HOOFDSTUK 3

Datatheorie

3.1 Nabijheid en afstand

De gegevens die in het vorige hoofdstuk met behulp van MDS geanalyseerd werden, waren *afstandsgegevens* in de letterlijke zin van het woord. In dit geval waren het de geografische afstanden in kilometers tussen een aantal Nederlandse steden. De term afstandsgegevens hoeft echter niet uitsluitend op echte, objectieve afstanden betrekking te hebben, het kan ook gaan om afstanden in een meer overdrachtelijke zin. Bijvoorbeeld: twee personen hebben een kleine afstand als zij elkaar aardig vinden en/of veel met elkaar omgaan (veel interactie met elkaar hebben). Of: voor iemand die op de Partij van de Arbeid stemt is er een kleinere afstand tussen hem en de PvdA dan tussen hem en het CDA. Ook: twee 'objecten' (bijvoorbeeld kleuren, of politieke partijen) hebben een kleinere afstand tot elkaar naarmate ze meer op elkaar lijken of gemakkelijker met elkaar verward worden.
In bovenstaande voorbeelden gaat het dus om allerlei soorten *objecten* en om gegevens over de *relaties* tussen die objecten. Met behulp van MDS willen we deze objecten als punten *afbeelden in een ruimtelijk model*, dat wil zeggen, we willen ze afbeelden door middel van coördinaten op een verzameling dimensies. Er is hier dus sprake van twee systemen: een *empirisch systeem* van objecten en hun onderlinge relaties (zoals onderlinge gelijkenis, preferentie, interactie, verwarring) en een *ruimtelijk model* waarin de objecten als punten worden afgebeeld. In de afbeelding die gezocht wordt, komen de afstanden tussen de punten-in-het-model qua grootte zoveel mogelijk overeen met de sterkte van de geobserveerde relaties tussen de objecten.
Elk soort gegeven over de onderlinge relaties tussen objecten kan in principe opgevat worden als een indicatie voor de afstand tussen die objecten in een

ruimtelijke afbeelding of kan daartoe herleid worden. Zulk soort gegevens noemen we in het algemeen *nabijheidsdata* (*proximities*). Twee in de sociale wetenschappen veelvuldig voorkomende soorten nabijheidsdata zijn *gelijkenisgegevens* (gelijkenissen, similarities, dissimilarities) en *preferenties*. Gelijkenissen en preferenties hebben een inverse relatie tot de afstanden in een model: hoe groter de gelijkenis of preferentie, des te kleiner de afstand. Dissimilarities, daarentegen, hebben een positieve relatie met afstanden. Om empirisch geobserveerde nabijheidsgegevens aan te duiden zullen we, in navolging van Young (1987), het symbool *O* gebruiken: de geobserveerde nabijheid van object *i* en object *j* is dan o_{ij}. De afstanden tussen de punten *i* en *j* in de uiteindelijke afbeelding worden aangeduid met d_{ij}.

3.2 DE DATATHEORIE VAN COOMBS

In het algemeen wordt er geen onderscheid gemaakt tussen de termen gegevens, observaties en data. Dit werd wel gedaan door Coombs (1964). Observaties worden pas *data* doordat ze geïnterpreteerd worden in termen van een model. Dezelfde observaties kunnen namelijk in termen van verschillende modellen geïnterpreteerd worden en daardoor tot verschillende klassen van data behoren. Coombs heeft een typologie van data ontworpen die uit acht categorieën bestaat. Deze indeling ontstaat uit de kruising van drie dichotomieën (tweedelingen), die betrekking hebben op de volgende drie vragen:

1 Gaat het bij de observaties om de paarsgewijze *relaties tussen punten* (bijvoorbeeld A lijkt op B) of om de paarsgewijze *relaties tussen paren van punten* (bijvoorbeeld A en B lijken meer op elkaar dan C en D)?
2 Gaat het bij de observaties om punten of puntenparen uit *één en dezelfde verzameling* of om punten (c.q. tweetallen van punten) uit *twee verschillende verzamelingen*? Bijvoorbeeld: de observatie 'Van Mierlo overlegt met Kok' betreft twee punten uit één verzameling, namelijk de verzameling van Nederlandse politici. De observatie 'Kok is lid van de PvdA' betreft twee punten uit twee verschillende verzamelingen: politici en politieke partijen.
3 Gaat het bij de observaties om een *dominantierelatie* of om een *nabijheidsrelatie*?[1] De observatie 'Kok is lid van de PvdA' is een nabijheidsrelatie: Kok en de PvdA liggen dicht bij elkaar. Anders gezegd: de afstand tussen Kok en de PvdA is kleiner dan een bepaalde kritische waarde ε. Was die afstand groter dan de drempelwaarde ε, dan zou Kok geen lid zijn van de PvdA. De observatie 'D66 heeft meer voorkeur voor de PvdA dan voor het CDA' is een dominantierelatie. Deze observatie impliceert dat de afstand tussen D66 en het CDA groter is dan de afstand tussen D66 en de PvdA. De afstand D66-CDA *domineert* dan

1 Coombs (1964; p. 19) sprak van *order relations* versus *proximity relations*; als alternatief voor order relations noemde hij de hierboven gebruikte term *dominance relations*.

de afstand D66-PvdA. Opgemerkt moet worden dat de term nabijheidsrelaties (bij Coombs *proximity relations*) een veel specifiekere betekenis heeft dan de term nabijheidsdata (*proximities*) zoals we die in Hoofdstuk 1 en in het begin van dit hoofdstuk hebben geïntroduceerd. In dit boek is 'nabijheidsdata' een algemene term voor allerlei soorten data die op wat voor manier dan ook iets zeggen over de (ruimtelijke) relaties tussen objecten. De term nabijheidsrelatie uit Coombs' datatheorie is gereserveerd voor een heel specifiek geval van nabijheidsdata: namelijk voor een observatie die impliceert dat een bepaalde afstand kleiner is dan een drempelwaarde ε.

	paren punten	paren tweetallen
twee verzamelingen	IIa	Ia
	IIb	Ib
één verzameling	IIIa	IVa
	IIIb	IVb

a = dominantierelatie; b = nabijheidsrelatie

Figuur 3.1 De vier kwadranten van Coombs' datatheorie

De typologie van Coombs is weergegeven in het schema dat is afgebeeld in Figuur 3.1. Een toelichting wordt gegeven in Tabel 3.1. Daarin wordt voor ieder kwadrant van het Coombsiaanse schema een voorbeeld gegeven van twee verbaal geformuleerde observaties met hun bijbehorende interpretaties in termen van nabijheids- of dominantierelaties. In de rechterkolom van Tabel 3.1 staat in formulevorm de informatie die elke observatie oplevert over de schaalwaarden (θ_i, θ_j, enzovoort) van de objecten of over de afstanden (d_{ij}, d_{ik}, enzovoort) tussen de objecten. Bijvoorbeeld: de uitspraak 'spruitjes zijn lekkerder dan andijvie' kan worden uitgelegd als 'de schaalwaarde θ_s van spruitjes op een lekkerheidsschaal is groter dan de schaalwaarde θ_a van andijvie' (Kwadrant IIIa). De uitspraak 'John houdt meer van spruitjes dan van andijvie' wordt daarentegen geïnterpreteerd als 'de afstand tussen John en spruitjes is kleiner dan de afstand tussen John en andijvie' (Kwadrant Ia).

Tabel 3.1 Voorbeelden van observaties en hun interpretatie

Kwadrant	Observatie en *interpretatie*	Informatie
Ia	John houdt meer van spruitjes dan van andijvie. *De afstand tussen John en spruitjes is kleiner dan de afstand tussen John en andijvie.*	$d_{js} < d_{ja}$
Ib	John heeft een voorkeur wanneer hij moet kiezen tussen spruitjes en andijvie. *De afstand tussen John en spruitjes verschilt (meer dan de drempelwaarde) van de afstand tussen John en andijvie.*	$\mid d_{js} - d_{ja} \mid > \varepsilon_j$
IIa	Marja vindt andijvie lekker (genoeg om te eten). *De 'lekkerheid' van andijvie is groter dan de door Marja gestelde minimumwaarde voor 'lekkerheid'.*	$\theta_a > \theta_m$
IIb	Heleen vindt spruitjes lekker. *De afstand tussen Heleen en spruitjes is klein genoeg; kleiner dan de drempelwaarde om ze lekker te vinden.*	$d_{hs} < \varepsilon_h$
IIIa	Andijvie is bitterder dan spruitjes. *De bitterheid van andijvie is groter dan de bitterheid van spruitjes.*	$\theta_a > \theta_s$
IIIb	Spruitjes en andijvie zijn even groen. *Het verschil tussen de 'groenheid' van spruitjes en de 'groenheid' van andijvie is kleiner dan de drempelwaarde om verschillende tinten groen te kunnen onderscheiden.*	$\mid \theta_s - \theta_a \mid < \varepsilon$
IVa	Spruitjes en andijvie lijken meer op elkaar dan doperwtjes en lof. *De afstand tussen spruitjes en andijvie is kleiner dan de afstand tussen doperwtjes en lof.*	$d_{sa} < d_{dl}$
IVb	Spruitjes en andijvie verschillen evenveel van elkaar als doperwtjes en lof. *De afstand tussen spruitjes en andijvie is even groot als de afstand tussen doperwtjes en lof (het verschil tussen beide afstanden is kleiner dan de drempelwaarde).*	$\mid d_{sa} - d_{dl} \mid < \varepsilon$

3.3 HET SYSTEEM VAN CARROLL, ARABIE EN YOUNG

Hoewel Coombs' datatheorie conceptueel van groot belang is (hij liet immers zien hoe men allerlei soorten observaties kan 'vertalen' in ruimtelijke termen), is zijn indelingsschema niet erg praktisch. Dat komt omdat er geen unieke relatie bestaat tussen de kwadranten van Figuur 3.1 en de verschillende schaaltechnieken die er voor de analyse van nabijheidsgegevens beschikbaar zijn. Het zou handiger zijn observaties of data zodanig in te delen dat bij iedere categorie een bepaalde klasse van schaaltechnieken hoort. Zo'n systeem is de hieronder te behandelen indeling met betrekking tot de *vorm* van de datamatrix. Deze indeling is voorgesteld door Carroll en Arabie (1980) en is inmiddels algemeen geaccepteerd. Young (1984) heeft deze indeling uitgebreid door ook de *meetkenmerken* van de data erbij te betrekken. Ook dat aspect wordt hieronder gedetailleerd besproken.

De vorm van de datamatrix

De hieronder volgende classificatie van data en schaalproblemen berust in eerste instantie op een indeling van de verzamelde afstandsgegevens met betrekking tot twee aspecten: het aantal *wegen* en het aantal *modi* van een tabel met nabijheidsgegevens.

Het aantal wegen. Een afstandentabel of nabijheidsmatrix heeft op zijn minst twee wegen: horizontaal een aantal rijen en verticaal een aantal kolommen. Met iedere rij en iedere kolom correspondeert een bepaald *object* (een stimulus, een persoon) en op het kruispunt van een bepaalde rij en een bepaalde kolom staat de 'nabijheid' die er tussen het rij-object en het kolom-object bestaat. Een voorbeeld van zo'n tweewegmatrix is Tabel 2.1 die de afstanden in kilometers tussen acht Nederlandse steden laat zien. In dit geval is de nabijheidsmatrix vanzelfsprekend *vierkant* en *symmetrisch*. Vierkant, omdat de rijen op dezelfde steden betrekking hebben als de kolommen. Symmetrisch, omdat de geobserveerde afstand o_{ij} tussen stad i en stad j gelijk moet zijn aan de geobserveerde afstand o_{ji}.
Een tweede voorbeeld van een tweewegmatrix is een tabel waarvan de rijen en kolommen op verschillende kleuren betrekking hebben en waarvan iedere cel (i,j) aangeeft hoeveel iemand de kleur van kolom j vindt lijken op de kleur die bij rij i hoort. In dit tweede geval is de matrix wel vierkant, maar hoeft hij niet altijd symmetrisch te zijn. Het oordeel van een proefpersoon kan namelijk afhangen van de kleur die het eerst wordt aangeboden; in dat geval is o_{ij} dus niet noodzakelijk gelijk aan o_{ji}.
Een derde voorbeeld van een tweewegmatrix is een tabel waarvan de rijen met verschillende personen corresponderen en de kolommen met verschillende activiteiten (bijvoorbeeld werken, lezen, sporten, huishouden). In iedere cel van de matrix staat nu hoeveel tijd de persoon van rij i per week besteedt aan de activiteit die bij kolom j hoort. Dit type matrix wordt *rechthoekig* genoemd, zelfs al zijn er in een bepaald geval evenveel personen als activiteiten. Als het

aantal rijen niet gelijk is aan het aantal kolommen kan de matrix nooit symmetrisch zijn. Alleen als de matrix evenveel rijen en kolommen heeft, zou hij bij toeval symmetrisch kunnen zijn; doorgaans zal dat echter niet het geval zijn.
Naast bovengenoemde tweewegmatrices met nabijheidsgegevens zijn er ook driewegmatrices met zulke gegevens te verkrijgen, bijvoorbeeld wanneer men de onderlinge gelijkenis van een verzameling kleuren door meerdere personen laat beoordelen. Voor iedere persoon hebben we dan een individuele tweewegmatrix van kleuren bij kleuren. Door die matrices als de plakjes van een cake achter elkaar te zetten, ontstaat een driewegmatrix, met de personen als de derde weg. Een andere driewegmatrix ontstaat als we personen een aantal politieke partijen laten beoordelen op een aantal kenmerken. Zouden we dit onderzoek enkele maanden later herhalen (met dezelfde personen, partijen en kenmerken) dan ontstaat een vierwegmatrix, waarbij de tijd de vierde weg vormt.

Het aantal modi. In het laatste voorbeeld zien we dat elk van de vier wegen van de nabijheidsmatrix overeenkomt met een andere categorie objecten. Anders gezegd: de objecten van de verschillende wegen zijn uit vier verschillende verzamelingen afkomstig, respectievelijk, de verzamelingen personen, partijen, kenmerken en tijdstippen. Elk van deze verzamelingen is een *modus* van de nabijheidsmatrix. In dit voorbeeld is er dus sprake van vier modi, en omdat er vier wegen zijn, spreken we van vierweg/viermodale data. Dit soort data komt niet zoveel voor.[2] Voorbeelden van data die wèl vaak voorkomen (zie ook Tabel 3.2), zijn:

– *tweeweg/éénmodale data*, bijvoorbeeld de onderlinge gelijkenissen van objecten uit één en dezelfde verzameling (bijvoorbeeld kleuren). Afhankelijk van het soort observaties, kunnen zulke data symmetrisch zijn of asymmetrisch;
– *drieweg/tweemodale data*, bijvoorbeeld de onderlinge gelijkenissen van objecten uit één verzameling, geobserveerd onder verschillende condities, bij verschillende personen, of op verschillende tijdstippen. Er is hier dus sprake van verschillende tweeweg/éénmodale matrices die als de plakjes van een cake achter elkaar zijn gezet. Elk van de 'plakjes' kan op zichzelf symmetrisch zijn of niet;
– *tweeweg/tweemodale data*, bijvoorbeeld de door verschillende personen gegeven rangordeningen van hun voorkeuren voor een aantal objecten;
– *drieweg/driemodale data*, bijvoorbeeld de beoordeling van een aantal objecten op verschillende kenmerken door meerdere personen, of de voorkeursrangordeningen van personen voor objecten op verschillende tijdstippen.

2 Wel mogelijk maar ook niet veel voorkomend, zijn drieweg/éénmodale data; bijvoorbeeld schattingen van de kans dat de gebeurtenissen A én B én C tegelijkertijd voorkomen.

Replicaties. Soms gebruikt men voor de elementen van een bepaalde modus de term replicaties. In dat geval is men niet geïnteresseerd in de verschillen tussen de elementen van die modus; men beschouwt die elementen alsof ze onderling verwisselbaar zijn. Bijvoorbeeld: men vraagt een aantal proefpersonen elk afzonderlijk de onderlinge gelijkenissen van een aantal stimuli te beoordelen. Voor elke proefpersoon krijgt men dan een tweeweg/éénmodale matrix met gelijkenissen. Zet men die matrices achter elkaar, dan ontstaat een drieweg/tweemodale matrix met de proefpersonen als derde weg en tweede modus. Is men geïnteresseerd in de eventuele verschillen tussen de – in principe unieke – proefpersonen, dan zal men een type MDS-analyse uitvoeren dat met zulke individuele verschillen rekening houdt (men spreekt dan van *individual differences scaling*). Is men niet geïnteresseerd in zulke verschillen, dan vat men de proefpersonen op als elementen uit een (aselecte) steekproef die – in principe – voor elkaar gesubstitueerd kunnen worden. De proefpersonen worden dan als replicaties van elkaar beschouwd; hun individuele eigenschappen hoeven dus niet als afzonderlijke parameters in de schaaloplossing te worden gemodelleerd. In dat geval noemt men drieweg/tweemodale data ook wel tweeweg/éénmodale data met replicaties. Deze term replicaties zegt niets over de vorm van de data, maar zegt wel iets over de betekenis die men aan de betreffende modus toekent (wat implicaties kan hebben voor de methoden die men kan gebruiken om de data te analyseren). Hierbij gaat het om de vraag of men *alle* elementen van *alle* modi afzonderlijk wil afbeelden, of dat men een afbeelding wil waarin de gegevens van een van de modi op een bepaalde manier worden samengenomen.

Tabel 3.2 Voorbeelden van matrices met onderzoeksdata

tweeweg/éénmodaal:	paarsgewijze beoordeling van gelijkenis tussen 23 automobielen
tweeweg/tweemodaal:	beoordeling van 14 geuren op zes schalen door 1 proefpersoon
drieweg/tweemodaal:	paarsgewijze beoordeling van de onderlinge gelijkenis van 23 automobielen door 14 proefpersonen (eventueel: tweeweg/éénmodaal met 14 replicaties)
drieweg/driemodaal:	beoordeling van 20 tonen op 8 schalen door 12 proefpersonen (eventueel: drieweg/tweemodaal met 12 replicaties, of drieweg/tweemodaal met 20 replicaties, of drieweg/tweemodaal met 8 replicaties!)

Symmetrie. Zoals al eerder is opgemerkt, maakt het uit of een tweeweg/éénmodale nabijheidsmatrix (die onderdeel kan zijn van een drieweg/tweemodale matrix) symmetrisch is of niet. Aangezien de afstand tussen object P en Q

identiek is aan de afstand tussen Q en P zou een matrix met afstandsdata in principe symmetrisch moeten zijn. In de praktijk echter kan men allerlei nabijheidsmaten gebruiken die allerminst symmetrisch hoeven te zijn. Ook bij subjectieve beoordelingen van gelijkenis kan het voorkomen dat iemand de gelijkenis tussen *i* en *j* anders beoordeelt dan die tussen *j* en *i*, zodat $o_{ij} \neq o_{ji}$. Wil men dit probleem vermijden, dan kan men beter methoden van dataverzameling gebruiken die van tevoren uitsluiten dat men asymmetrische afstandsdata kan krijgen. Gebruikt men dataverzamelingsmethoden die wel asymmetrische data kunnen opleveren, dan moet men daarmee in de analyse rekening houden. MDS-methoden voor asymmetrische data worden in Hoofdstuk 8 besproken.

Meetkenmerken van de data

Voor de MDS-analyse van een verzameling nabijheidsdata is niet alleen de vorm van de datamatrix van belang, maar moet men ook rekening houden met de meetkenmerken van de geobserveerde gegevens, die – uiteindelijk – uit getallen bestaan. De meetkenmerken hebben betrekking op de vraag wat de precieze, numerieke betekenis van de geobserveerde getallen is. Wat drukken ze precies uit en met welke andere getallen mogen ze vergeleken worden? Hierbij gaat het om drie aspecten: het *meetniveau*, het *meetproces* en de *conditionaliteit* van de observaties.

Meetniveau. Wat betreft het niveau waarop de te analyseren nabijheidsgegevens gemeten zijn, kunnen we het bekende onderscheid maken tussen metingen op ratio-, op interval-, op ordinaal, en op nominaal niveau (zie bijvoorbeeld Meerling, 1989). Het gaat hier dus om de *veronderstelde relatie* tussen de geobserveerde nabijheden $\{o_{ij}\}$ en de 'ware' afstanden $\{d_{ij}\}$ tussen de objecten. Onder 'ware' afstanden verstaan we hier de afstanden die de objecten in het model, in de afbeelding ten opzichte van elkaar hebben. De veronderstelde relatie tussen geobserveerde nabijheden en ware afstanden kunnen we heel algemeen weergeven als $o_{ij} = h(d_{ij})$: de geobserveerde nabijheden zijn een functie van de 'echte' afstanden. Deze relatie kunnen we ook omkeren: $d_{ij} = f(o_{ij})$, dat wil zeggen, afstand is een functie van waargenomen nabijheid. Om *D* te bepalen uit *O* moeten we de nabijheden *O* dus *transformeren* volgens de functie *f*. In de praktijk zal er zelden een perfecte relatie tussen de afstanden en de observaties bestaan. We zullen genoegen moeten nemen met een *optimale transformatie* die de relatie tussen *D* en *O* zo goed mogelijk beschrijft. Dus zoeken we transformaties waarvoor geldt dat $d_{ij} \approx f(o_{ij})$, dat wil zeggen dat $f(O)$ *bij benadering* gelijk is aan *D*.

De functie *f* in $d_{ij} \approx f(o_{ij})$ kan verschillende vormen aannemen. Zijn de metingen op rationiveau, dan zoeken we transformaties van de vorm $d_{ij} \approx k.o_{ij}$; de functie *f* bestaat hier dus uit een multiplicatieve transformatie van de $\{o_{ij}\}$. Zijn de data op intervalniveau, dan willen we dat $d_{ij} \approx k.o_{ij} + c$; *f* beschrijft nu een lineaire transformatie van de $\{o_{ij}\}$. Zijn de metingen op ordinaal niveau gedaan, dan geldt voor de gezochte transformatie dat $d_{ij} \approx g(o_{ij})$ waarbij *g* een

monotoon stijgende of dalende functie van de nabijheidsdata is. Deze functie is dalend als grotere waarden van de geobserveerde gegevens een grotere mate van nabijheid (grotere gelijkenis) uitdrukken: hoe groter de nabijheid, hoe kleiner de afstand. De functie *g* is stijgend als grotere waarden van de data een kleinere mate van nabijheid (grotere dissimilarity) uitdrukken. Zijn de data op nominaal niveau gemeten, dan houdt de functie *u* uit $d_{ij} \approx u(o_{ij})$ alleen maar in dat (a) objectparen met gelijke geobserveerde nabijheidsscores gelijke onderlinge afstanden zouden moeten hebben en (b) dat objectparen met verschillende geobserveerde nabijheden verondersteld worden verschillende onderlinge afstanden te hebben. Verder kunnen we niets zeggen over welke afstanden er bij welke nabijheden horen.

Meetproces. In een verzameling nabijheidsgegevens kunnen zogenaamde *ties* voorkomen, dat wil zeggen dat voor verschillende paren van objecten gelijke nabijheidswaarden geobserveerd zijn. Enerzijds kan men nu wensen dat in de uiteindelijke MDS-afbeelding deze paren ook gelijke afstanden moeten hebben; anderzijds kan men toestaan dat deze paren niet *per se* op gelijke afstanden van elkaar hoeven te liggen. Het eerste geval noemt men de *discrete behandeling van ties* (Kruskals *secondary approach*), het tweede geval is de *continue behandeling* (Kruskals *primary approach*). De keuze tussen deze twee mogelijkheden is vooral van belang bij de analyse van ordinale nabijheidsdata; bij metrische analyses impliceren de functies $d_{ij} \approx k.o_{ij}$ en $d_{ij} \approx k.o_{ij} + c$ altijd een discrete behandeling van *ties*. Zijn de metingen op ordinaal niveau, dan krijgt men in de discrete benadering een andere functie $g(o_{ij})$ dan in de continue benadering. Op deze kwesties wordt nader ingegaan in de paragraaf over monotone transformaties.

Conditionaliteit. Iedere tweewegmatrix bestaat uit een aantal rijen, en iedere drie- of meerwegmatrix bestaat uit een aantal tweewegmatrices. Nu kan het voorkomen dat de getallen die in één stukje van de matrix staan niet zonder meer vergeleken kunnen worden met de getallen die in een ander stuk van de matrix staan. Stel dat we twee matrices hebben: de ene matrix bevat de afstanden in kilometers tussen de acht Nederlandse steden uit ons eerdere voorbeeld, terwijl de andere matrix de reistijden in minuten bevat om van de ene stad per trein naar de andere te reizen. Als we beide matrices afzonderlijk zouden analyseren, dan zouden we in beide gevallen natuurlijk een nagenoeg identieke afbeelding van de kaart van Nederland vinden. Maar hoewel de onderliggende structuur van de nabijheidsgegevens van beide matrices identiek is, zijn de getallen uit de ene matrix niet zonder meer te vergelijken met de getallen uit de andere. Immers, het getal 80 (kilometer) in de ene matrix betekent niet hetzelfde als het getal 80 (minuten) in de andere. In dit geval zegt men dat de getallen in de afzonderlijke matrices verschillende *partities* vormen; de gegevens zijn dan *matrix-conditioneel*. Alleen de getallen binnen een en dezelfde partitie kunnen zinvol met elkaar vergeleken worden. Zijn de getallen van alle matrices wèl zonder meer met elkaar te vergelijken, dan zijn de data *onconditioneel*.

Een voorbeeld van matrixconditionele data uit de psychologie is het volgende. Stel dat een drieweg/tweemodale matrix bestaat uit een aantal achter elkaar geplaatste tweeweg/éénmodale gelijkenismatrices die verkregen zijn door een aantal personen de onderlinge gelijkenis van een aantal objecten te laten beoordelen op een rating scale. Het is dan mogelijk dat alle personen in principe dezelfde onderlinge relaties tussen de objecten 'waarnemen' of 'in hun hoofd hebben', maar toch allemaal op een eigen, individuele manier getallen toekennen aan die relaties. Waar de ene persoon het getal 8 gebruikt, gebruikt een andere persoon bijvoorbeeld het getal 5 om dezelfde mate van nabijheid van een en hetzelfde paar objecten aan te geven. Ook kan het voorkomen dat de ene proefpersoon een 7 geeft aan objecten die slechts matig op elkaar lijken, terwijl een ander de 7 gebruikt voor objecten met grote onderlinge gelijkenis (de 7 van de een heeft niet dezelfde betekenis als de 7 van de ander). In zulke gevallen heeft het geen zin de getallen van de ene persoon zonder meer met die van een ander te vergelijken.
Een andere vorm van conditionaliteit doet zich voor als men van iedere matrix alleen die getallen met elkaar mag vergelijken die binnen een en dezelfde rij staan. Deze data noemen we dan *rijconditioneel*. Een traditioneel voorbeeld van rijconditionele data is een matrix met zogenaamde voorkeursrangordeningen. Dit is een matrix met als rijen een aantal personen en als kolommen een aantal voorwerpen, met in elke rij de rangordecijfers die een persoon aan de voorwerpen toekent. Ook hier hoeft de 8 van de ene persoon niet hetzelfde te betekenen als de 8 van een ander. Zelfs al zeggen twee personen allebei dat object X hun grootste voorkeur heeft, dan nog kan de mate van voorkeur van de ene proefpersoon voor dit object veel groter zijn dan die van de andere proefpersoon. In Coombsiaanse termen: de afstand tussen X en persoon P kan veel kleiner zijn dan de afstand tussen X en persoon Q (het omgekeerde kan ook). Dus zelfs al is $o_{PX} = o_{QX}$ dan is het nog steeds mogelijk dat $d_{PX} < d_{QX}$ of dat $d_{PX} > d_{QX}$.
De conditionaliteit van de data heeft consequenties voor de transformatiefunctie f uit $d_{ij} \approx f(o_{ij})$. In een onconditioneel tweeweg MDS-probleem (onconditioneel is in dit geval hetzelfde als matrixconditioneel omdat er maar één matrix is) is er maar één transformatie $f(o_{ij})$ voor alle getallen uit de matrix. In een rijconditioneel probleem is er voor de getallen uit iedere rij i ($i = 1,...,m$; m is het aantal objecten) een afzonderlijke transformatiefunctie $f_i(o_{ij})$. In een onconditioneel drieweg MDS-probleem is er slechts één functie f voor alle getallen, en in een matrixconditioneel driewegprobleem zijn er n verschillende transformatiefuncties $f_k(o_{ijk})$ ($k = 1,...,n$; n is het aantal elementen van de derde weg). In een rijconditioneel driewegprobleem zijn er $m \times n$ transformatiefuncties $f_{ik}(o_{ijk})$ ($i = 1,...,m$; $k = 1,...,n$; m is het aantal elementen van de eerste weg; n is het aantal elementen van de derde weg).

Implicaties

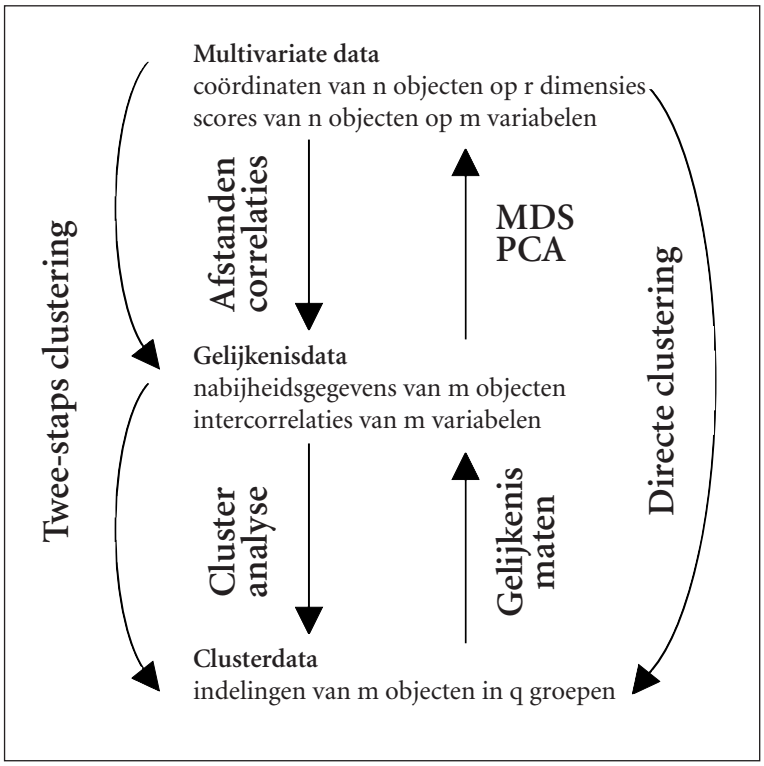
Het hierboven beschreven systeem van Carroll en Arabie (1980) en Young (1984) maakt het mogelijk alle voorkomende verzamelingen van nabijheidsdata in te delen naar de vorm van de datamatrix en naar hun meetkenmerken.

Bij elkaar zijn er zeven aspecten (aantal wegen, aantal modi, aantal replicaties, (a)symmetrie, meetniveau, meetproces en conditionaliteit) die in verschillende combinaties met elkaar voor kunnen komen. In principe levert iedere combinatie een uniek type schaal*probleem* op dat men met een bijbehorende schaal*techniek* zou willen analyseren.
Gelukkig kunnen niet alle combinaties van alle aspecten voorkomen. In de praktijk onderscheiden we negen typen schaalproblemen, die bepaald worden door de vormaspecten van de datamatrix. Deze typen, die weer worden onderverdeeld aan de hand van de meetkenmerken, zullen meer gedetailleerd besproken worden in de paragraaf *Een indeling van schaalproblemen*. Daarbij zullen we aandacht besteden aan de verschillende transformatiefuncties die de relatie tussen observaties en afstanden kunnen beschrijven. Eerst bespreken we echter nog een derde systeem waarmee men observaties kan classificeren: de typologie van Kruskal.

3.4 KRUSKALS TYPOLOGIE VAN DATA

Hierboven zijn twee classificatiesystemen van observaties en data besproken: Coombs datatheorie uit 1964 en het door Young (1984) aangevulde systeem van Carroll en Arabie (1980). Daarnaast is er een derde typologie van data die heel handig is: de door Kruskal (1977b) voorgestelde indeling van data in multivariate data, gelijkenisdata en clusterdata.
Multivariate data zijn data die bestaan uit de coördinaten van een aantal objecten op een aantal dimensies. Zulke coördinaten kunnen het resultaat zijn van een of andere multivariate analyse of MDS-techniek. Typische voorbeelden zijn de ladingen van variabelen op een aantal factoren verkregen via factoranalyse, en de bijbehorende scores van de proefpersonen op die factoren. Dergelijke 'coördinaten' kunnen echter ook 'direct geobserveerd' zijn en bijvoorbeeld bestaan uit de scores die proefpersonen aan objecten toekennen op een aantal beoordelingsschalen (rating scales). In de terminologie van Carroll en Arabie zijn dit dus tweeweg/tweemodale data.
Gelijkenisdata zijn tweeweg/éénmodale data die de onderlinge gelijkenissen van een verzameling objecten weergeven. Zulke data kunnen 'direct geobserveerd' worden, bijvoorbeeld door proefpersonen de objecten paarsgewijs te laten beoordelen, maar kunnen ook het resultaat zijn van bepaalde analysemethoden. Bijvoorbeeld: bij multivariate data kan men correlaties berekenen tussen de variabelen waarop de objecten beoordeeld zijn. Zulke correlatiecoëfficiënten zijn een vorm van gelijkenisdata. Ook kan men met behulp van een afstandsfunctie (zie Hoofdstuk 4) afstanden definiëren tussen de variabelen of tussen de objecten. Daarmee ontstaan nabijheidsdata.
Clusterdata zijn data die aangeven hoe een verzameling objecten in groepen (clusters) is onderverdeeld. Stel dat de objecten A, B, C, D, E, F, en G in drie groepen uiteenvallen: {A, B, C}, {D, E}, en {F, G}. Van zulke gegevens zijn – primitieve – gelijkenisdata te maken door een matrix van objecten × objecten te

construeren waarvan de cellen enen of nullen bevatten: in cel (*i,j*) staat een 1 als object *i* en object *j* samen in één subgroep zitten en een 0 als dat niet het geval is. In principe is zo'n gelijkenismatrix (of een afgeleide daarvan; zie Hoofdstuk 13) met MDS te analyseren; dat levert als resultaat een matrix met multivariate data op. Ook kunnen clusterdata direct als multivariate data genoteerd worden, namelijk door een matrix te construeren met evenveel rijen als er objecten zijn en evenveel kolommen als er clusters zijn. Ook deze matrix bestaat uit enen en nullen; een nul geeft aan dat het desbetreffende object niet in het betreffende cluster zit, een één geeft aan dat dit wel het geval is. Zulk soort matrices kunnen, met name wanneer er meerdere clusterindelingen van dezelfde objecten voorhanden zijn, zinvol geanalyseerd worden met allerlei (niet-lineaire) multivariate technieken (zie Hoofdstuk 10, 11 en 13). Clusterdata kunnen ofwel direct geobserveerd worden, bijvoorbeeld door proefpersonen een verzameling objecten te laten sorteren, of kunnen zelf het resultaat zijn van een eerdere analyse. Clusterdata worden namelijk als resultaat verkregen door clusteranalyse toe te passen op een matrix met multivariate data of op een matrix met gelijkenisdata.



Figuur 3.2 Kruskals classificatie van data

Het bijzondere van Kruskals indeling is dat hij laat zien hoe verschillende typen data, al of niet via een of andere analysetechniek, in elkaar omgezet kunnen worden. Ook wordt hier benadrukt dat de uitkomsten van analysetechnieken in formeel opzicht vergelijkbaar zijn met data die direct geobserveerd worden. De relaties tussen de verschillende datatypen zijn schematisch weergegeven in Figuur 3.2.
Tot slot moet opgemerkt worden dat Kruskal alleen over tweewegdata spreekt. Het is natuurlijk ook mogelijk driewegdata in Kruskals systeem onder te brengen. Het aantal manieren waarop het ene type data in het andere getransformeerd kan worden neemt dan drastisch toe.

3.5 EEN INDELING VAN SCHAALPROBLEMEN

Hoewel de indelingen van Coombs en Kruskal beide elegante en nuttige classificaties van data zijn, is het in hun systemen niet zo dat bij één bepaalde klasse data vanzelfsprekend één bepaald type data-analyse hoort. Een classificatie waarbij de aansluiting tussen observaties en analyse veel duidelijker is, is de door Young aangevulde indeling van Carroll en Arabie die eerder in dit hoofdstuk besproken is. De eerste vier aspecten van deze indeling – wegen, modi, replicaties en (a)symmetrie – leiden tot een classificatie van negen typen schaal*problemen* die elk weer onderverdeeld kunnen worden met betrekking tot hun meetkenmerken (meetniveau, meetproces en conditionaliteit).
Zoals al eerder is opgemerkt, zouden we voor ieder type schaalprobleem een geschikte schaal*techniek* willen hebben. In zo'n techniek moeten steeds twee onderscheidbare problemen worden opgelost: hoe moet men de observaties transformeren om schattingen van echte afstanden te krijgen en hoe moet men deze schattingen van afstanden bewerken om een goede afbeelding te krijgen? In het eerste probleem gaat het erom een *optimale, toegestane transformatie* te vinden waarmee het veronderstelde verband tussen observaties en afstanden wordt weergegeven. Daarbij moeten we ons realiseren dat er in de praktijk waarschijnlijk geen perfecte relatie (*fit*) tussen data en afstanden bestaat. Dit gebrek aan *fit* is in het algemeen groter naarmate men *sterkere aannamen* over het meetniveau van de data maakt, dat wil zeggen, naarmate men veronderstelt dat de observaties op een hoger niveau (interval of ratio) gemeten zijn. Hetzelfde geldt voor aannamen met betrekking tot meetproces en conditionaliteit. Discreet en onconditioneel zullen in het algemeen beide een lagere *fit* opleveren dan continue en matrix- of rijconditioneel. In de praktijk moeten we dus zoeken naar die transformatie $f(o_{ij})$ die de relatie tussen observaties en afstanden zo goed mogelijk beschrijft en tegelijkertijd voldoet aan de eisen die door meetniveau en meetproces gesteld zijn: vandaar de termen 'optimaal' en 'toegestaan'.
Hieronder bespreken we de genoemde negen typen schaalproblemen en gaan we in op de verschillende effecten van de conditionaliteit van de observaties.

De andere meetkenmerken (meetniveau en meetproces) komen later in de paragraaf *Transformaties* aan de orde.

1 *CMDS*. Het klassieke meerdimensionale schaalprobleem (*classical multidimensional scaling*): de analyse van één enkele symmetrische tweeweg/éénmodale datamatrix. Zo'n matrix kan eventueel verkregen zijn door meerdere tweeweg/éénmodale matrices te middelen, of door een oorspronkelijk asymmetrische matrix symmetrisch te maken. De essentie van CMDS is echter dat het uiteindelijk om de analyse van één enkele symmetrische matrix met nabijheidsdata o_{ij} gaat. De gezochte MDS-oplossing beeldt de objecten in een ruimte af als punten met onderlinge afstanden d_{ij}. De optimale transformatie die gevonden moet worden is

$$d_{ij} \approx f(o_{ij}) \qquad [3.1]$$

waarin $\approx$ betekent dat niet alle d_{ij}-waarden exact gelijk aan $f(o_{ij})$ zullen zijn, maar dat de gevonden transformatie f de verschillen tussen d_{ij} en $f(o_{ij})$ op een bepaalde manier minimaliseert[3].
Indien de datamatrix *rijconditioneel* is, dan kunnen de getallen uit verschillende rijen van de matrix niet zonder meer met elkaar vergeleken worden. Dit betekent dat we de getallen *per rij* mogen transformeren en dus voor elke rij i een afzonderlijke transformatiefunctie f_i moeten zoeken. De relatie waar het om gaat is dan

$$d_{ij} \approx f_i(o_{ij}). \qquad [3.2]$$

Een rijconditionele aanpak zou men kunnen kiezen als de matrix van gelijkenissen niet symmetrisch is, dus wanneer $o_{ij} \neq o_{ji}$. Als we nu toch aannemen dat $d_{ij} = d_{ji}$ dan komt dezelfde afstand tussen i en j in rij i met een andere geobserveerde nabijheidsscore overeen dan in rij j. Dat kan alleen maar als de transformatiefunctie $f(o_{ij})$ in rij i er anders uitziet dan $f(o_{ij})$ in rij j, zodat we de data rijconditioneel moeten transformeren.

2 *ASYMSCAL*. Het asymmetrische meerdimensionale schaalprobleem (*asymmetric multidimensional scaling*): de analyse van één enkele asymmetrische matrix van nabijheidsdata. In tegenstelling tot de rijconditionele aanpak van asymmetrische data berust ASYMSCAL op een afstandsmodel (zie Hoofdstuk 4) dat het mogelijk maakt dat niet alleen $o_{ij} \neq o_{ji}$ maar dat ook $d_{ij} \neq d_{ji}$ kan zijn. Dit gebeurt door middel van speciale parameters die het verschil tussen d_{ij} en d_{ji} (en dus ook tussen o_{ij} en o_{ji} kunnen verklaren (zie Hoofdstuk 8). De asymmetrie van de afstanden $d_{ij} \neq d_{ji}$ kunnen we ook formuleren als $d_{ij/i} \neq d_{ij/j}$ (de werkelijke afstand tussen i en j gezien vanuit i is niet gelijk aan de werkelijke afstand tussen i en j gezien vanuit j). De transformatiefunctie wordt dan

3 Bijvoorbeeld door ervoor te zorgen dat $\mathfrak{L} = \sum_i \sum_j \{d_{ij} - f(o_{ij})\}^2$ minimaal is: het kleinste-kwadratencriterium.

$$d_{ij/i} \approx f_i(o_{ij/i}). \qquad [3.3]$$

In principe is het mogelijk ook in het ASYMSCAL-model een rijconditionele analyse met $d_{ij/i} \approx f_i(o_{ij/i})$ uit te voeren. Het is echter niet gemakkelijk een voorbeeld te bedenken waarin dit de meest voor de hand liggende analyse zou zijn.

3 *RMDS*. Het klassieke MDS-probleem met replicaties (*replicated multidimensional scaling*): hier gaat het om de analyse van een drieweg/tweemodale nabijheidsmatrix waarvan de derde weg uit 'objecten' bestaat die als replicaties van elkaar worden opgevat. Het doel van RMDS is dat van de objecten uit de eerste modus een afbeelding verkregen wordt die zo goed mogelijk past bij de nabijheidsmatrices van *alle* replicaties tegelijkertijd.
In het RMDS-geval is er sprake van een drieweg/tweemodale datamatrix met observaties o_{ijk} waarin i en j de elementen van de eerste modus en k de elementen van de derde weg (de tweede modus) aanduiden. De relatie tussen afstanden en data kan nu drie vormen aannemen:

$$d_{ij} \approx f(o_{ijk}) \qquad [3.4]$$

$$d_{ij} \approx f_k(o_{ijk}) \qquad [3.5]$$

$$d_{ij} \approx f_{ik}(o_{ijk}). \qquad [3.6]$$

In het eerste geval is er één transformatiefunctie voor de hele drieweg/tweemodale datamatrix. Alle getallen van alle replicaties worden op dezelfde manier getransformeerd. Deze aanpak houdt in dat de data onconditioneel zijn: elk getal uit de driewegmatrix mag met elk ander getal vergeleken worden. In het tweede geval krijgt iedere replicatie k een eigen transformatiefunctie f_k; de data uit de verschillende matrices worden dus matrixconditioneel opgevat. Dit is de juiste aanpak als het geen zin heeft getallen uit verschillende matrices met elkaar te vergelijken. In het derde geval wordt er voor iedere rij i van elke matrix k een afzonderlijke transformatiefunctie gezocht. De data worden dus rijconditioneel opgevat.

4 *RAMDS*. Asymmetrische MDS met replicaties (*replicated asymmetric multidimensional scaling*): de combinatie van ASYMSCAL en RMDS op een drieweg/tweemodale nabijheidsmatrix. RAMDS zoekt een bij alle replicaties passende oplossing die de objecten van de eerste modus afbeeldt, en berekent daarbij aparte parameters die de asymmetrie van de afzonderlijke nabijheidsmatrices verklaren. Hier zijn dezelfde transformatiefuncties als bij RMDS mogelijk, met dien verstande dat er een afstandsmodel gebruikt wordt waarin het mogelijk is dat $d_{ij/i} \neq d_{ij/j}$. Dus $d_{ij/i} \approx f(o_{ijk/i})$ in het onconditionele geval, $d_{ij/i} \approx f_k(o_{ijk/i})$ in het matrixconditionele geval, en $d_{ij/i} \approx f_{ik}(o_{ijk/i})$ in het rijconditionele geval.

5 *WMDS*. Gewogen MDS (*weighted multidimensional scaling*): de analyse van een drieweg/tweemodale nabijheidsmatrix waarbij parameters (zogenaamde *gewichten*) worden berekend die aangeven op welke manier de objecten van de derde weg (de tweede modus) van elkaar verschillen. Omdat de objecten van

de derde weg vaak personen zijn, wordt dit type analyse ook wel *individual differences scaling* genoemd. Anders dan in RMDS worden in WMDS de elementen van de derde weg (tweede modus) niet als simpele replicaties van elkaar opgevat. WMDS gaat uit van een afstandsmodel dat het toelaat dat er bij verschillende matrices uit de derde weg (bijvoorbeeld van verschillende proefpersonen) verschillende configuraties van punten horen. De afstand tussen twee punten in de ene configuratie hoeft dan niet gelijk te zijn aan de afstand tussen die punten in een andere configuratie. Omdat er in dit afstandsmodel (zie Hoofdstuk 4) zogenaamde *gewichten* een rol spelen, heet dit type analyse *gewogen MDS*. Niet alleen de observaties verschillen hier per matrix, maar dus ook de bijbehorende afstanden. Vandaar de notatie d_{ijk} voor de afstand tussen i en j zoals 'gezien door element k'. Ook hier kan men een onderscheid maken tussen het onconditionele, het matrixconditionele en het rijconditionele geval. De bijbehorende functies zien er als volgt uit:

$$d_{ijk} \approx f(o_{ijk}) \qquad [3.7]$$

$$d_{ijk} \approx f_k(o_{ijk}) \qquad [3.8]$$

$$d_{ijk} \approx f_{ik}(o_{ijk}) \,. \qquad [3.9]$$

Meestal wordt bij WMDS aangenomen dat de gegevens matrixconditioneel zijn. Een rijconditionele analyse zou men onder andere kunnen toepassen als de nabijheidsgegevens van de afzonderlijke tweeweg/éénmodale observaties asymmetrisch zijn.

6 *WAMDS*. Gewogen, asymmetrische MDS (*weighted asymmetric multidimensional scaling* ook wel *asymmetric individual differences scaling* of *ASINDSCAL*): een vorm van WMDS die rekening houdt met asymmetrie van de individuele nabijheidsmatrices. Het toegepaste afstandsmodel laat toe dat er per matrix verschillende configuraties met verschillende afstanden berekend worden die bovendien asymmetrisch kunnen zijn, dus: $d_{ijk/i} \neq d_{ijk/j}$. De transformaties kunnen voor het onconditionele, het matrixconditionele en het rijconditionele geval als volgt genoteerd worden:

$$d_{ijk/i} \approx f(o_{ijk/i}) \qquad [3.10]$$

$$d_{ijk/i} \approx f_k(o_{ijk/i}) \qquad [3.11]$$

$$d_{ijk/i} \approx f_{ik}(o_{ijk/i}). \qquad [3.12]$$

Meestal neemt men aan dat dit soort data matrixconditioneel is.

7 *CMDU*. Klassieke meerdimensionale ontvouwing (*classical multidimensional unfolding)*: de analyse van een tweeweg/tweemodale datamatrix die de nabijheid aangeeft van de objecten uit de ene modus tot de objecten uit de andere modus. De door Coombs (1964) beschreven methode om voor dit soort gege-

vens een ruimtelijke afbeelding te krijgen werd door hem *unfolding* (ontvouwing) genoemd. In beginsel zijn bij CMDU dezelfde transformaties mogelijk als bij CMDS. In het typische CMDU-geval wordt meestal aangenomen dat de data rijconditioneel zijn. In dat geval geldt dus: $d_{ij} \approx f_i(o_{ij})$.

8 *RMDU*. Meerdimensionale ontvouwing met replicaties (*replicated multidimensional unfolding*): de analyse van een drieweg/driemodale nabijheidsmatrix waarin de elementen van de derde weg als replicaties van elkaar worden opgevat en niet afzonderlijk worden afgebeeld. Bijvoorbeeld: politieke partijen en kenmerken worden zodanig weergegeven dat hun afstanden zo goed mogelijk overeenkomen met alle beoordelingen van een aantal verschillende personen (de derde weg en derde modus) tegelijkertijd. Bij RMDU kan men in beginsel dezelfde transformaties toepassen als bij RMDS, zij het dat er meestal van wordt uitgegaan dat de data rijconditioneel zijn.

9 *WMDU*. Gewogen meerdimensionale ontvouwing (*weighted multidimensional unfolding*): de analyse van een drieweg/driemodale datamatrix, bijvoorbeeld een matrix met voorkeursgegevens van een aantal proefpersonen voor een aantal objecten op verschillende tijdstippen. Een WMDU-analyse levert een ruimtelijke afbeelding op van de elementen van de eerste twee modi, samen met een verzameling parameters (gewichten) die aangeven hoe de elementen van de derde modus van elkaar verschillen. Bij WMDU zijn dezelfde transformaties mogelijk als bij WMDS; in het typische WMDU-geval wordt aangenomen dat de datamatrix rijconditioneel is.

Deze negen klassen van analyseproblemen zijn elk weer onder te verdelen aan de hand van de meetkenmerken van de nabijheidsdata, met name het meetniveau, de conditionaliteit en het meetproces waarmee gemeten is. Met deze aspecten moet rekening worden gehouden in het algoritme (de verzameling rekenregels) dat gebruikt wordt om het betreffende schaalprobleem op te lossen. Dit gebeurt met name bij het zoeken naar een optimale transformatie van de data om een zo goed mogelijke overeenstemming te krijgen tussen afstanden in de oplossing en de getransformeerde nabijheidsgegevens. Deze kwestie wordt in de volgende paragraaf behandeld.

3.6 TRANSFORMATIES

Bij MDS van nabijheidsdata kan men in principe steeds twee problemen onderscheiden. In de eerste plaats: hoe moet men de observaties transformeren om schattingen van echte afstanden te krijgen? Ten tweede: hoe moet men deze schattingen bewerken om een goede afbeelding te krijgen, dat wil zeggen een afbeelding waarin de afstanden tussen de punten overeenkomen met de geobserveerde schattingen van die afstanden? Een oplossing voor het tweede probleem, de Young-Householdermethode, is al besproken in het vorige hoofdstuk. Deze methode is geschikt als de data op rationiveau gemeten zijn en dus verondersteld worden evenredig te zijn met echte afstanden. De observaties

zelf zijn dan al schattingen van afstanden: het is dan niet nodig een transformatie te vinden die voorafgaat aan de analyse. Dat ligt anders bij data die op intervalniveau gemeten zijn. Om zulke observaties in (schattingen van) afstanden om te zetten is een lineaire transformatie $f(o_{ij}) = k.o_{ij} + c$ nodig, waarbij het in het bijzonder gaat om het bepalen van c, de additieve constante. Hierbij kunnen we twee strategieën volgen, een tweestapsprocedure en een directe methode. In eerstgenoemde aanpak maken we een of andere schatting van c, transformeren vervolgens de observaties tot (schattingen van) afstanden en passen de Young-Householdermethode toe. In de directe methode maken we gebruik van een MDS-techniek waarin tegelijkertijd een optimale transformatiefunctie én een optimale afbeelding van de objecten verkregen wordt.
Als de nabijheden op ordinaal niveau gemeten zijn, moet er gezocht worden naar een *monotone transformatie* van de data. Daarbij zijn er weer twee verschillende gevallen te onderscheiden: (a) tussen d_{ij} en o_{ij} bestaat een kromlijnig, regelmatig verband dat door een (eenvoudige) wiskundige functie te beschrijven is (bijvoorbeeld $d_{ij} \approx e^{oij}$ of $d_{ij} \approx \log(o_{ij}) + c$); (b) er is sprake van een onbekende functie f die de vorm heeft van een trap met treden van ongelijke hoogte en ongelijke diepte. In het eerste geval is het (soms) mogelijk schattingen van de $\{d_{ij}\}$ te maken om daarna Young-Householder toe te passen. In het tweede geval moet men gebruikmaken van een van de moderne, *niet-metrische schaaltechnieken*, die onder andere in Hoofdstuk 6 worden behandeld.
De hierboven geschetste tweestapsaanpak is typisch voor wat men in het algemeen *metrische MDS* noemt. Tot het begin van de jaren zestig was metrische analyse de enige mogelijkheid om MDS-afbeeldingen te maken. In 1962 en 1964 zorgden Shepard (1962a,b), Kruskal (1964a,b) en Coombs (1964) voor wat men de *niet-metrische doorbraak* in de MDS is gaan noemen. Deze doorbraak was van groot belang omdat de eerdere, metrische analysemethoden eigenlijk ongeschikt zijn voor afstandsdata die op ordinaal niveau gemeten zijn en slechts qua rangorde overeenkomen met 'echte' afstanden.

Monotone transformaties

Zoals we hierboven zagen zijn er twee factoren die de vorm van de optimale transformaties $f(o_{ij})$ bepalen. Dat zijn enerzijds de aannamen over het meetniveau en anderzijds die over het meetproces van de observaties. Qua meetniveau maakten we onderscheid tussen nominale, ordinale, interval-, en ratiodata (zie bijvoorbeeld Meerling, 1989). Wat betreft het meetproces maakten we onderscheid tussen discreet (observaties met gelijke waarden, zogenaamde *ties*, moeten na transformatie gelijke waarden houden) en continu (*ties* in de observaties mogen na transformatie verschillende waarden krijgen).
Van observaties op interval- en rationiveau nemen we aan dat ze via een of andere, exact te omschrijven functie corresponderen met de 'echte' afstanden tussen de objecten. Niet alleen functies van het type $f(o_{ij}) = a + b.o_{ij}$ (lineaire functies) of functies van het type $f(o_{ij}) = c.o_{ij}$ (multiplicatieve functies) zijn daar een voorbeeld van, maar ook $f(o_{ij}) = \log(o_{ij})$ en $f(o_{ij}) = e^{Oij}$ (respectievelijk

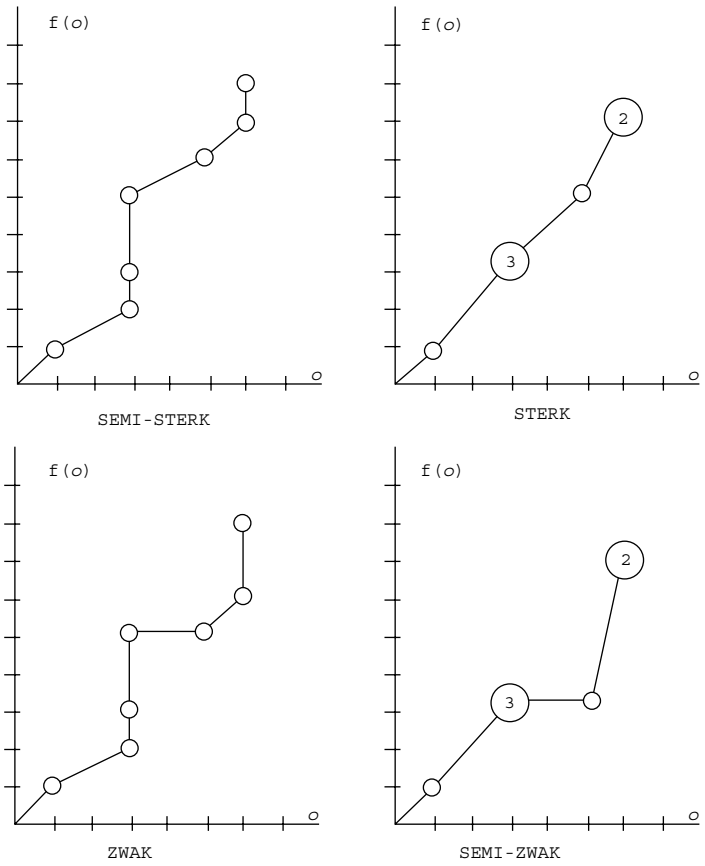
logaritmische en exponentiële functies). De mogelijke transformaties van data op interval- en rationiveau zullen we in dit hoofdstuk verder buiten beschouwing laten. Ook op de mogelijke transformaties van nominaal gemeten nabijheidsdata gaan we in dit hoofdstuk niet verder in.
Bij ordinale observaties zijn de nabijheidsrelaties tussen de objecten uitgedrukt in getallen waarvan alleen de volgorde zinvolle informatie bevat. Zulke getallen mag men op iedere willekeurige manier monotoon transformeren, omdat daarmee de volgorde-informatie behouden blijft. Wel moet men erop letten of de transformatie monotoon stijgend moet zijn (er wordt een positief verband tussen observaties en afstanden verondersteld) of monotoon dalend (er wordt een negatief verband tussen data en afstanden verondersteld). Binnen de monotone transformaties kan men twee vormen onderscheiden: *sterke* en *zwakke*. Voor een sterke monotone transformatie geldt dat als $o_{ij} < o_{hk}$ dan ook altijd $f(o_{ij}) < f(o_{hk})$ moet zijn; dit heet de eis van sterke monotoniciteit. De eis van zwakke monotoniciteit houdt slechts in dat als $o_{ij} < o_{hk}$ er voor de transformaties alleen maar hoeft te gelden dat $f(o_{ij}) \leq f(o_{hk})$.
Ook het veronderstelde meetproces stelt eisen aan de toegestane transformaties. Als $o_{ij} = o_{hk}$ dan moet in het discrete geval ook altijd gelden dat $f(o_{ij}) = f(o_{hk})$. In het continue geval is dan niet beslist vereist dat $f(o_{ij}) = f(o_{hk})$ maar is het ook toegestaan dat $f(o_{ij}) > f(o_{hk})$ of $f(o_{ij}) < f(o_{hk})$ is. Combineren we de verschillende mogelijkheden met elkaar dan kunnen we vier soorten monotone transformaties onderscheiden:

1 *zwak monotone transformaties*: als $o_{ij} < o_{hk}$ dan mag $f(o_{ij}) \leq f(o_{hk})$ en als $o_{ij} = o_{hk}$ dan geldt niet noodzakelijk dat $f(o_{ij}) = f(o_{hk})$
2 *semi-zwak monotone transformaties*: als $o_{ij} < o_{hk}$ dan mag $f(o_{ij}) \leq f(o_{hk})$ maar als $o_{ij} = o_{hk}$ dan moet gelden $f(o_{ij}) = f(o_{hk})$
3 *semi-sterk monotone transformaties*: als $o_{ij} < o_{hk}$ dan moet gelden $f(o_{ij}) < f(o_{hk})$ maar als $o_{ij} = o_{hk}$ dan geldt niet noodzakelijk dat $f(o_{ij}) = f(o_{hk})$
4 *sterk monotone transformaties*: als $o_{ij} < o_{hk}$ dan moet gelden $f(o_{ij}) < f(o_{hk})$ en als $o_{ij} = o_{hk}$ dan moet gelden $f(o_{ij}) = f(o_{hk})$.

De eerste en de derde transformatie zijn toegestaan onder aanname van een continu meetproces. Het zoeken naar dit soort transformaties heet *Kruskal's primary approach to ties* (ties mogen opengebroken worden). De tweede en de vierde transformatie zijn vereist onder aanname van een discreet meetproces. Het vinden van dit soort transformaties heet *Kruskal's secondary approach to ties* (ties mogen niet opengebroken worden).

Tabel 3.3 Een getallenvoorbeeld van vier monotone transformaties

transformatie				
observaties	zwak	semi-zwak	semi-sterk	sterk
1	1	1	1	1
3	2	3.33	2	3.33
3	3	3.33	3	3.33
3	5	3.33	5	3.33
5	5	3.33	6	5
6	6	7	7	7
6	8	7	8	7



Figuur 3.3 Grafieken van zwakke, semi-zwakke, semi-sterke en sterke monotoon stijgende transformaties

Een getallenvoorbeeld staat in Tabel 3.3; daarin worden zeven observaties op vier verschillende manieren monotoon getransformeerd. De grafieken van deze vier transformaties zijn weergegeven in Figuur 3.3. Aan dit getallenvoorbeeld is te zien dat in de zwakke en semi-sterke transformaties gelijke observaties verschillende transformatiewaarden mogen krijgen. In de semi-zwakke en sterke transformaties is dat niet zo. In de zwakke en semi-zwakke transformaties mogen verschillende observaties gelijke transformatiewaarden krijgen (zie $o = 3$ en $o = 5$), terwijl dat in de semi-sterke en sterke transformaties niet is toegestaan. Let wel: deze transformaties zijn uiteraard niet uniek; er zijn oneindig veel getallen te bedenken die de observaties op de betreffende vier manieren transformeren! Een voorbeeld van de manier waarop zulke transformaties gevonden kunnen worden, wordt besproken in Blok 3.1.

BLOK 3.1 HET VINDEN VAN MONOTONE TRANSFORMATIES

Stel dat we iemand vragen om uit het hoofd aan te geven welke van de afstanden tussen Amsterdam (A), Rotterdam (R), Den Haag (H), Utrecht (U) en Eindhoven (E) het kleinst is, welke afstand het op één na kleinst is, welke afstand daarna komt, enzovoort, en ten slotte welke afstand het grootst is. Stel nu dat onze proefpersoon de volgende rangordening geeft: $1 = o_{AU}$, $2 = o_{HU}$ en o_{RU}, $3 = o_{AH}$ en o_{HR}, $4 = o_{EU}$, $5 = o_{AR}$ en o_{RE}, $6 = o_{EH}$ en $7 = o_{AE}$. Er zitten dus drie *ties* in de geobserveerde nabijheidsrangordening van deze proefpersoon. Zoeken we de bijbehorende afstanden op in Tabel 2.1, dan zien we dat de werkelijke afstanden tussen deze vijf steden als volgt zijn: $d_{AU} = 37$, $d_{AH} = 57$, $d_{HR} = 21$, $d_{HU} = 61$, $d_{RU} = 57$, $d_{EU} = 87$, $d_{AR} = 73$, $d_{RE} = 111$, $d_{EH} = 132$ en $d_{AE} = 120$. We zien nu dat de door de proefpersoon gegeven rangordening niet volledig correspondeert met de rangorde van de echte afstanden. De vraag is nu, hoe moeten we de geobserveerde waarden (dus de getallen 1 tot en met 7) transformeren om een zo groot mogelijke overeenstemming tussen de observaties-na-transformatie en de echte afstanden te krijgen?

Hiervoor wordt veelal de door Kruskal voorgestelde monotone regressiemethode gebruikt. Van deze methode bestaan twee varianten: (a) de *primary approach*, ook wel de continue benadering genoemd, waarbij *ties* in de data mogen worden opengebroken, en (b) de *secondary approach* (de discrete benadering) waarbij *ties* in de observaties bewaard moeten blijven. In de continue variant van deze methode vindt men $f(o_{ij})$ door in eerste instantie de geobserveerde o-s te rangordenen van klein naar groot. Aan de kleinste waarde, zeg o_{ih}, kent men dan de waarde d_{ih}, de afstand tussen de objecten i en h, toe. Dat wordt dan de waarde $f(o_{ih})$. Vervolgens geeft men de op één na kleinste waarde van O, stel o_{kl}, de waarde van de

bijbehorende afstand d_{kl}. Is $d_{kl} \geq d_{ih}$ dan is alles in orde en gaan we naar de op twee na kleinste waarde van O. Is daarentegen $d_{kl} < d_{ih}$ dan staan de afstanden d_{kl} en d_{ih} niet in dezelfde volgorde als o_{kl} en o_{ih}. Om nu toch een niet-dalende transformatie te krijgen kennen we zowel aan o_{kl} als aan o_{ih} de gemiddelde waarde van d_{kl} en d_{ih} toe. Daarna gaan we naar de op twee na kleinste waarde van O, stel o_{ht}, en kijken we of de bijbehorende afstand d_{ht} inderdaad groter is dan de waarde die aan o_{kl} is toegewezen. Als dat zo is, dan wordt $f(o_{ht}) = d_{ht}$; is $f(o_{kl})$ echter groter dan d_{ht}, dan wordt $f(o_{ht}) = (d_{ht} + d_{kl})/2$ tenzij deze waarde weer kleiner is dan $f(o_{ih})$. In dat laatste geval kiezen we de waarde $f(o_{ht}) = (d_{ht} + d_{kl} + d_{ih})/3$. Zo gaat men net zolang verder tot aan alle waarden van O een transformatiewaarde is toegekend en al deze waarden in dezelfde volgorde staan als de oorspronkelijke observaties.

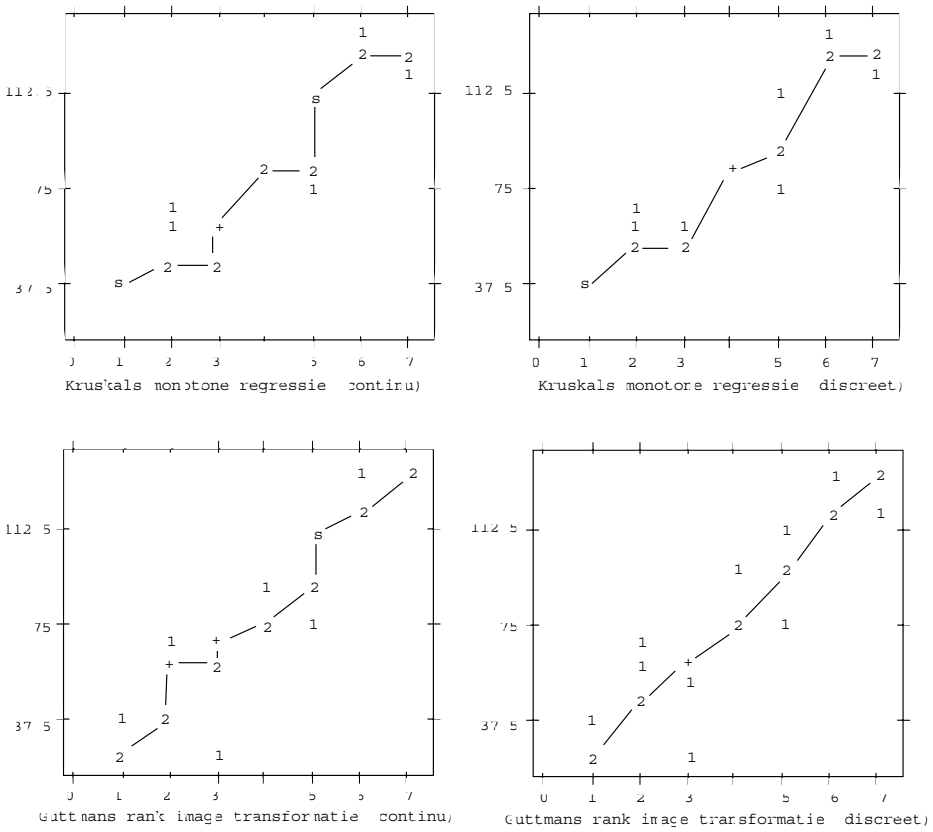
In de discrete variant van Kruskals methode kent men aan de observaties uit een bepaalde groep *ties* de waarde toe van het gemiddelde van de afstanden in de desbetreffende groep. Daarna gaat men verder te werk als in de continue benadering, met dien verstande dat de observaties binnen een bepaalde *tie*-groep na transformatie allemaal dezelfde waarden krijgen. Deze procedures worden toegelicht in Tabel 3.4.

Naast Kruskals monotone-regressiemethoden om transformaties te vinden, wordt in sommige computerprogramma's de *rank image*-transformatie van Guttman toegepast. Deze komt er simpelweg op neer dat zowel de observaties als de afstanden op volgorde van klein naar groot worden gezet. De observatie die het k-de rangnummer heeft, krijgt als transformatie de afstand met het k-de rangnummer toegewezen, ook al hoort die k-de afstand bij een heel ander puntenpaar dan de k-de observatie. De rank image-transformatie heeft net als Kruskals monotone regressie twee varianten: een discrete en een continue. Ook deze twee varianten worden in Tabel 3.4 toegelicht. In Figuur 3.4 zijn de vier transformatiecurven $f(o_{ij})$ ingetekend in vier grafieken waarin D is uitgezet tegen O. Deze grafieken worden Shepard-diagrammen genoemd; in Hoofdstuk 6 komen we hierop terug.

Tabel 3.4 Vier methoden voor het vinden van een monotone transformatie

paren geordend naar de waarden van de observaties										
	AU	RU	HU	HR	AH	EU	AR	RE	EH	AE
o_j	1	2	2	3	3	4	5	5	6	7
d_j	37	57	61	21	57	87	73	111	132	120
Kruskals monotone regressie (continu proces)										
Stap 1	37	57	61	> 21[a]						
Stap 2	37	57	> 41	41						
Stap 3	37	46.33	46.33	46.33	57	87 > 73				
Stap 4	37	46.33	46.33	46.33	57	80	80	111	132	>120
Stap 5	37	46.33	46.33	46.33	57	80	80	111	126	126
Kruskals monotone regressie (discreet proces)										
Stap 1	—	59	= 59	39	= 39	—	92 =	92	-	-
Stap 2	37	59	= 59	> 39	= 39					
Stap 3	37	49	= 49	= 49	= 49	87	92 =	92	132	>120
Stap 4	37	49	= 49	= 49	= 49	87	92 =	92	126	= 126
Guttmans rank image transformatie (continu proces)										
Stap 1	21	37	57	57	61	73	87	111	120	132
Guttmans rank image transformatie (discreet proces)										
Stap 1	21	37	57	57	61	73	87	111	120	132
Stap 2	21	47	= 47	59	= 59	73	99 =	99	120	132

[a] *De vetgedrukte getallen zijn transformatiewaarden die in de verkeerde volgorde staan; in de volgende stap worden zij gemiddeld.*



Figuur 3.4 Vier transformatiecurven in de grafieken van D tegen O. De punten met coördinaten {o,d} zijn aangegeven met het symbool 1, de punten {o,f(o)} met het symbool 2.

3.7 RECAPITULATIE

In Hoofdstuk 2 hebben we gezien hoe men een MDS-techniek kan gebruiken om gegevens over afstanden tussen objecten in een landkaartachtige configuratie af te beelden. In het voorbeeld dat we behandelden, bestonden de gegevens uit echte, geografische afstanden, gemeten in kilometers, die werden gebruikt om een aantal steden in een landkaart te plaatsen. Die kaart kan door de cartograaf net zo groot of net zo klein gemaakt worden als in een bepaald geval gewenst is; in het spoorboekje is een kleine kaart voldoende, maar voor in de auto is een tamelijk grote kaart wel handig. De steden kunnen dus op verschillende schaalgrootte worden afgebeeld, het enige dat van belang is, is dat

de afstanden tussen de steden in de kaart (meestal enkele centimeters) evenredig zijn met de werkelijke afstanden in kilometers. Als dit zo is, dan is er een goede overeenkomst tussen ons model (de kaart, de configuratie van steden) en de werkelijkheid (de geobserveerde afstanden). De transformatiefunctie die men nodig heeft om van de afstanden die geobserveerd zijn afstanden in de kaart te krijgen is dan $d_{ij} = a.o_{ij}$ waarbij a de 'schaal' van de kaart is, bijvoorbeeld 1/50000. Ook als er meetfouten in de geobserveerde afstanden zitten – zoals in ons voorbeeld – dan zal men toch willen dat $d_{ij} \approx a.o_{ij}$. Bij een echte landkaart zoeken we dus een multiplicatieve transformatie van o_{ij}.
In dit hoofdstuk hebben we gezien dat men, afhankelijk van het soort data, ook andere transformatiefuncties kan veronderstellen. Uitgaande van meetniveau, meetproces en conditionaliteit kan men in ieder MDS-probleem één of meer toegestane transformaties trachten te vinden, die sterk, semi-sterk, semizwak of zwak kunnen zijn. Hoe zwakker de eisen die aan een transformatie gesteld worden, des te meer transformaties er mogelijk zijn en des te gemakkelijker het is een goede overeenstemming tussen de (getransformeerde) observaties en de afstanden in de oplossing te krijgen.
In Hoofdstuk 6 zal een algoritme behandeld worden waarmee een oplossing van het klassieke CMDS-probleem gevonden wordt. Daarin zullen we zien hoe een configuratie verkregen wordt en hoe de data getransformeerd moeten worden om een optimale overeenstemming met de afstanden in die configuratie te bereiken. Voor we daar aan toe zijn, moeten we ons echter eerst nog bezighouden met de precieze betekenis van het begrip afstand en met de verschillende manieren waarop men uit de coördinaten van punten in een configuratie afstanden kan berekenen. Dit gebeurt in Hoofdstuk 4.