correleert.
Voor de volledigheid zijn in Tabel 2 bovendien de correlaties vermeld van de geschaalde variabelen met de canonische scores van de eerste groep. Voor de interpretatie zijn deze echter minder van belang.

Samenvatting en conclusies

In dit voorbeeld is gezocht naar een één-dimensionale oplossing, naar een enkele canonische as. De resulterende oplossing is te interpreteren als de samenhang die het geheel van verklarende kenmerken heeft met een bepaalde groep minder ernstige ongevallen, waarbij vooral het nekletsel is te onderscheiden van de andere letseltypen die optreden

bij ongevallen met minder ernstige schade. Kennelijk is deze groep ongevallen het meest ondubbelzinnig te onderscheiden van de andere ongevallen. Of het de enige interpreteerbare onderscheiding is, zou moeten blijken uit een analyse waarbij een oplossing in meer dan één dimensie wordt gezocht, of een waarbij deze groep ongevallen is verwijderd.

We zien dat de genoemde groep ongevallen de volgende kenmerken bezit: de schade is aan de achterzijde van de auto geconstateerd, de botspartner is geen vrachtauto, boom of paal, de snelheid van het voertuig is laag, de bestuurder is niet uit het voertuig geslingerd, het ongeval vond voornamelijk binnen de bebouwde kom plaats, terwijl vaker geen gordel is gedragen.

Voor alle genoemde variabelen van groep 1 en groep 2 is de schaling consistent met dit algemene beeld. Voor de andere variabelen lijkt de schaling tamelijk arbitrair, hetgeen niet verwonderlijk is als we, op grond van de gevonden gewichten en correlaties, aannemen dat die variabelen geen samenhang hebben met het bovenomschreven ongevallenpatroon en bijbehorend letsel.

Hoe stabiel de schaling van de relevante kenmerken is en hoe betrouwbaar de gewichten en correlaties zijn, is, zoals eerder is opgemerkt, niet direct aan te geven. De interne consistentie van het beeld echter dat wordt opgeroepen, zowel door de schalingen als door de gewichten, lijkt overtuigend. Hierbij kan worden aangetekend dat de analysetechniek alleen toewerkt naar een zo goed mogelijke beschrijving binnen de ruimte die in het model gegeven is. Bij het zoeken naar een oplossing speelt de interpreteerbaarheid geen rol. Hierin verschilt de analyse in gunstige zin van het op een heuristische wijze zoeken door grote aantallen kruistabellen, waarbij op een vaak inconsistente wijze alleen die verbanden worden opgemerkt en benadrukt die gemakkelijk te interpreteren zijn.

Willen we echter verder gaan en van het bovenstaande beeld de betrouwbaarheid vaststellen, dan kunnen we dit bijvoorbeeld doen door willekeurige deelgroepen van gegevens te maken en deze te analyseren, of door op nieuwe gegevens gerichte toetsen uit te voeren waarbij strikt gedefinieerde model-assumpties worden gehanteerd, die uit de bovenstaande beschrijving zijn afgeleid.

8. DE ANALYSE VAN KRUISTABELLEN

Een kruistabel, of liever zoals de Engelse benaming luidt een contingency table, is een tabel met aantallen observaties in de cellen. Observaties worden geclassificeerd naar een of meer kenmerken. Na de classificatie kan een tabel worden gemaakt waarin voor elke cel wordt geteld hoeveel observaties behoren bij de desbetreffende combinatie van klassen van de kenmerken.
Bij de analyse van een dergelijke tabel zijn we vooral geïnteresseerd in de verdeling van de observaties over de cellen. In het algemeen geldt dat het toeval een rol speelt bij het tot stand komen van de tabel. De observaties zijn op te vatten als steekproefelementen uit een populatie. Daardoor zijn de celaantallen aan schommelingen onderhevig, wanneer hetzelfde verschijnsel een aantal malen steekproefsgewijs wordt onderzocht. Dit wordt weergegeven door de Engelse benaming "contingency table".
Bij een analyse van kruistabellen willen we de systematische effecten scheiden van de toevallige fluctuaties die gezamenlijk resulteren in een bepaalde tabel. Daarvoor is een model nodig waarin de systematische effecten worden beschreven en een foutentheorie waarin rekening wordt gehouden met de afwijkingen van de geobserveerde celaantallen ten opzichte van de celaantallen die op grond van het model te verwachten zijn. Het gaat de onderzoeker uiteindelijk om de systematische effecten. Hij wil nagaan of de veronderstellingen die uit het model of de modelstructuur kunnen worden afgeleid met betrekking tot de tabel houdbaar zijn, gegeven de geobserveerde aantallen. Dit proces wordt aangeduid met hypothesetoetsing. Daarnaast zoekt hij vaak naar nadere specificaties van de modelstructuur om te komen tot een meer specifiek model. Men spreekt dan van parameterschatting. Zowel voor dit schatten van parameters als voor het toetsen van hypothesen is de toevalscomponent van belang. Ten aanzien van de toevalscomponent is in het verleden veel onderzoek gedaan. Wat de modelspecificatie betreft, hierin is door recent onderzoek zeer veel verbeterd. Dit heeft de analyse van kruistabellen opnieuw in de belangstelling gebracht. Het resultaat van dit recente onderzoek is ook voor het onderzoek naar de verkeersveiligheid van groot belang.

Verkeersveiligheid wordt beschreven met behulp van verkeersongevallen. Soms gebruikt men andere middelen zoals het aantal verkeersconflicten, als er geen bruikbare ongevallengegevens zijn. Dit is echter niet essentieel voor hetgeen volgt. De opmerkingen betreffen zowel ongevallen als conflicten.
Zelfs als ongevallenratio's (ongevallen per inwoner, per weglengteeenheid, per verreden voertuigkilometer) worden gebruikt, dan hebben we te maken met een maat voor de verkeersveiligheid die in feite berust op telgegevens.
Wanneer er de nadruk op wordt gelegd dat aan deze maten tellingen ten grondslag liggen, dan zal de analyse tenderen in de richting van het zoeken naar een verklaring van het gevonden aantal observaties in die specifieke cel van de tabel. Dit leidt tot vragen zoals "Wat is de kans op dat aantal observaties in die bepaalde cel?".
Een totaal andere benadering krijgen we als benadrukt wordt dat we te maken hebben met een verkeersveiligheidsmeting, een score voor verkeersveiligheid. De relevante vraag lijkt dan "Waarom is die score zo hoog of laag voor die bepaalde cel?".
Bij de analyse van kruistabellen zijn we, zoals gezegd, voornamelijk geïnteresseerd in de verdeling van de observaties over de cellen.
De tweede benaderingswijze leidt in de meeste gevallen tot analyses waarbij van het lineaire model wordt uitgegaan, zoals lineaire-regressie- of variantie-analyse of andere analyses van deze aard. De beide benaderingen leiden echter tot totaal verschillende modelaannamen. Wanneer we bijvoorbeeld te maken hebben met een twee-wegtabel van ongevallenratio's dan wordt bij een variantie-analysemodel de te verwachten score voor de cel <i,j> behorend bij rij i en kolom j in het algemeen beschreven als de som van een rijscore en een kolomscore. In formulevorm:

$$E(x_{ij}) = r_i + k_j$$

Bij de analyse van kruistabellen wordt meestal uitgegaan van een multiplicatief model, hoewel ook additieve modellen (of liever

modellen met multiplicatieve en additieve componenten) bestaan. In zo'n multiplicatief model geldt dan:

$$E(n_{ij}) = r_i \cdot k_j$$

In de praktijk hangt de keuze van de modelstructuur vaak minder af van aard van de onderzoekgegevens dan van factoren als de statistische eenvoud van de modellen en de kennis van of de beschikbaarheid van de analysetechnieken.
Vanuit de statistiek bezien heeft het lineaire model veel voordelen. Voor elke component uit het lineaire model zijn vrij eenvoudig de bijbehorende statistische grootheden aan te geven. Verder sluit de modelopbouw zodanig aan op de veronderstelde structuur van de gegevens dat ook de ingewikkelde structuren eenvoudig zijn weer te geven als een som van modelcomponenten. De presentatie van het multiplicatieve model als een lineair model voor de logaritme van de celaantallen, resulteert in dezelfde voordelen. Dit verklaart dan ook de grote toename in belangstelling hiervoor.
Alvorens over te gaan tot een beschrijving van deze ontwikkeling lijkt het nuttig eerst in te gaan op het model dat ten grondslag ligt aan de analyse van kruistabellen.

Het multinomiale model

Als we willen beschrijven hoe een bepaalde kruistabel is verkregen, dan gaan we uit van een statistisch model, dat is gebaseerd op een denkbeeldig experiment.
De formulering van het gekozen denkbeeldige experiment leidt tot toepasbaarheid van het multinomiale model op de uitkomsten van het experiment. Het experiment bestaat uit een aantal onafhankelijke trekkingen van een steekproefelement uit een populatie. Elke trekking resulteert in één uitkomst uit een aantal mogelijke uitkomsten. Om dit model op de besproken kruistabellen te kunnen toepassen zijn de volgende aannamen van belang:
1. De kans dat een bepaalde observatie wordt geclassificeerd in een bepaalde cel is onafhankelijk van de classificatie van andere observaties.

2. Deze kans is voor alle observaties gelijk.
3. De observaties worden in één en slechts één cel geclassificeerd, met andere woorden de gebeurtenissen zijn wederzijds uitsluitend en uitputtend.
4. Er gelden geen aannamen over de kans dat er gebeurtenissen optreden, men gaat ervan uit dat de gebeurtenissen plaatsvinden.

Het multinomiale model kan bij verkeersveiligheidsonderzoek worden gebruikt voor de analyse van aantallen ongevallen. Gaat het echter om aantallen bij ongevallen betrokken voertuigen, of om aantallen slachtoffers, dan is het model niet toepasbaar. Meerdere observaties kunnen dan bij een ongeval behoren en de in klassen ingedeelde observaties zijn niet meer onafhankelijk.
Indien het multinomiale model wordt toegepast bij de analyse van kruistabellen, dan dient men zich te realiseren dat de gebeurtenissen in feite als samengestelde gebeurtenissen worden opgevat. Classificatie van een observatie in cel <i,j> van een twee-wegtabel betekent in feite classificatie in zowel rij i als in kolom j. Vaak wordt daarbij verondersteld dat de indeling in rij i onafhankelijk is van de indeling in kolom j. Dit is echter een geheel andere soort onafhankelijkheid dan de onder 1. omschreven aanname van het multinomiale model, de tweede voorwaarde geldt alleen als het model wordt toegepast op een kruistabel waarbij elke uitkomst wordt opgevat als een samenstel van twee andere uitkomsten die onafhankelijk van elkaar zijn. Van twee gebeurtenissen A en B wordt gezegd dat zij statistisch onafhankelijk van elkaar zijn, als de kans $P(A \cap B)$ dat zij zich tegelijk voordoen gelijk is aan het produkt $P(A).P(B)$ van de kansen dat elk van de gebeurtenissen zich afzonderlijk voordoet. In de toepassing hiervan op de analyse van kruistabellen geldt, bij onafhankelijkheid van rijen en kolommen, dat voor elke cel <i,j> de kans p_{ij} op een observatie in die cel is te beschrijven als het produkt $p_{i.}.p_{j}$ van de kansen op een observatie in rij i en in kolom j. Dus de hypothese van geen samenhang of interactie tussen rijen en kolommen van een tabel komt neer op de aanname van een multiplicatief model voor de celkansen en leidt derhalve tot het eerder genoemde multiplicatieve

model voor de celaantallen. De gebruikelijke Chi-kwadraattoets voor het toetsen van de hypothese dat er geen samenhang is tussen het rijkenmerk en het kolomkenmerk van een twee-wegtabel is hiervan een voorbeeld. De kansen $p_{i.}$ en $p_{.j}$ worden daarbij geschat met behulp van de rij- en kolomfrequenties van de tabel.

Het Poissonmodel

Naast het multinomiale model voor analyse van kruistabellen bestaat er een ander, in feite sterker, model dat vaak bij verkeersveiligheidsgegevens wordt gebruikt. Voor toepassing hiervan is naast de voorwaarden die zijn genoemd bij het multinomiale model nog een extra aanname nodig. Deze heeft betrekking op het optreden van gebeurtenissen. De observaties worden nu niet als gegeven verondersteld, maar er wordt aangenomen dat het optreden van gebeurtenissen met een Poissonproces beschreven kan worden. Het aantal observaties is het resultaat van een kansverschijnsel.
Als we er van uitgaan dat voor elke cel van de tabel geldt dat de aantallen gebeurtenissen Poisson verdeeld zijn, dan geldt ook dat het totale aantal gebeurtenissen die in de tabel zijn opgenomen Poisson verdeeld is. We kunnen verder afleiden dat de voorwaardelijke kans op de aantallen in de cellen, gegeven het totale aantal gebeurtenissen, beschreven kan worden met behulp van de multinomiale verdeling. In feite is het multinomiale model dan ook te beschouwen als een bijzonder geval het het Poissonmodel.
Om na te gaan of de extra assumptie gerechtvaardigd is, zullen we eerst ingaan op de meest gebruikelijke interpretatie van het Poissonmodel. Ook hier betreft het weer de beschrijving van een denkbeeldig proces, waarvan de uitkomst nu te beschrijven valt met behulp van de Poissonverdeling.
De essentiële assumpties zijn dat het optreden van gebeurtenissen niet afhangt van de voorgeschiedenis, met andere woorden dat de gebeurtenissen onafhankelijk van elkaar optreden, en dat de kans dat een gebeurtenis optreedt niet verandert in de tijd.
De eerste aanname vindt weinig weerstand bij het verkeersveiligheidsonderzoek: ongevallen zijn zeldzame gebeurtenissen en zeer

zelden is een ongeval oorzaak voor een ander ongeval. In de meeste gevallen waarbij dit zich toch voordoet wordt zo'n aaneenschakeling van ongevallen beschouwd als een (gecompliceerd) ongeval.
De tweede assumptie echter houdt veel onderzoekers bezig. De verkeersstroom verandert soms snel in de tijd en de ongevallenkans wordt verondersteld mee te veranderen.
Echter, de homogeniteitsassumptie behoeft niet noodzakelijk voor een lange tijdperiode te gelden. Geldt voor een aantal korte tijdperioden dat de ongevallen binnen elke periode Poisson verdeeld zijn, dan geldt ook voor het totaal van de korte perioden dat het aantal ongevallen Poisson verdeeld is, ook al verschillen de Poissonparameters.
In veel gevallen heeft beschrijving met behulp van een Poissonverdeling tot bevredigende resultaten geleid. In enkele gevallen, waar een Poissonproces verondersteld werd voor een aantal situaties elk met een andere ongevallendichtheid, dus met een andere Poissonparameter, bleek de verdeling van alle gegevens te beschrijven met behulp van een negatieve-binomiaalverdeling, zoals te verwachten is door aan de Poissonaanname een extra aanname omtrent de verdeling van de Poissonparameters toe te voegen.
De Poissonverdeling wordt in veel verkeersveiligheidsonderzoek gebruikt. Vooral de laatste jaren is het multiplicatieve Poissonmodel vaak toegepast.
Rasch (1973) paste het model toe op verkeersongevallen uitgesplitst naar wegcategorieën en dagen van de week. Verder gebruikte hij het model om de ongevalsvatbaarheid van verschillende bestuurders vast te stellen en om na te gaan hoe deze verandert in de tijd.
Hamerslag & Huisman (1977) gebruikten het multiplicatieve Poissonmodel om de onveiligheidsparameters te schatten voor de klassen van diverse ongevalskenmerken, onder de hypothese dat deze kenmerken onderling onafhankelijk zijn. Deze vrij sterke aanname kan slechts gedeeltelijk worden ondervangen door nieuwe kenmerken te construeren uit combinaties van oorspronkelijke kenmerken.
In De Leeuw & Oppe (1976) is een gewogen versie van het Poissonmodel gebruikt. Hiermee is het mogelijk, uitgaande van het loglineaire model, een tabel te beschrijven met behulp van de diverse

kenmerken waarnaar de tabel is uitgesplitst en hun onderlinge samenhangen. Ten aanzien van de weging is dit model te vergelijken met het "multiplicative Poissonmodel with unequal cellrates" van Andersen (1977).

9. RECENTE ONTWIKKELINGEN IN DE ANALYSE VAN KRUISTABELLEN: LOG-LINEAIRE MODELLEN

De meeste toepassingen van de analyse van kruistabellen zijn voorbeelden van het testen van de "geen-interactie" hypothese in tweewegtabellen. De kansen op rij- en kolomclassificatie worden geschat met behulp van de rij- en kolomtotalen. Hieruit worden verwachte aantallen $E_{ij} = n.p_{i.}p_{.j}$ berekend voor de cellen van de tabel. Deze worden vergeleken met de geobserveerde aantallen O_{ij}. Een discrepantiemaat tussen beide series waarden, Chi-kwadraat genoemd, is als volgt gedefinieerd.

$$X^2 = \sum_{i,j} (O_{ij} - E_{ij})^2 \Big/ E_{ij}$$

Onder de aanname dat er geen interactie is, berust de grootte van X^2 alleen op de invloed die het toeval heeft op de waarden van O. Daarom wordt verondersteld dat de verdeling van X^2-waarden van de multinomiale verdeling kan worden afgeleid.
Voor ieder totaal aantal observaties en voor elk model waaruit de celkansen zijn af te leiden, is er een discrete verzameling van mogelijke waarden voor X^2. De kans op elk van de uitkomsten berust op de kans dat een bepaalde verzameling celwaarden wordt gevonden. Berekening van deze exacte kansen vraagt echter veel tijd.
Alleen in eenvoudige gevallen is een zo te definieren exacte test praktisch toepasbaar. Fisher's exacte test voor 2x2 tabellen is hiervan een voorbeeld. Daarom worden in de praktijk extra aannamen gemaakt omtrent de verdeling van de X^2-waarden. In feite wordt ervan uitgegaan dat de X^2-waarde een som is van een aantal gekwadrateerde standaard-normaal verdeelde scores. Hierbij wordt gebruik gemaakt van een bekende limietstelling voor de multinomiale verdeling. Waarden voor deze theoretische Chi-kwadraatverdeling, aangeduid met χ^2, zijn eenvoudiger te berekenen en kunnen uit tabellen worden afgelezen. De waarden berusten uitsluitend op het aantal vrij te kiezen celwaarden van een tabel, gegeven het totale aantal observaties en de in het model gespecificeerde restricties. Dit aantal wordt het aantal vrijheidsgraden genoemd. Het gebruik van

deze tabellen is echter alleen toegestaan bij relatief grote aantallen observaties. Slechts dan is de limietstelling te gebruiken. Daarom spreken we vaak van tests voor grote steekproeven (large sample tests).
Er is nog niet zoveel bekend over de bruikbaarheid van de χ^2-verdeling bij kleine steekproeven, voor andere dan de 2x2 tabel. Er zijn een aantal vuistregels voor bruikbaarheid. Cochran (1952) geeft hiervan een overzicht. De laatste tijd wordt hieraan echter meer aandacht besteed. Ook bij de SWOV is onderzoek op dit terrein in het programma opgenomen. Het gaat hier met name om een vergelijking van de bruikbaarheid van diverse schattingsmethoden voor de parameters van het log-lineaire model bij variaties van de marginale kansen en de grootte van de steekproef.
De noodzaak van grote steekproeven brengt onderzoekers soms in een moeilijke positie. Vaak zijn voor statistische analyses slechts kleine aantallen ongevallen ter beschikking. Dit is een van de redenen waarom de Chi-kwadraattest lange tijd zo weinig aandacht heeft gekregen. Er is nog een reden. De Chi-kwadraattest, zoals boven omschreven, wordt gebruikt als een test voor interacties tussen kenmerken. Indien de test aangeeft dat er van samenhang sprake is, dan vertelt ze niet hoe deze samenhang eruit ziet. De test constateert slechts de aanwezigheid ervan. Met andere woorden de Chi-kwadraattest is een slecht hulpmiddel als het gaat om het formuleren van theorieën.
Een ander probleem dat vaak naar voren komt bij het testen van hypothesen met de Chi-kwadraattoets, is dat kleine, onbelangrijke samenhangen leiden tot het verwerpen van de nul-hypothesen. Dit is het geval als zeer grote steekproefaantallen worden verzameld. Naast de vraag of een samenhang <u>significant</u> is, is het dus de vraag of deze <u>relevant</u> is.
Dit probleem en het voorgaande benadrukken de behoefte aan parameterschatting en het berekenen van de betrouwbaarheidsgrenzen van parameters als aanvullende informatie naast het toetsen van hypothesen.
De reden waarom de belangstelling voor de analyse van kruistabellen vooral sinds 1960 zo is toegenomen, is niet omdat het probleem van

de kleine aantallen is opgelost, maar vooral omdat een meer gedetailleerde toetsing van het model mogelijk is, gepaard aan het schatten van parameters en het aangeven van betrouwbaarheidsintervallen hiervan.
De ontwikkelingen zijn onder te brengen onder de volgende punten:
1. De toepasbaarheid van de Chi-kwadraattoets als toets van de hypothese dat er geen interactie is tussen kenmerken, is uitgebreid tot tabellen met meer dan twee kenmerken. Foldvary & Lane (1974) gebruikten in hun onderzoek naar het effect van de wet op de draagplicht van gordels, deze methode om de totale Chi-kwadraatwaarde te splitsen in eerste, tweede en derde orde-interacties in een vierwegtabel.
2. Toetsen beperken zich niet meer tot totale effecten. Chi-kwadraten kunnen worden gesplitst in Chi-kwadraatwaarden voor deelhypothesen. Dit niet alleen voor het interactieniveau van kenmerken zoals bij Foldvary & Lane, maar ook binnen kenmerken, bijvoorbeeld in een twee-wegtabel om de invloed van een bepaalde klasse, of combinatie van klassen, op het interactie-effect na te gaan. Het voordeel van deze methode is, dat hierbij geen overlappende toetsen van 2x2 deeltabellen worden gebruikt, maar dat een exacte opsplitsing plaats vindt van de totale Chi-kwadraat in Chi-kwadraten voor deelhypothesen.
3. Als gevolg van de laatstgenoemde ontwikkeling is meer aandacht besteed aan het schatten van parameters. Bovengenoemde deelhypothesen worden geformuleerd door restricties op te leggen aan bepaalde parameters van het model, zoals lineaire restricties of kwadratische restricties ten aanzien van de rij-parameters enz.
4. De analyse-eenheid is uitgebreid tot gewogen of genormeerde aantallen, bijvoorbeeld ongevallen per km weg. Hierdoor is het onder andere soms mogelijk om behalve interactie-effecten ook hoofdeffecten te toetsen, bijvoorbeeld als we willen nagaan of er een significant verschil is tussen het aantal ongevallen met dodelijke afloop in verschillende landen per hoofd van de bevolking.

De belangrijkste reden om deze uitbreidingen van de Chi-kwadraattoets te gebruiken is echter het gevolg van de presentatie van de theorie in termen van het lineaire model. Het lineaire model wordt

geacht toepasbaar te zijn op de logaritme van de aantallen, vandaar de naam log-lineair model. In het geval van een twee-wegtabel zegt het log-lineair model dat de logaritme van de verwachte aantallen als volgt is opgebouwd:

$$\ln(E(x_{ij})) = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

Dus uit een voor alle cellen identieke parameter μ, voor iedere rij i is een vaste parameter α_i en voor iedere kolom j een β_j en een voor iedere cel unieke parameter γ_{ij}.
Als de parameters onbekend zijn en moeten worden geschat uit de geobserveerde celaantallen, dan is het altijd mogelijk een perfecte oplossing voor de bovenstaande parameters te vinden. Het model wordt daarom wel (over)verzadigd genoemd. Het testen van de hypothese dat rijen en kolommen onafhankelijk van elkaar zijn, betekent het toetsen of voor alle γ_{ij}'s geldt dat deze gelijk aan nul zijn.
Striktere modellen zoals $\ln(E(x_{ij})) = \mu + \alpha_i$ zijn nog minder verzadigd en kunnen dan ook binnen het eerder genoemde model worden getoetst. Het is verder mogelijk restricties op parameters te toetsen, zoals de restrictie dat de α's alle op een rechte lijn liggen, hetgeen overeenkomt met de hypothese dat er een exponentieel verband is tussen de aantallen in de rij-klassen.
Door deze wijze van presentatie is de generalisering naar hogere-ordetabellen direct duidelijk: we voegen voor elke combinatie van kenmerken nieuwe parameters toe.
Modellen voor deze parameters en tests hiervoor zijn conceptueel gezien, gemakkelijk te formuleren.
Een uitvoerige beschrijving van het log-lineaire model en de mogelijkheden ervan vindt men in Bishop, Fienberg & Holland (1975), verder geven Gokhale & Kullback (1978) een zeer heldere beschrijving vanuit het gezichtspunt van de informatieverwerking.
Soms wordt het log-lineaire model verward met het log-normale model. Het log-normale model wordt o.a. gebruikt bij variantieanalyse om de variantie van de observaties van scores te stabiliseren.
De sterke gelijkenis met variantie-analyse wordt het best duidelijk gemaakt door Nelder & Wedderburn (1972). In hun veel toegepaste

computerprogramma GLIM (Generalized Linear Models) hebben zij het log-lineaire model opgenomen als een speciaal geval van het lineaire model. Het lineaire model en de vereiste transformatie om het model toe te kunnen passen zijn hierbij gescheiden.
Dit brengt ons terug tot het probleem van de additieve en multiplicatieve modellen voor kruistabellen. Darroch (1974) die beide modellen vanuit een statistisch standpunt bekeek, geeft de volgende presentaties van interactie voor een drie-wegtabel.
Voor het multiplicatieve model geldt:

$$P_{ijk} = \alpha_{ij} \cdot \beta_{ik} \cdot \gamma_{jk}$$

en voor het additieve model:

$$P_{ijk} \Big/ P_{i}.P_{j}.P_{k}) = a_{ij} + b_{ik} + c_{jk}$$

Het additieve model is geïntroduceerd door Lancaster (1951). Hij gebruikte het model om de totale Chi-kwadraat (die kan worden opgevat als een residuwaarde) te splitsen in effecten die samenhangen met de a-, b- en c-parameters. Zodoende worden de residuwaarden verder getest met behulp van een additief model.
Dit compromis tussen interactie, gedefinieerd in termen van statistische onafhankelijkheid en een ontbinden van de residuwaarden met hulp van een additief model, leidt bij tabellen met meer dan twee kenmerken tot andere resultaten dan bij het log-lineaire model. Ondanks sommige voordelen van het additieve model besluit Darroch tot een lichte voorkeur voor het multiplicatieve interactiemodel, uitsluitend bekeken vanuit de statistische eigenschappen van de modellen.
Oppe (1978) paste een gegeneraliseerd lineair model toe om een tabel met ongevallenratio's te analyseren. De ongevallen zijn uitgesplitst naar de kenmerken uurintensiteit en stroefheid van het wegdek. Er is een additief model gebruikt:

$$E(a^{*}_{ij}.) = r_{i} + c_{j}$$

Dit model wordt echter niet direct toegepast op de ratio's a zelf, maar op die monotone transformatie a^{*}_{ij} van de ratio's die zo goed mogelijk met het model overeenkomt. We spreken dan wel van "additief conjunct meten". Een beschrijving van dit type regressie-model vindt men in Kruskal (1965). De transformatie bleek achteraf de gedaante van een log-transformatie te bezitten. Dit suggereert dus een multiplicatief model in plaats van een additief model. Dit leidde tot toepassing van het log-lineaire model op de ongevallen-ratio's, waarbij de noemer als gewicht werd opgevat van het in de teller gegeven aantal.
Behalve de methodologische aantrekkelijkheid van het log-lineaire model en sommige statistische voordelen zijn er dus ook empirische resultaten die het multiplicatieve model ondersteunen voor dit type onderzoekgegevens.
In veel gevallen wordt een dergelijke test op het lineaire model achterwege gelaten en wordt al te gemakkelijk beslist dat afwijkingen van het model op toevalsfluctuaties berusten. Zo is men in lineaire-regressie-analyses zelden geïnteresseerd in de omvang van de foutcomponent, maar vraagt men zich vooral af of de model-parameters significant van nul verschillen.
Bij log-lineaire tests is dit niet het geval. De toets van het model impliceert aannamen omtrent de mate waarin de gegevens afwijken van het model hetgeen een groot voordeel is. Dit leidt sneller tot het verwerpen van een onjuist model.

10. VOORBEELD VAN EEN LOG-LINEAIRE ANALYSE

Als voorbeeld van een log-lineaire analyse is gekozen voor een drie-wegtabel. Hierin zijn voor twee jaren bloed-alcoholgegevens uitgesplitst naar geslacht (zie Tabel 3). De gegevens zijn afkomstig uit het SWOV-onderzoek naar Rij- en drinkgewoonten (SWOV, 1978).
De aantallen in de tabel betreffen bestuurders van personenauto's die in het najaar van 1975 en 1977 tijdens weekeindnachten van 10 uur 's avonds tot 4 uur 's nachts zijn onderzocht op hun bloed-alcoholgehalte (BAG). In Tabel 3a staan de aantallen weergegeven. In Tabel 3b de gewichtsfactoren.

BAG	1975		1977	
	M	V	M	V
<0,2	2275	448	1838	452
0,2-0,5	339	33	350	38
0,5-1,0	263	11	247	20
>1,0	163	10	145	9

Tabel 3a. Aantal onderzochte bestuurders, uitgesplitst naar geslacht en BAG, voor de jaren 1975 en 1977.

BAG	1975		1977	
	M	V	M	V
<0,2	.275	.265	.236	.233
0,2-0,5	.268	.199	.251	.280
0,5-1,0	.317	.229	.286	.273
>1,0	.372	.556	.291	.425

Tabel 3b. Gewichten afgeleid uit de percentages onderzochte bestuurders ten opzichte van het aantal passerende bestuurders.

De gewichten zorgen in dit geval voor een correctie van de aantallen op grond van het percentage van de weggebruikers dat in de steekproef is opgenomen. De samenhang tussen de gewichten en de BAG-klassen is een gevolg van het feit dat in de laat-nachtelijke uren een hoger alcoholgebruik wordt geconstateerd, terwijl dan tevens bijna alle weggebruikers die de onderzoekplaats passeren kunnen worden onderzocht. Voor de vroegere uren is het percentage onderzochten geringer.
Door de aantallen te delen door de gewichten worden gecorrigeerde aantallen verkregen waarop de analyse dient te worden toegepast. Met behulp van het computerprogramma "WPM" (De Leeuw & Oppe, 1976), zijn de parameters van het log-lineaire model berekend voor deze gewogen aantallen. Van deze parameters, die asymptotisch normaal verdeeld zijn met gemiddelde 0, is tevens een standaardscore berekend die getoetst kan worden tegen de gebruikelijke grenswaarden die gelden voor de standaard normaal verdeling: $\pm$ 1,96 voor het 5% significantieniveau en $\pm$ 2,58 voor het 1% niveau. Naast deze parameters zijn Chi-kwadraatwaarden berekend voor groepen van parameters overeenkomstig de mogelijke bronnen van interactie. In Tabel 4 zijn de diverse interactiebronnen gegeven, met hun Chi-kwadraatwaarden en vrijheidsgraden. Tevens zijn de parameters en bijbehorende standaardscores opgenomen.

Interactiebron	Chi-kwadraat	df	Contrast	Parameter	Standaard score
jaar x geslacht	0.21	1	1 1 0	0.08	0.46
jaar x BAG	1.70	3	1 0 1	0.06	0.58
			1 0 2	0.19	1.22
			1 0 3	-0.05	-0.24
geslacht x BAG	102.71	3	0 1 1	-1.00	-10.10
			0 1 2	-0.58	-3.65
			0 1 3	-0.36	-1.69
jaar x geslacht x BAG	4.04	3	1 1 1	0.07	0.72
			1 1 2	-0.20	-1.28
			1 1 3	0.12	0.55

Tabel 4. Analyse van rij- en drinkgewoontegegevens, uitgesplitst naar jaar, geslacht en bloedalcoholgehalte (BAG).

Bij de interpretatie van de X^2-waarden wordt eerst gekeken naar de hoogste-orde-interactie. Indien deze significant is, dan zijn de andere effecten niet goed meer te interpreteren, omdat ze door het hogere-orde-effect beïnvloed worden.
We zien dat de Chi-kwadraatwaarde van de interactie voor jaar x geslacht x BAG niet significant is (X^2 = 4.04, df = 3). De verhouding tussen het drinken van mannen en vrouwen is dus niet veranderd in 1977 t.o.v. 1975. Indien deze interactie wel significant zou zijn geweest, dan zou een verdere analyse van de relatie tussen BAG en geslacht apart voor 1975 en 1977 dienen te geschieden, of van de relatie tussen BAG en jaar apart voor mannen en vrouwen. Hier is dit niet nodig. Er is ook geen significant verschil te constateren in drinken tussen 1975 en 1977 (X^2 = 1.70). Wel vinden we een zeer significant verschil tussen het BAG van mannen en vrouwen over de beide jaren gezamenlijk bekeken (X^2 = 102.71). De rij- en drinkgewoonten van mannen verschillen zeer significant van die van vrouwen.
Indien we meer informatie wensen over de bijdrage van de verschillende klassen aan de interactie-effecten, dan is dit voor jaar en geslacht direct duidelijk. Er is slechts een contrast: klasse 1 vs klasse 2. Het effect wordt dus veroorzaakt door het verschil tussen de twee klassen.
Voor het BAG is dit niet zo duidelijk. We kunnen ons afvragen of het effect te maken heeft met het drinken versus het niet-drinken, of eerder is toe te schrijven aan het meer gaan drinken van de drinkers, of misschien aan beide factoren. Om dit te onderzoeken kunnen we de volgende contrasten tussen de BAG-klassen kiezen:
contrast 1: klasse 1 vs klasse 2 en 3 en 4 (niet drinken vs drinken)
contrast 2: klasse 2 vs klasse 3 en 4 (weinig drinken vs matig of veel drinken)
contrast 3: klasse 3 vs klasse 4 (matig vs veel drinken).
Deze contrasten zijn onafhankelijk van elkaar te interpreteren en leveren alle informatie ten aanzien van het interactie-effect.
Voor BAG zijn ook andere contrasten mogelijk, bijvoorbeeld
contrast 4: klasse 1 en 2 vs klasse 3 en 4 (geoorloofd vs ongeoorloofd BAG)

contrast 5: klasse 1 vs klasse 2
contrast 6: klasse 3 vs klasse 4.
Ook deze contrasten zijn onafhankelijk van elkaar, maar niet van de eerder genoemde contrasten. Ze leveren eveneens alle informatie. De keuze van de contrasten wordt dus vooraf bepaald door de onderzoeker afhankelijk van zijn interesse. Voor de eerstgenoemde contrasten zijn de parameters berekend en vervolgens uitgedrukt in standaard-scores.
Voor het interactie-effect geslacht x BAG betreft dit bijvoorbeeld 1 x 3 = 3 parameters. Eén parameter voor elke combinatie van contrasten. Ook deze zijn gegeven in Tabel 4.
We zien dat de grootste parameter wordt gevonden voor combinatie 0 1 1 (z = -10.10), daarna voor 0 1 2 (z = -3.65), tenslotte voor 0 1 3 (z = -1.69). Het teken geeft de richting van het effect aan: mannen drinken vaker, meer en mogelijk zwaarder dan vrouwen.
Uit deze parameters kunnen we tevens parameters afleiden voor de klassen zelf. We vinden dan 2 x 4 klasseparameters. Deze zijn echter wel afhankelijk van elkaar. De parameters voor de BAG-categorieën van mannen en vrouwen hebben een tegengesteld teken. Verder ligt de laatste parameter vast met de eerste drie parameters: de parameters tellen op tot nul.
In Afbeelding 11 zijn deze parameters afgebeeld.

Tenslotte kunnen we nog zoeken naar het meest efficiënt beschrijvende model voor de gegevens. Dit blijkt uit bovenstaande gegevens het model te zijn waarin de interacties jaar x geslacht x BAG, jaar x BAG en jaar x geslacht afwezig worden verondersteld. In formulevorm:

$$E(\ln X_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + \delta_{jk}$$

Voor dit model wordt een X^2-waarde gevonden van 13.10 (df = 7).
Tot zover de bespreking van de resultaten van de analyse van de tabel.

Tenslotte kan nog worden opgemerkt dat de lineaire modellen die ten grondslag liggen aan HOMALS en CANALS weliswaar modellen zijn

die lineair zijn in de parameters, maar daaruit mag niet worden geconcludeerd dat de onderzoekvariabelen perse lineair moeten samenhangen. Via schaling van de kenmerken is het mogelijk om een verondersteld multiplicatief model ten aanzien van de onderzoekvariabelen door een log-transformatie als lineair model te presenteren. Het bovengenoemde "additief conjunct meetmodel" kan als zodanig worden opgevat. De oplossing kan ook worden gevonden door de variabelen met behulp van CANALS te analyseren en hierbij de onafhankelijke variabelen als nominaal te beschouwen en de afhankelijke variabele als ordinaal. Hierin komt een zekere verwantschap tussen de gegeneraliseerde lineaire modellen op deze wijze toegepast en de log-lineaire analysetechnieken duidelijk tot uitdrukking.

LITERATUUR

Andersen, E.B. (1977). Multiplicative Poisson models with unequal cell rates. Scand. J. Statist. 4.

Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). Discrete Multivariate Analysis; Theory and Practice. MIT-Press, London, 1975.

Cochran, W.G. (1952). The X^2-test of goodness of fit. Ann. of Math. Statist. (23).

Darroch, J.N. (1974). Multiplicative and additive interaction in contingency tables. Biometrika, 1974, p. 207.

Datatheorie R.U. Leiden (1980). Niet-lineaire multivariate analyse. Gifi, Leiden, 1980.

De Leeuw, J. & Oppe, S. (1976). Analyse van kruistabellen: Log-lineaire Poisson modellen voor gewone aantallen. R-76-8. SWOV, Voorburg, 1976.

Foldvary, L.A. & Lane, J.C. (1974). The effectiveness of compulsory wearing of seat belts in casualty reduction. Accid. Anal. Prev., 6, p. 59-81.

Gokhale, D.V. & Kullback, S. (1978). The information in contingency tables. Marcel Dekker Inc., New York, 1978.

Hamerslag, R. & Huisman, M.C. (1977). Het gebruik van het multi-proportionele schattingsmodel bij ongevallenanalyse. Ingenieurs-bureau Dwars, Heederik en Verhey, Amersfoort.

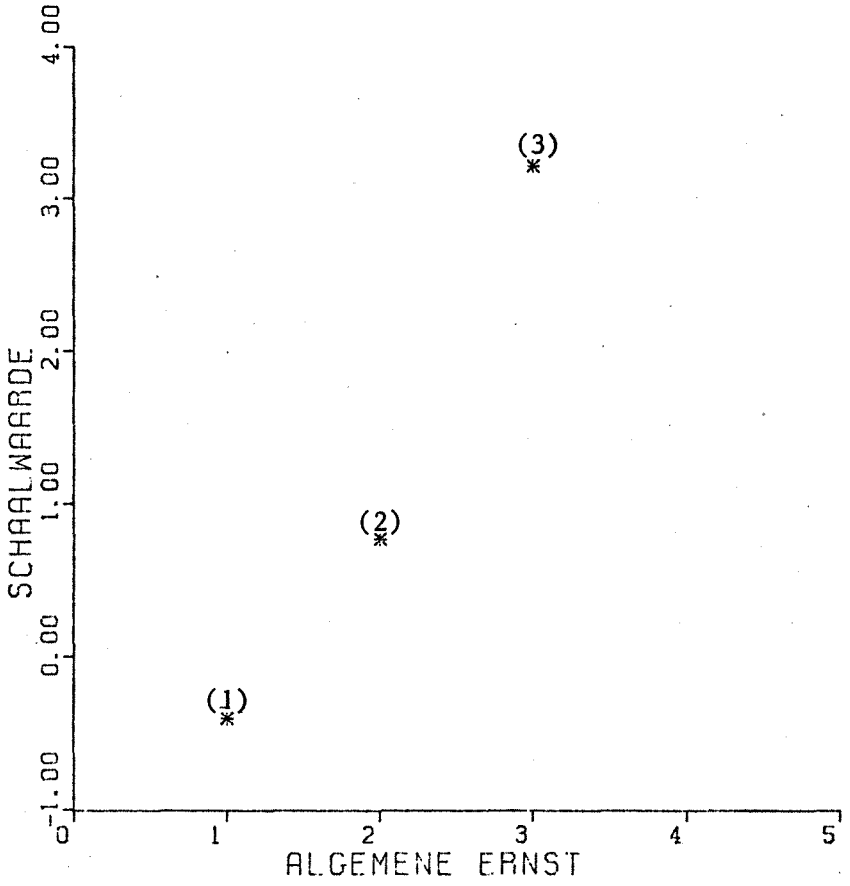
Kruskal, J.B. (1965). Analysis of factorial experiments by estimating monotone transformations of the data. J.R. Stat. Stoc., Serie B, 27, 1965.

Nelder, J.A. & Wedderburn, R.W.M. (1972). Generalized linear models. J.R. Stat. Soc., Serie A, 1972, p. 370.

Oppe, S. (1978). The use of multiplicative models for analysis of road safety data. R-78-18. SWOV, Voorburg, 1978. Ook in: Accid. Anal. Prev., 11, 1979.

Rasch, G. (1973). Two applications of the multiplicative Poisson models in road accidents statistics. In: Proc. of the 38th Session of the ISI, Wien, 1973.

SWOV (1978). Alcoholgebruik onder automobilisten. Verslag en resultaten van het onderzoek Rij- en drinkgewoonten van Nederlandse automobilisten in weekeindnachten in het najaar van de jaren 1970, 1971, 1973, 1974, 1975 en 1977. 2de herziene en uitgebreide druk. R-78-19. SWOV, Voorburg, 1978.

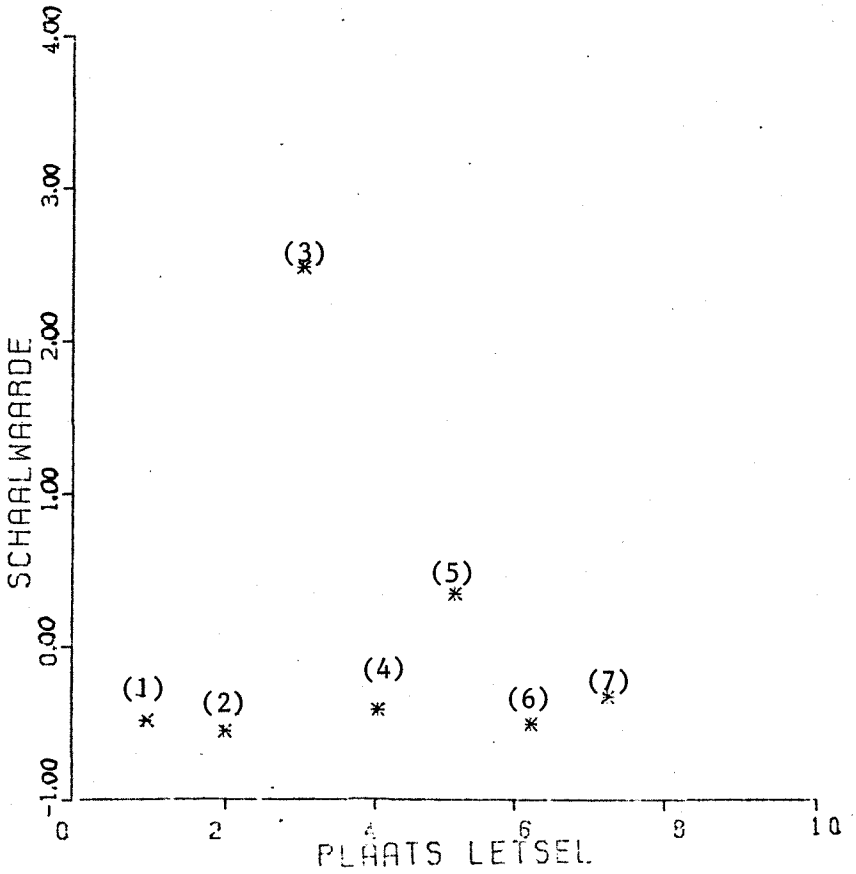


(1) licht + matig

(2) ernstig

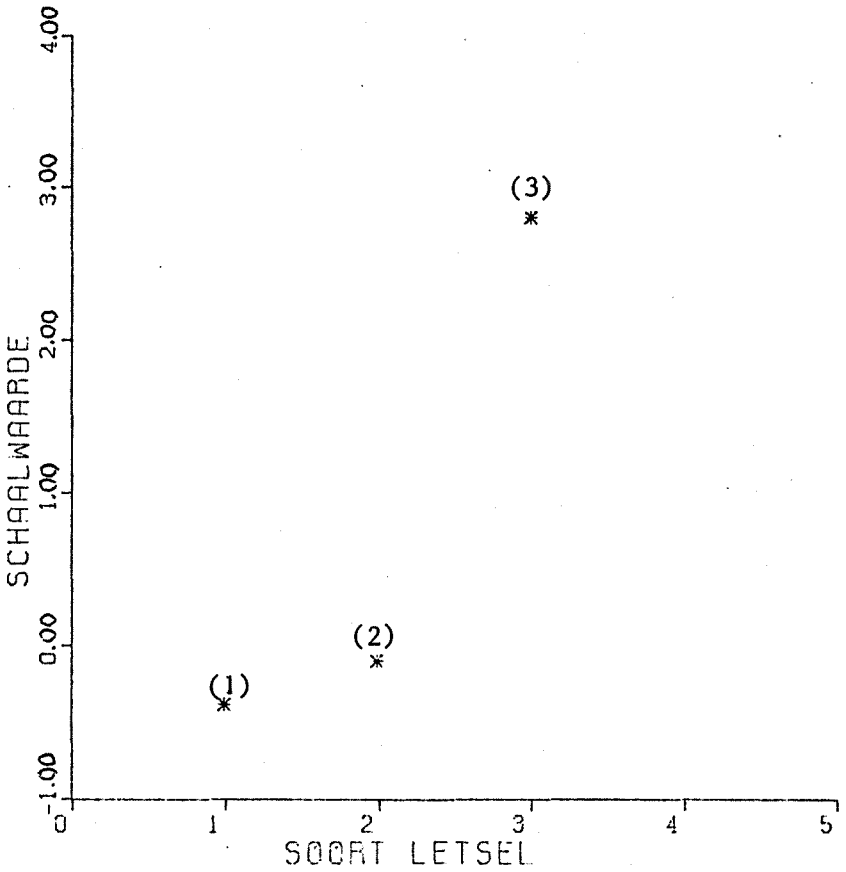
(3) zwaar + dood

Afbeelding 1.



(1) schedel + hersenen
(2) gelaat
(3) nek
(4) borst
(5) buik + rug + bekken
(6) armen
(7) benen

Afbeelding 2.

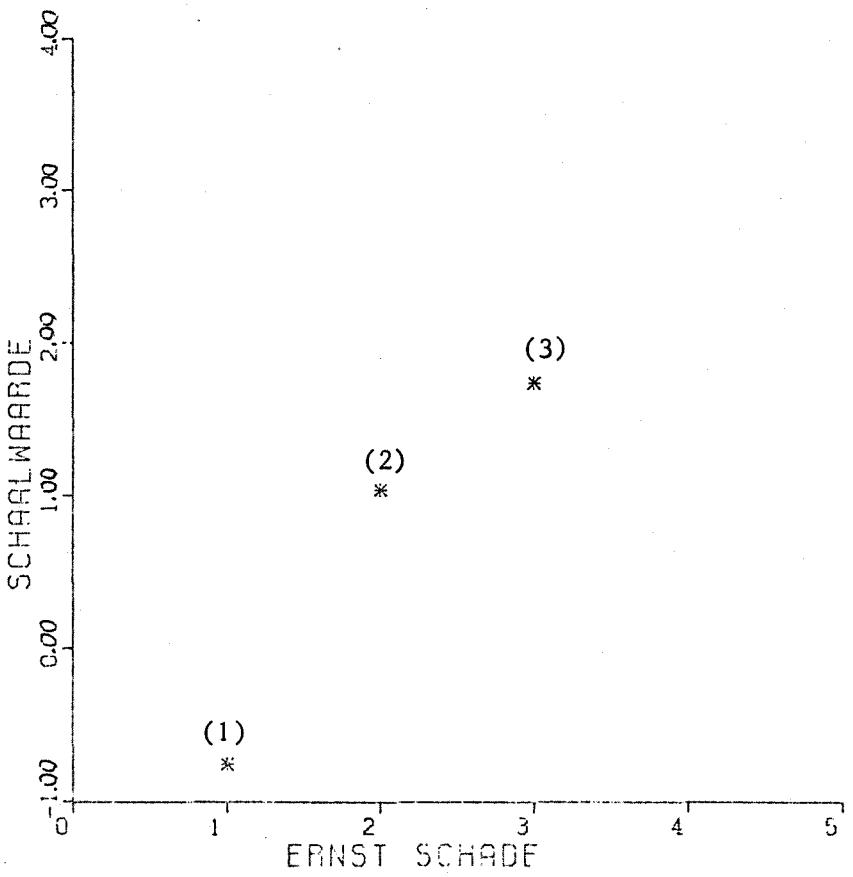


(1) wonden

(2) fracturen

(3) inwendig letsel

Afbeelding 3.

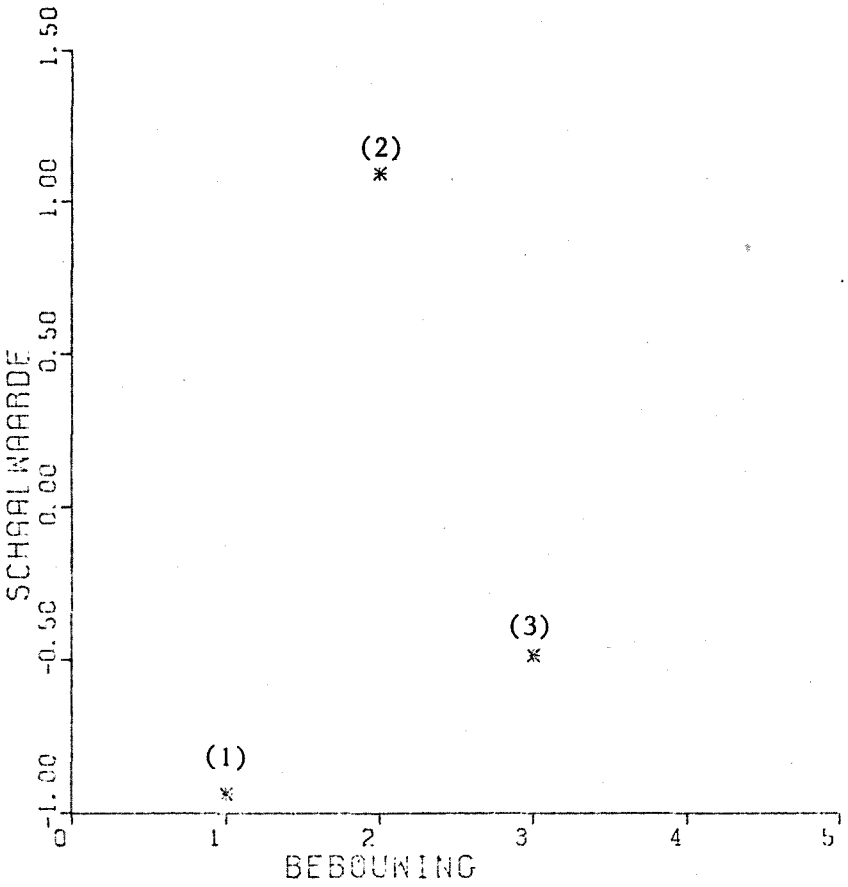


(1) lichte plaatschade

(2) lichte compartimentschade

(3) zware compartimentschade

Afbeelding 4.

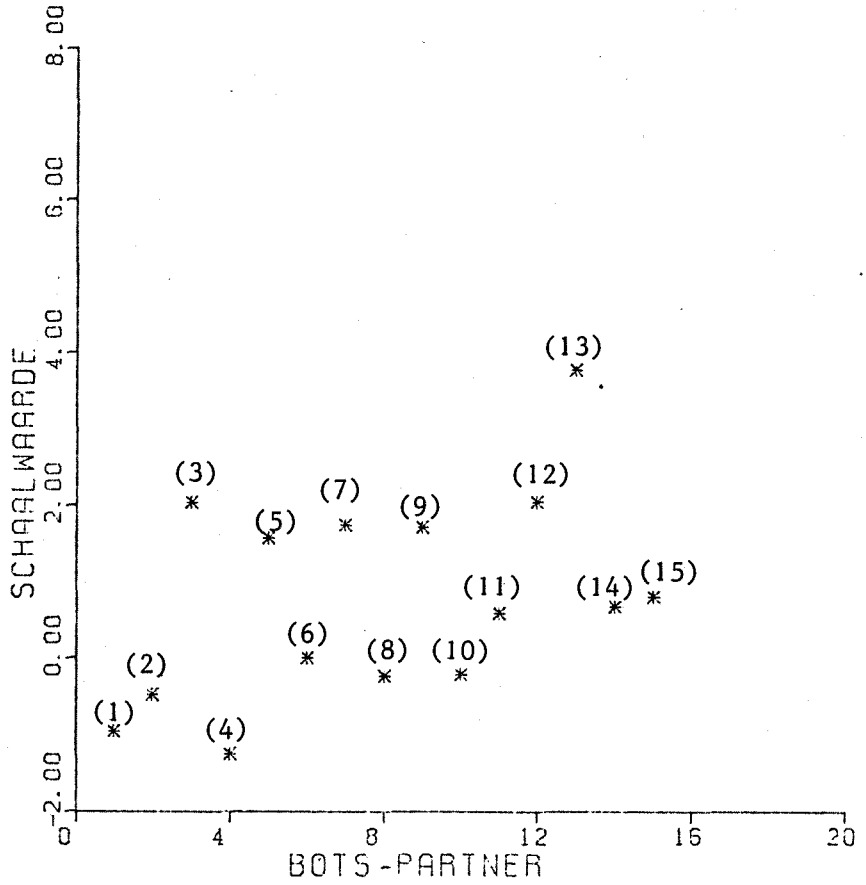


(1) binnen bebouwde kom

(2) buiten bebouwde kom

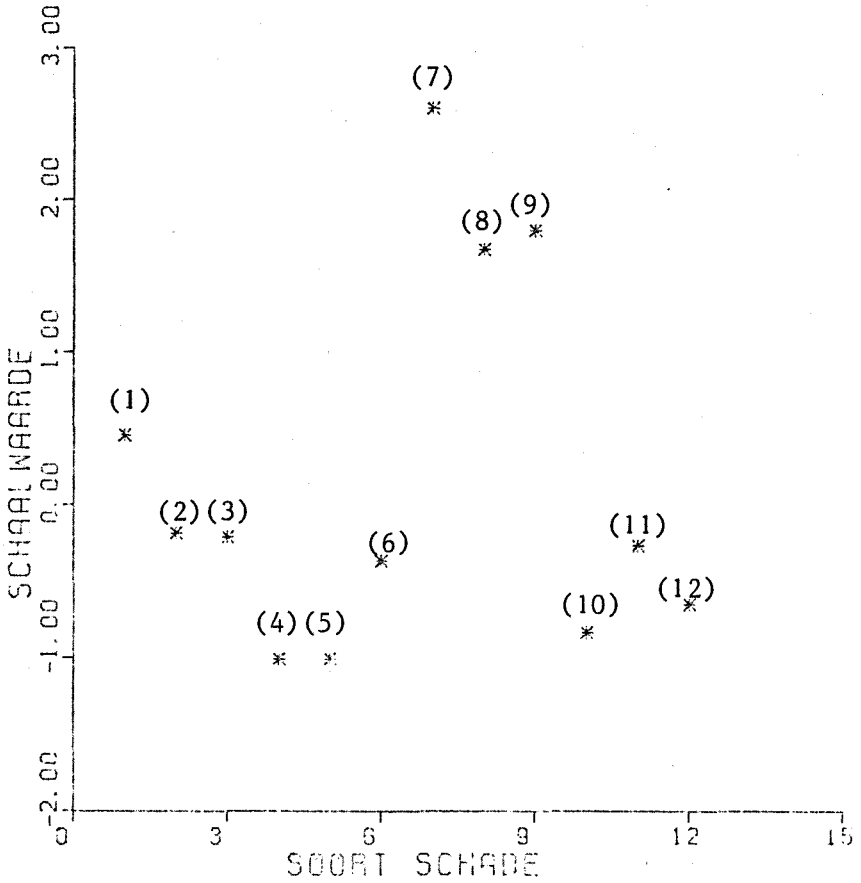
(3) onbekend

Afbeelding 5.



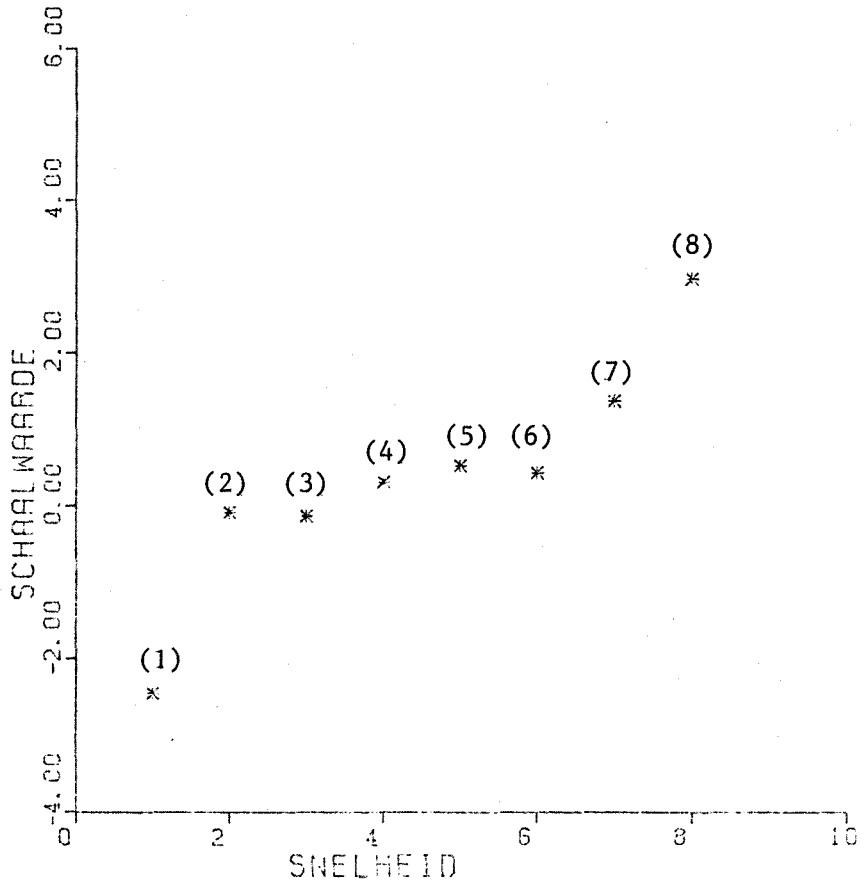
(1) geslipt
(2) personenauto
(3) vrachtauto
(4) combinatie voertuigen
(5) ander voertuig
(6) lichte verkeersdeelnemers
(7) boom + paal
(8) geleiderail
(9) pers.auto + paal
(10) pers.auto + geleiderail
(11) pers.auto + voorwerp
(12) vrachtauto + paal/geleiderail
(13) vrachtauto + voorwerp
(14) overig
(15) onbekend

Afbeelding 6.



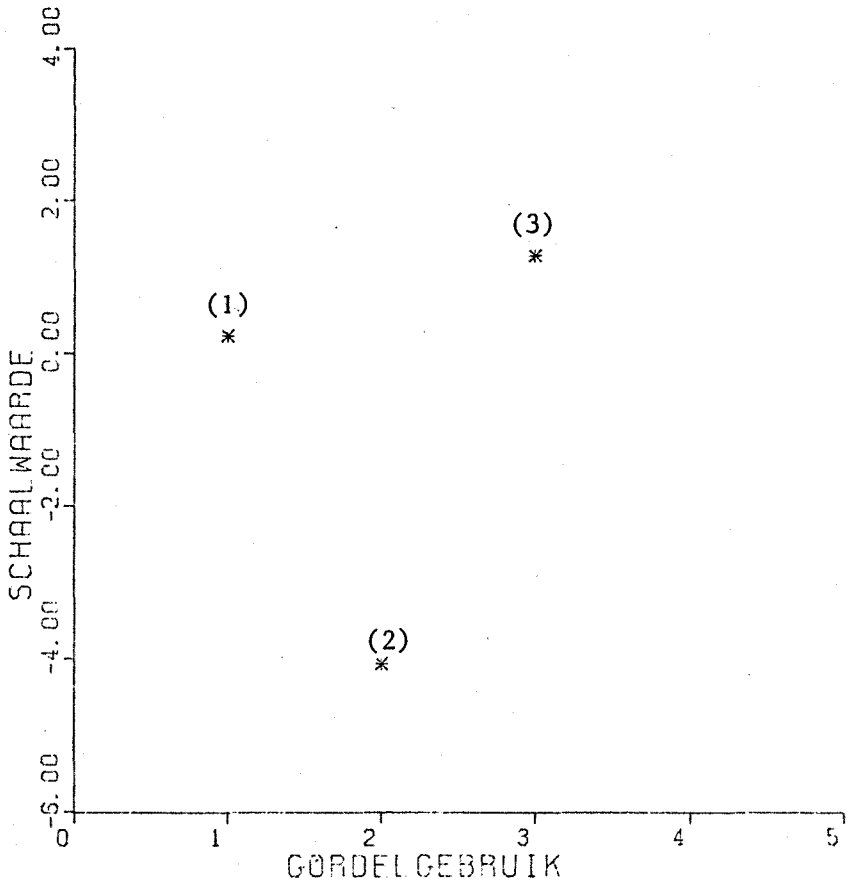
(1) licht
(2) zuiver front
(3) overig front
(4) linker flank
(5) rechter flank
(6) overig flank
(7) zuiver achter
(8) achter + voor
(9) overig achter
(10) over de kop
(11) anders
(12) onbekend

Afbeelding 7.



(1) 0 km/u
(2) 1-25 km/u
(3) 26-50 km/u
(4) 51-75 km/u
(5) 76-100 km/u
(6) 101-125 km/u
(7) >125 km/u
(8) onbekend

Afbeelding 8.

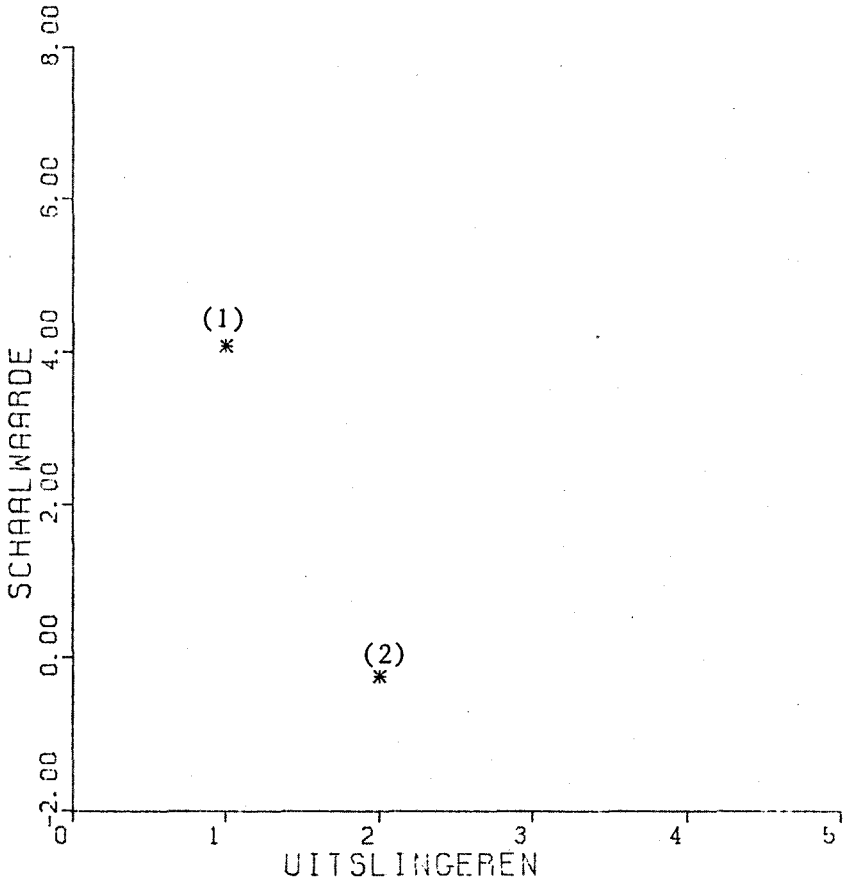


(1) geen

(2) wel

(3) onbekend

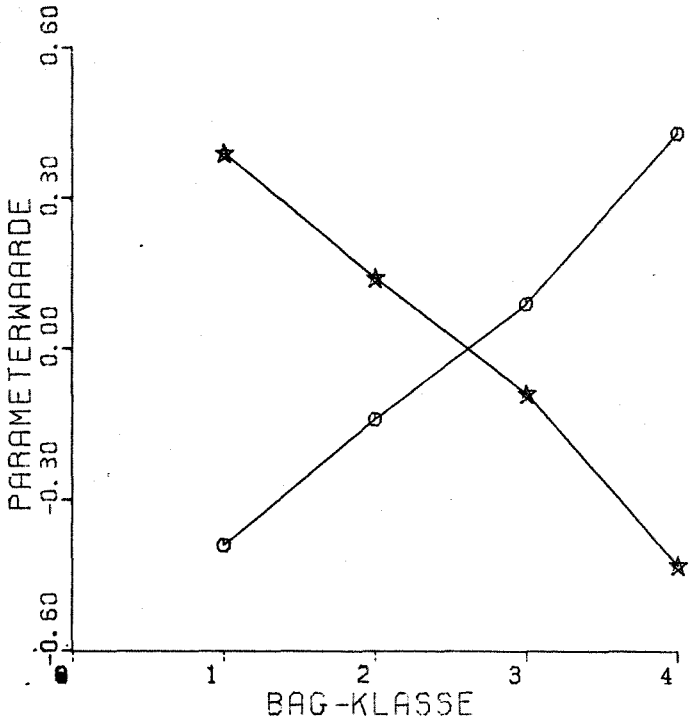
Afbeelding 9.



(1) wel

(2) niet

Afbeelding 10.



● : mannen ★ : vrouwen

Afbeelding 11. De interactie-parameters voor het effect: geslacht x BAG.