De toegepaste datatheorie en de verwondering

Meulman, J.J.

De toegepaste datatheorie en de verwondering

Rede uitgesproken door

Jacqueline J. Meulman

bij de aanvaarding van het ambt van bijzonder hoogleraar
in de toegepaste datatheorie, in het bijzonder de multivariate
analyse van kwalitatieve data aan de Universiteit Leiden
op 3 september 1999.

Het feit dat de tekst van deze oratie pas veel later in drukvorm is verschenen dan hij is uitgesproken, is voornamelijk te danken aan mijn ambitie om ondanks mijn beperkte muzikale kennis het lied *Wat heb ik nu aan algebra* zelf in notenschrift te zetten.

Mijnheer de Rector Magnificus,

Zeer gewaardeerde toehoorders,

De titel van mijn oratie *De toegepaste datatheorie en de verwondering* verdient enige uitleg. Toegepaste datatheorie is de naam van mijn leerstoel. Maar vanwaar die verwondering? Ik noem eerst een secundaire reden: de herinnering aan het boek *De verwondering* van Hugo Claus [1]. Was het boek voor mij als middelbare scholier niet altijd even gemakkelijk te doorgronden, de schoonheid van de titel is mij altijd bijgebleven. De concrete aanleiding is van recentere datum. Het betreft een artikel in het gratis studentenblad SUM. Bij het doorbladeren werd mijn aandacht getrokken door de volgende passage, die ik indertijd gelijk heb uitgeknipt. Aan het woord is Dr. R. Meijer, destijds studieadviseur bij psychologie aan de Universiteit van Amsterdam:

> "Methodologie, functieleer en kansberekening kennen aan onderzoek een grotere mate van objectiviteit toe. Ik pleit niet voor psychologie als harde wetenschap maar de relatie tussen psychologie en wiskunde is onontbeerlijk. Hoewel wiskunde tot de ingangseisen behoort, of een verplichte cursus voor mensen die daaraan niet voldoen, leidt dat mensen niet van de veronderstelling af een studie te volgen die zich alléén over menselijke gedragingen buigt. Uiteindelijk is psychologie natuurlijk een studie over de mens. De interesse daarvoor bestaat uit een verwondering en niet uit getallen ....".

Volgens Van Dale betekent verwondering: de toestand van het binnenste van de mens, als zetel van zijn geestelijk gevoel, zijn neigingen, hartstochten en stemmingen, die ontstaat wanneer men iets gewaarwordt dat men *niet* of *anders* had verwacht. Ik heb mij tot doel gesteld u in deze oratie uit te leggen dat de kwantitatieve benadering in de sociale en gedragswetenschappen niet alleen onontbeerlijk is, maar wel degelijk óók verwondering teweeg kan brengen; met andere woorden, ik ga u uitleggen waarom de toegepaste datatheorie mijn hartstocht oproept.

Het meten in de gedragswetenschappen

Sir Francis Galton, die leefde van 1822 tot 1911, wordt doorgaans erkend als een van de uitvinders van de statistische begrippen regressie en correlatie, welke hij ontwikkelde bij zijn studie naar de overerfelijkheid van intelligentie. Hoewel de geschiedenis van de statistiek altijd mijn warme belangstelling heeft gehad, wil ik het vandaag niet over de statistische bijdragen van Galton hebben. Wat ik ooit geweten moet hebben, maar in de loop der tijd blijkbaar weer vergeten, is dat Galton óók als de uitvinder van de psychologische rating scale te boek staat. Zo staat geschreven: "er is weinig twijfel dat de eerste rating scale gebruikt in een psychologisch probleem die

van Galton was, gebruikt bij het meten van de levendigheid van herinneringen". Deze uitspraak van Guilford [2] wordt geciteerd in een artikel in *Psychological Bulletin* uit 1953 van de hand van Elson en Elson, die hun bezoek beschrijven aan een museum in New Harmony, Indiana, USA [3].

De New Harmony kolonie werd gesticht in 1925 door Robert Owen. Deze zou zeer geïnteresseerd zijn in wat we vandaag de dag pedagogiek of ontwikkelingspsychologie zouden noemen. In het museum zagen Elson en Elson een koperen plaat, waarop, volgens Owen, een kind zijn of haar vaardigheden kon aangeven. De plaat heeft als opschrift "schaal van de menselijke vermogens en eigenschappen bij de geboorte", en bestaat uit tien (sub)schalen met de volgende benamingen: zelfgehechtheid, gevoelens, inzicht, verbeelding, geheugen, reflectie, perceptie, prikkelbaarheid, moed en kracht. Er kan een schuif op en neer bewogen worden om op een schaalverdeling de sterkte van de eigenschappen aan te geven. Als de schuiven op de verschillende schalen op een bepaalde positie zijn gezet, geven zij gezamenlijk weer wat wij vandaag de dag een persoonlijkheidsprofiel zouden noemen [4].

In de laatste 70 jaar zijn klassieke statistische methoden op verschillende manieren aangepast om ze geschikt te maken voor de specifieke eigenschappen van data verkregen in sociaal-wetenschappelijk onderzoek. Laatstgenoemd onderzoek resulteert namelijk zeer vaak in data die niet-kwantitatief zijn. Bij kwantitatieve informatie kan men denken aan gegevens verkregen door het meten van lengte en gewicht. In de sociale wetenschappen worden metingen vastgelegd op schalen, bijvoorbeeld op de zojuist genoemde rating scale, waarbij de meeteenheid onzeker is. Op een rating scale wordt met waarden tussen 1 en 5 of tussen 1 en 7 bijvoorbeeld aangegeven of een persoon een bepaalde eigenschap in grote mate dan wel helemaal niet bezit, of de respondent het met een bepaalde uitspraak helemaal eens of helemaal oneens is, en of dat iets heel erg belangrijk is of helemaal niet. Zulke data worden kwalitatieve of categorische variabelen genoemd. Het nulpunt van dergelijke schalen is onbepaald, en de relatie tussen de categorieën is ook niet helemaal bekend. Vaak kan wel worden aangenomen dat de categorieën geordend zijn (3 is meer dan 2 en 1, en 5 is meer dan 4), maar de onderlinge afstand tussen b.v. 1 en 2, en 4 en 5 is onbekend. Het meten in de natuurwetenschappen is gebaseerd op het tellen van standaard meeteenheden; het feit dat het in de gedragswetenschappen in veel gevallen om kwalitatieve gegevens gaat, maakt het meten hier wezenlijk anders. Het moge duidelijk zijn dat de onzekerheid m.b.t. de meeteenheid niet slechts een kwestie van meetfouten is.

De term datatheorie is een afkorting van het begrip relationele datatheorie. We beschouwen data als bepaalde relaties tussen individuen, groepen van individuen, en tussen individuen en variabelen of attributen [5]. In de datatheorie heeft altijd voorop gestaan dat kwalitatieve data geen geweld mogen worden aangedaan door ze te behandelen alsof ze van dezelfde aard zijn als data in de natuurwetenschappen.

Tegelijkertijd is het *wel* de bedoeling om méér met de data te doen dan ze slechts op *inhoudelijk* niveau te analyseren. Een belangrijke ontwikkeling in het vakgebied der datatheorie bestond er nu uit dat het mogelijk bleek te zijn om optimale kwantitatieve waarden voor dergelijke kwalitatieve schalen te vinden. Deze vorm van kwantificatie wordt dan ook wel 'optimal scaling' [6] genoemd.

Optimal scaling in de multivariate analyse kent historisch gezien twee belangrijke voorlopers. Aan de ene kant hebben we de geschiedenis van de techniek die tegenwoordig met correspondentieanalyse wordt aangeduid. Belangrijke vroege bijdragen tussen 1930 en 1960 zijn van de hand van Horst [7], Fisher [8], Guttman [9], en Hayashi [10]. Op welk idee voor de analyse van kwalitatieve gegevens waren deze gekomen? Laten we naar het klassieke voorbeeld van Fischer uit 1940 kijken. Stel, je weet van een groep personen de haarkleur en de oogkleur, en je wilt weten hoe sterk de samenhang tussen deze twee variabelen is. We kunnen moeilijk rekenen met de categorieën bruin, blauw, groen en grijs voor oogkleur en zwart, bruin, rood, en blond voor haarkleur. Maar laten we nu eens eerst willekeurige scores geven aan respectievelijk bruin, blauw, groen en grijs voor ogen, en vervolgens aan zwart haar een gemiddelde score toekennen die samengesteld wordt uit de scores van oogkleur naar evenredigheid met het aantal zwartharigen met bruine, blauwe, groene en grijze ogen. We doen hetzelfde voor bruin, rood, en blond haar. Vervolgens rekenen we nieuwe scores uit voor iedere kleur ogen op basis van de scores verkregen voor haarkleur. Enzovoort. We kunnen bewijzen dat een dergelijke iteratieve strategie tot een stabiel punt zal convergeren, waarbij de scores voor haar- en oogkleur op een bepaald moment niet meer veranderen, en dit is de optimale oplossing. In plaats van haar- en oogkleur kunnen we natuurlijk ook denken aan verschillende dagbladen en wijken in de stad, of, ook een echt voorbeeld, aan de samenhang tussen wijken en het bespelen van bepaalde muziekinstrumenten [11]. Met niet al te veel categorieën en veel geduld is een dergelijke oplossing nog wel met potlood en papier uit te rekenen. Voorwaar, een lumineus idee dat tegelijkertijd uitblinkt in eenvoud en genialiteit: we beginnen met kwalitatieve data (b.v. kleuren), en we eindigen met kwantitatieve data (optimale scores).

Een tweede, en voor de datatheorie misschien nog wel belangrijker invloed komt van het werk op het gebied van de meerdimensionele schaaltechnieken (MDS), waar in het begin van de jaren zestig niet-kwantitatieve, ordinale, gegevens geanalyseerd werden, door de pioniers Shepard [12], Kruskal [13], en Guttman [14]. In MDS worden een set gelijkenissen (bijvoorbeeld tussen personen) afgebeeld door een set afstanden in een laagdimensionele ruimte (het liefst in het platte vlak). De zogenaamde 'niet-metrische doorbraak' bestond eruit dat de in de optimale kwantificatie van de gelijkenissen de oorspronkelijke volgorde gehandhaafd bleef. Het betrof hier ingewikkelde te optimaliseren functies, waarbij het echt niet meer mogelijk was om de oplossing met de hand uit te rekenen [15]. Op de rol van de voortschrijdende, of accurater gezegd, voorthollende computertechnologie komen we later nog terug, aangezien deze het vakgebied kwalitatief veranderd heeft.

Sinds deze doorbraak in de MDS wordt de ordinale schaaltechniek, waarbij de oorspronkelijke volgorde van de variabele bewaard bleef, ook in de multivariate analyse toegepast. De datatheorie zegt nu dat we in het optimale schaalproces een geschikt kwantificatie niveau moeten *kiezen*; let op dat dit niet hetzelfde hoeft te zijn als het niveau waarop de data *gemeten* zijn. Twee variabelen kunnen ieder voor zich een ordinale of zelfs numerieke schaal vormen, zoals leeftijd en inkomensklasse. De relatie tussen deze twee variabelen is echter niet-lineair: jonge mensen verdienen eerst helemaal niets, en daarna weinig; in de daarop volgende leeftijdsklassen neemt het inkomen toe, om op een gegeven moment een top te bereiken, en na de, eventueel vervroegde, pensionering loopt het inkomen meestal weer terug. Een dergelijk krom, of niet-lineair, verband kan door het gebruik van de combinatie van nominaal en ordinaal schalingsniveau worden blootgelegd, waarbij de categorieën van de nominale variabele herordend mogen worden. Beroemde vroege bijdragen van het gebruik van optimale schaaltechnieken in de multivariate analyse zijn wederom van de hand van Kruskal [16] en Shepard [17], maar ook in *ons* land werd al in 1968 een belangrijke bijdrage geleverd door de in Leiden gepromoveerde, in Nijmegen beroemd geworden, en helaas te vroeg overleden, Eddy Roskam [18]. In de jaren '70 en '80 zijn er talloze bijdragen te vinden in de psychometrische literatuur. De Leidse 'Albert Gifi' groep ontwikkelde een veelomvattend multivariate analyse aanpak onder de naam Gifi-system [19].

Sinds de helft van de tachtiger jaren verschijnen de optimale schaalmethoden ook in de mainstream statistische literatuur, waarbij ik met name Breiman en Friedman [20], Ramsay [21], Buja [22], en Hastie en Tibshirani [23] wil noemen. In publicaties van hun hand in toonaangevende statistische tijdschriften zoals de *Annals of Statistics* en de *Journal of the American Statistical Association* wordt het Leidse Datatheorie werk erkend. In de negentiger jaren zijn de optimale schaaltechnieken verder uitgebreid en ondergebracht in een raamwerk van analysemethoden dat DTSS, Data Theory Scaling System, genoemd wordt. Zoals zijn voorgangers, is de hoofddoelstelling van DTSS het tegemoet komen aan specifiek sociaal wetenschappelijke data en hun eigenschappen. Qua nieuwe technieken kan men denken aan de afstandanalyse van kwalitatieve multivariate data, de analyse van sociale netwerk data door middel van grafen, en de combinatie van optimal scaling en het vinden van optimale één-dimensionele ordeningen [24].

Samengevat: In de datatheorie worden in eerste instantie zo weinig mogelijk assumpties gedaan, en wordt getracht aan kwalitatieve gegevens kwantitatieve schaalwaarden toe te kennen waarbij gestreefd wordt om de complexe relaties tussen individuen onderling, en tussen individuen en variabelen gezamenlijk, weer te geven in overzichtelijke representaties [25].

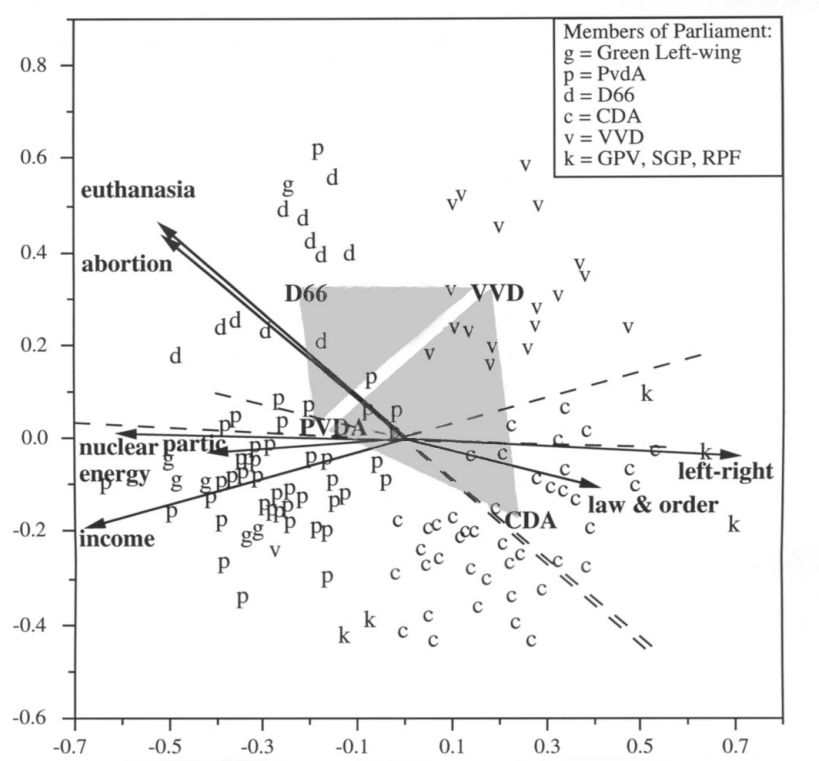
Te maken keuzes in de datatheoretische praktijk:
Een vector-model of een ideaalpunt-model?

De datatheorie maakt een onderscheid tussen beoordeling van bijvoorbeeld risico enerzijds en prestatie anderzijds. Voor de laatste variabele geldt gewoonlijk: hoe groter de prestatie hoe beter, en de te construeren schaal wordt het best door een monotoon stijgende functie weergegeven. De visualisering met het zogenaamde vectormodel geeft individuen weer als punten en variabelen als pijlen, en hun onderlinge relatie wordt gesymboliseerd door de projectie van de een op de ander [26]. We zullen hier een voorbeeld van zien. Voor een variabele als risico nemen we aan dat individuen verschillen, maar dat het overgrote deel der mensen een gemiddeld risico niveau zullen verkiezen boven een positie aan een van de extremen (helemaal geen risico of enorm veel risico). In plaats van een monotoon stijgende functie zegt de datatheorie dat dit soort gegevens met een enkeltoppige functie dienen te worden weergegeven: preferentie, bijvoorbeeld, neemt eerst toe om na een zeker niveau weer af te nemen. Een dergelijke enkeltoppige functie kan geometrisch het best worden weergegeven door een zogenaamd ideaalpunt-model. Hierin wordt een individu weergegeven als een punt met afstanden tot een stel andere punten die de opties weergeven waarvoor de voorkeur is uitgesproken. Een grote voorkeur in de data moet overeenkomen met een kleine afstand in de representatie. Een dergelijk ideaalpunt-model is zeker geschikt voor de preferentie voor bijvoorbeeld gezinssamenstelling. Gegeven allerlei combinaties van aantal meisjes en aantal jongens, zal een ieder na het krijgen van een bepaald aantal kinderen het gezin compleet achten, ook al is de ideale verhouding tussen meisjes en jongens niet bereikt. (Al bestaat er een gerucht dat, nadat Prinses Margriet geboorte had gegeven aan alwéér een jongen, de heer Pieter van Vollenhoven gezegd zou hebben: "We gaan door tot we een meisje hebben"; maar dit terzijde.)

Ik wil u graag een voorbeeld van een analyse laten zien waarin we het ideaalpunt-model en het vector-model gecombineerd hebben [27]. Het betreft een dataset uit 1990 waarin 136 leden van de Tweede Kamer zich hebben uitgesproken over hun sympathie voor de op dat moment vier grootste partijen in het parlement, te weten, de PvdA, het CDA, de VVD, en D66. In de gevonden twee-dimensonale oplossing, die gegeven wordt in Figuur 1, zien we een grote hoeveelheid informatie samengevat.

In de eerste plaats zijn daar de kamerleden zelf, weergegeven met hun ideaalpunt. Hoe kleiner de afstand tussen een kamerlid en een van de vier partijpunten, hoe groter de sympathie, en ik citeer uit het Hillebrand en Meulman hoofdstuk uit het boek *De geachte afgevaardigd*e: "In het algemeen bevinden kamerleden zich op niet al te grote afstand van hun eigen partij, al lijkt een enkel kamerlid enigszins 'verdwaald' te zijn. Dit is een rechtstreeks gevolg van het feit dat deze kamerleden hun eigen partij niet de hoogste sympathiescore gaven. Van de kamerleden van de VVD bevindt een groot gedeelte zich op ongeveer gelijke afstand van PvdA en CDA en ook

van de kamerleden van de PvdA bevindt een groot gedeelte zich op ongeveer gelijke afstand van de andere twee partijen. Onder de CDA-kamerleden is sprake van wat meer spreiding, maar het grootste deel bevindt zich iets dichter bij de PvdA dan bij de VVD. De kamerleden van D66 kunnen in twee groepen worden onderscheiden. Vijf kamerleden worden op kleine afstand van de PvdA aangetroffen. De overige zes D66-kamerleden bevinden zich iets dichter bij de VVD" [28]. "Tussen PvdA, CDA en VVD is sprake van een driehoeksrelatie, waarbij de drie partijen zich op vrijwel even grote afstand van elkaar bevinden. Ook tussen PvdA, D66 en VVD is sprake van een driehoeksrelatie, maar D66 neemt veel meer dan het CDA een positie tussen PvdA en VVD in. D66 en het CDA staan duidelijk diametraal tegenover elkaar" [29]. En kunnen we van verwondering spreken: in tegenstelling tot wat op *dat* moment (1990) veelal werd aangenomen, vallen het CDA en D66 *niet* samen in het centrum van de Nederlandse politiek, en is er al *helemaal* geen sprake van een één-dimensionale links-rechts structuur.



Figuur 1. Grafische representatie in een triplot: kamerleden, partijen en politieke issues in de Tweede kamer in 1990.

Dit blijkt ook uit de positie van de kamerleden op verschillende politieke issues. Deze zijn door middel van pijlen (vectoren) in de figuur weergegeven. Het begin en het eind van een pijl komt overeen met de uiteinden van de rating scales waarop de issues gedefinieerd waren. Zo wijst de abortuspijl in de richting van 'iedere vrouw heeft het recht hierover zelf te beslissen', en staat het begin van de pijl voor 'de overheid moet abortus onder alle omstandigheden verbieden'. Zo kan de antagonistische verhouding tussen D66 en het CDA verklaard worden uit de verschillen van inzicht inzake ethische vraagstukken als abortus en euthanasie. Ten aanzien van het vraagstuk van de inkomensverschillen vormen de PvdA (inkomenverschillen moeten kleiner) en de VVD (inkomensverschillen moeten zo blijven) elkaars tegenpolen. CDA en VVD liggen op de links-rechts dimensie dicht bij elkaar [30].

De gevonden twee-dimensionele structuur was aanleiding voor Hillebrand en Meulman om voorzichtig te opperen dat de tijd wel eens rijp zou kunnen zijn voor PvdA, D66 en VVD kabinet. Let wel, we schreven dit in 1991, toen het kabinet Lubbers/Kok aan het bewind was. Onze conclusie werd breed uitgemeten in een artikel in de Elsevier, dat als kop had "Kabinet zonder CDA komt dichterbij" [31]. Onze figuur werd in dit artikel letterlijk overgenomen, waarbij de journalist tot mijn grote verwondering een uitermate goede beschrijving gaf. De rest, inclusief de vorming van het eerste paarse kabinet, is geschiedenis.

Analyseren we de respondenten in de rijen of in de kolommen?

In de klassieke multivariate data analyse beschouwt men de kolommen van een datamatrix als de variabelen en de rijen als de personen, replicaties. Ook in de gedragswetenschappen is de onderzoeker veelal gevangen in dit paradigma. Mathematisch gezien lijkt het probleem geheel symmetrisch, en is er een klassieke aanpak via de lineaire algebra om rijen en kolommen geometrisch weer te geven zonder een enkele verdelingsassumptie te hoeven doen. Dit is dus *niet* de kwintessens van de datatheorie. Waar het in de datatheorie omgaat is dat het altijd de variabelen zijn die de objecten ordenen. In het klassieke psychometrische geval, ordenen items van een test de leerlingen naar hun talent, een niet-te-meten eigenschap die geschat wordt door de studie van het goed dan wel fout maken van bepaalde items. In het geval van beoordelingen of preferenties zijn het echter de *individuen* die ordenen. In de toegepaste datatheorie moet dus altijd eerst bepaald worden welke eenheden de rol van variabelen (de ordenende mechanismen) vervullen.

Een praktijk voorbeeld (uit de vroege Gifi tijd, omstreeks 1980). In een onderzoek geven respondenten op een rating scale aan in hoeverre bepaalde motieven een rol spelen om bloed te geven. Een 'geconditioneerde-reflex-analyse' met de respondenten in de rijen gaf vooral responsbias te zien. Sommige donoren geven allemaal scores van 3 of hoger aan alle motieven, variërend van het redden van mensenlevens, het regelmatig ondergaan van een geneeskundig onderzoek, tot het krijgen van een medaille. Andere respondenten lijken bijna niets echt belangrijk te vinden

(zij geven helemaal geen hoge scores). De datatheorie moet hier onmiddellijk roepen: wat zijn de objecten en wat zijn de variabelen in de analyse? Wie groepeert en ordent wat? Natuurlijk zijn het de donoren die de motieven scoren, en niet andersom. Analyseren we nu de data met dit uitgangspunt, dan zullen we een representatieve ordening te zien krijgen van de motieven die een rol spelen bij het geven van bloed bij de bloedbank.

Een ander mooi voorbeeld van deze 'omkering' zijn data verkregen voor een aantal pathologen die dezelfde uitstrijkjes beoordeeld hebben van 1 (geen afwijkingen) tot 5 (invasieve kanker in de baarmoeder). Hier zijn het de pathologen die de rol van variabelen moeten vervullen. Dan zal de data-analyse ons een consensus diagnose van de uitstrijkjes te zien geven, waarbij een patholoog die te vaak afwijkt van wat de anderen vinden, een kleiner gewicht in het uiteindelijke oordeel zal krijgen. In dit praktijkvoorbeeld liet optimale kwantificatie trouwens zien dat de meeste beoordelingsproblemen bestonden bij het onderscheiden van de categorieën 3 (carcinoom in situ) en 4 (zich verspreidende carcinoma cellen), hetgeen zeer belangrijke informatie is.

Definiëren nullen in data ook een homogene groep?

Naast rating scales wordt een zeer belangrijke plaats binnen de sociale wetenschappen ingenomen door zogenaamde binaire data, variabelen met twee categorieën, aangegeven met 1 en 0. Het prototypische voorbeeld zijn weer scores van geëxamineerden op een multiple-choice-test. Maken zij de vraag goed, dan krijgen ze een 1, en anders scoren we een nul. In dit geval is het niet moeilijk om aan te nemen dat geëxamineerden die op een bepaalde vraag een 1 scoren iets gemeenschappelijks hebben. Maar ook degenen die een 0 scoren vormen een homogene groep: zij maken immers allemaal de vraag fout. Door de analyse van de patronen in deze 1-0 data kan een grote kennis verkregen worden over de moeilijkheidsgraad van de items, en natuurlijk ook over de bekwaamheid van de geëxamineerden.

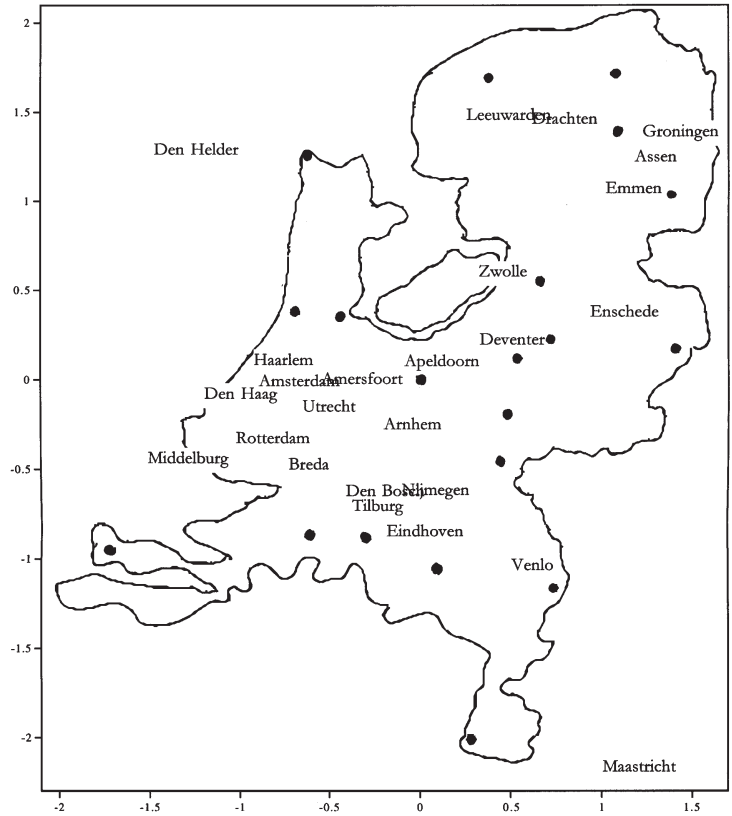
Er is echter een ander soort binaire data, waarbij respondenten in de 0 categorie géén homogene groep vormen. Denk bijvoorbeeld aan de historische stemming in de Tweede Kamer over de abortuswet tijdens het eerste kabinet Van Agt. Het wetsvoorstel kwam van de CDA-VVD regering, en had door de dominantie van de christen-democraten een aantal niet-liberale elementen. Bij de stemming in het parlement stemden de regeringspartijen vóór. De tegenstemmers waren kamerleden uit D66, de PvdA, de kleine linkse partijen en de kleine christelijke partijen. Vormden deze nu een homogene groep? Het antwoord is nee, want linkse kamerleden stemden tegen omdat zij een liberaler beleid t.a.v. abortus voorstonden, terwijl de klein-christelijke kamerleden vonden dat abortus *nooit* moest worden toegestaan. De aanname van homogeniteit van de tegenstemmers zal een desastreus effect op de data-analyse tot gevolg hebben.

Hoe kunnen we in verwondering geraken door de analyse van 1-0 data? Ik geef u een voorbeeld. Het betreft gelijkenissen tussen een set objecten, in dit geval steden in Nederland. De oorspronkelijke tabel geeft afstanden-op-de-weg tussen 25 steden [32]. In onze analyse hebben we de informatie drastisch samengevat door de gelijkenis tussen twee steden weer te geven met een 1 als hun afstand kleiner is dan de gemiddelde afstand in de hele 25x25 tabel, en met een 0 als hun afstand groter is dan het gemiddelde. De te analyseren data ziet u in Tabel 1.

Tabel 1. Gedichotomiseerde afstanden tussen 25 steden in Nederland. Een 1 geeft een kleine afstand aan en een 0 een grote afstand.

1	1	1	1	0	1	1	0	1	0	0	1	0	1	1	1	0	0	0	1	1	1	1	0	1
1	1	1	1	0	1	1	0	1	0	1	1	0	1	1	1	1	0	0	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	0	0	1	1	1	1	1	1
1	1	1	1	0	1	1	0	1	1	1	1	0	1	0	1	0	0	0	1	1	1	1	1	1
0	0	1	0	1	0	1	1	0	1	1	0	1	0	0	0	1	0	0	0	0	0	0	0	1
1	1	1	1	0	1	0	0	1	0	0	1	0	1	0	1	0	0	1	1	1	1	1	1	0
1	1	1	1	1	0	1	1	1	1	1	0	1	1	0	1	1	0	0	1	0	1	1	1	1
0	0	1	0	1	0	1	1	0	1	1	0	1	0	1	0	1	0	0	0	0	0	0	0	1
1	1	1	1	0	1	1	0	1	0	0	1	0	0	0	1	0	1	0	1	1	1	1	1	0
0	0	1	1	1	0	1	1	0	1	1	0	1	0	0	0	1	0	0	0	0	0	0	0	1
0	1	1	1	1	0	1	1	0	1	1	0	1	0	0	0	0	0	0	1	0	0	0	1	1
1	1	1	1	0	1	0	0	1	0	0	1	0	1	1	1	0	0	1	1	1	1	1	0	0
0	0	0	0	1	0	1	1	0	1	1	0	1	0	0	0	1	0	0	0	0	0	0	0	1
1	1	1	1	0	1	1	0	0	0	0	1	0	1	1	1	0	0	0	1	1	1	1	0	1
1	1	0	0	0	0	0	1	0	0	0	1	0	1	1	0	1	0	0	0	0	0	1	0	0
1	1	1	1	0	1	1	0	1	0	0	1	0	1	0	1	0	1	1	1	1	1	1	1	1
0	1	1	0	1	0	1	1	0	1	0	0	1	0	1	0	1	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	0
0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0	1	1	0	0	0
1	1	1	1	0	1	1	0	1	0	1	1	0	1	0	1	0	1	0	1	1	1	1	1	1
1	1	1	1	0	1	0	0	1	0	0	1	0	1	0	1	0	0	1	1	1	1	1	0	0
1	1	1	1	0	1	1	0	1	0	0	1	0	1	0	1	0	1	1	1	1	1	1	1	0
1	1	1	1	0	1	1	0	1	0	0	1	0	1	1	1	0	0	0	1	1	1	1	1	1
0	1	1	1	0	1	1	0	1	0	1	0	0	0	0	1	0	1	0	1	0	1	1	1	0
1	1	1	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	0	1	0	0	1	0	1

Als analyse methode is gekozen voor een correspondentie-analyse. Wordt regressie analyse doorgaans het werkpaard van de klassieke statistiek genoemd, dan is correspondentie-analyse het werkpaard van de grafische data analyse. Vaak zijn er voor specifieke situaties beter op de data toegesneden technieken, die echter veel gecompliceerder zijn. Correspondentie-analyse is relatief simpel, en doet het over de gehele linie nooit slecht. Het analyseresultaat vindt u in Figuur 2.



Figuur 2. Grafische representatie van de resultaten van een correspondentie analyse van de 1/0 gegevens uit Tabel 1. De plaatsnamen komen overeen met de coördinaten verkregen uit de analyse; de stippen komen overeen met de werkelijke kaart van Nederland die over de figuur heen getekend is.

De echte positie van de steden op de landkaart is over het resultaat van de correspondentie analyse van de 1-0 data getekend. Het zal duidelijk zijn dat de verminking van de precieze afstanden in kilometers tot enen en nullen een vertekening geeft. Dit geldt met name voor steden die overal ver vanaf liggen: kijk naar Middelburg, Den Helder, en Maastricht. Maar over het algemeen is de reconstructie van de kaart van Nederland opmerkelijk; en dit alles tot onze grote verwondering gekregen door alleen naar enen en nullen te kijken.

Het individu in de statistiek

In de toegepaste data theorie hebben we nooit te *weinig* individuen en ook nooit te *veel* variabelen. Een voorbeeld uit de praktijk dat mij erg lief is, is de data set die indertijd door Din van Strien en Thecla van der Ham aan mij werd overlegd. De data werden verzameld in de psychiatrische afdeling van het Academisch Ziekenhuis Utrecht, en betroffen vragen over het welbevinden van patiënten met verschillende vormen van eetstoornissen [33]. Deze vragen waren op vier verschillende tijdstippen voorgelegd, bij intake, en na 1, 2 en 4 jaar van behandeling. Het was niet bekend echter welke behandeling de patiënten gehad hadden. Beschikbaar waren alleen de data verkregen bij 56 jonge vrouwen, die op 16 vragen over hun welzijn geantwoord hadden op een 5-puntschaal van goed tot slecht, variërend van menstruatie status, gezinsleven, overgeven, en seksualiteit. Dus, we hebben 56 personen met hun scores op 16 x 4 (tijdstippen) = 64 variabelen. Een horror voor de klassieke statistische aanpak die zegt dat je altijd veel meer personen dan variabelen moet hebben. De toegepaste datatheorie is óók blij met veel personen t.a.v. het aantal variabelen, maar als de realiteit daar niet aan voldoet, omdat we b.v. nu eenmaal niet méér patiënten kunnen creëren, kan zij nog steeds veel met dergelijke data doen.

In het geval van de eetstoornispatiënten is het wonder dat er een prachtige twee-dimensionale structuur gevonden werd, waarbij de eerste dimensie het verloop in de tijd aangaf, en de tweede dimensie de verschillende vormen van eetstoornissen. Hierbij werd een clustering gevonden van anorexia en atypische eetstoornis patiënten enerzijds, en bulimische met anorexia en bulimia na anorexia patiënten anderzijds. De analyse liet zien dat er een aantal kernvariabelen en een aantal specifieke variabelen waren, en dat de patiënten in de tijd dichter bij elkaar kwamen, d.w.z. zich ieder verbeterden op hun specifieke probleemvariabelen (waarbij de een groep het wel veel beter bleef doen dan de andere groep), maar zij ook gezamenlijk op de gemeenschappelijke probleemvariabelen vooruitgingen.

Analyseren met de data die je hebt

Voordat er iets geanalyseerd kan worden, zijn de data eerst verzameld, meestal zonder veel aandacht voor de daarna te volgen analyse. Dit is jammer, maar het is niet anders. Nog erger is het als er bovendien in de dataverzameling zelf van alles mis is gegaan. Denk bijvoorbeeld aan een uiterst belangrijke ingrediënt in de wetenschappelijke methode, nl. de mogelijkheid tot replicatie. Hiermee wordt bedoeld het tot in detail herhalen van een oorspronkelijke onderzoek om te kunnen verifiëren of het resultaat hetzelfde is. In een dergelijke onderzoeksopzet moeten de onafhankelijke variabelen constant gehouden worden, maar in veel sociaal-wetenschappelijk onderzoek zorgt de weerbarstige werkelijkheid ervoor dat de geobserveerde situaties nooit precies hetzelfde zijn. Hier zullen we mee moeten leven. Maar ook in andere vakgebieden heeft men soms problemen om de condities constant te houden. Zie het volgende voorbeeld [34].

De Amerikaanse federale luchtvaartdienst gebruikt een bepaald apparaat om de sterkte van de voorruit in de cockpit van nieuw ontwikkelde vliegtuigen te testen. Het apparaat lijkt nog het meest op een kanon. Hiermee worden dode kippen tegen de voorruit van een stilstaand vliegtuig geschoten. Door de snelheid van het schieten te variëren kan de snelheid van het vliegtuig in de lucht worden nagebootst. Het idee is natuurlijk dat als de ruit niet breekt door de impact van de dode kip, de ruit ook een botsing met een levende vogel tijdens de vlucht kan doorstaan.

Britse ingenieurs lazen over deze test, en wilden hem aanpassen om de voorruit van een nieuwe hogesnelheidstrein te testen. Zij construeerden een soortgelijk kanon, plaatsten de eerste dode kip erin, en schoten met hoge snelheid op de voorruit van de trein. De kip versplinterde de ruit, verpletterde het instrumentenpaneel aan de achterzijde, en boorde zich in de achterwand van de cabine van de machinist. De Britten waren perplex en vroegen de Amerikaanse luchtvaartdienst om de door hen gevolgde procedure zorgvuldig te bekijken om te zien wat zij eventueel fout hadden gedaan. De Amerikaanse luchtvaartdienst bekeek de test zeer grondig en had slechts één advies: "voortaan de kip eerst ontdooien".

In het ideale geval worden bij herhaalde meting de variabelen constant gehouden. In de praktijk zien we ons geconfronteerd met bijvoorbeeld longitudinaal onderzoek waarbij sommige van de indicatoren door de tijd heen niet op dezelfde manier gemeten zijn. Moeten we deze data dan maar weggooien? Natuurlijk niet. In een studie naar de ontwikkeling van te-vroeg-geborenen, hebben we eerst een data matrix geanalyseerd, waarbij alle variabelen op alle tijdstippen voorkomen [35]. Hiermee konden we zien of variabelen op verschillende tijdstippen clusterden, om daaruit af te leiden dat zij hetzelfde meetten. Daarná werd de datamatrix zo herordend dat alle variabelen maar één keer voorkwamen, maar nu de *kinderen* meerdere malen, d.w.z. op ieder van de tijdstippen. Nu konden we wél het verloop van verschillende groepen kinderen in de tijd bestuderen. Een dergelijk aanpak gaat natuurlijk ook van allerlei aannamen uit, en het moge duidelijk zijn dat zo'n analyse-strategie alleen met de grootste zorgvuldigheid kan worden uitgevoerd.

Oude wijn in nieuwe zakken

Tegenwoordig hebben we bijna allemaal met de resultaten van geavanceerde data analyses te maken, al hebben we dit soms niet door, en weten we al helemaal niet hoe het werkt. Een mooi voorbeeld is de vaste klantenkaart die geïntroduceerd werd in de supermarkt van de Amerikaanse supermarktketen Safeway [36]. Men krijgt als klant een pasje, waarmee men korting op allerlei producten krijgt. Om het pasje te krijgen moet de klant een formulier invullen waarbij zijn of haar adres en allerlei demografische gegevens gevraagd worden. Iedere keer dat er boodschappen worden gedaan, wordt er via het pasje automatisch opgeslagen welke producten de vaste klant gekocht heeft. De supermarkt zelf is niet zozeer geïnteresseerd in het kop-

pelen van individuele gegevens met bepaalde producten, maar verkoopt deze kennis aan de fabrikanten van de producten, voor wie marktsegmentatie van het grootste belang is [37]. Deze gang van zaken doet natuurlijk denken aan het systeem dat nu ook bij Albert Hein in gebruik is.

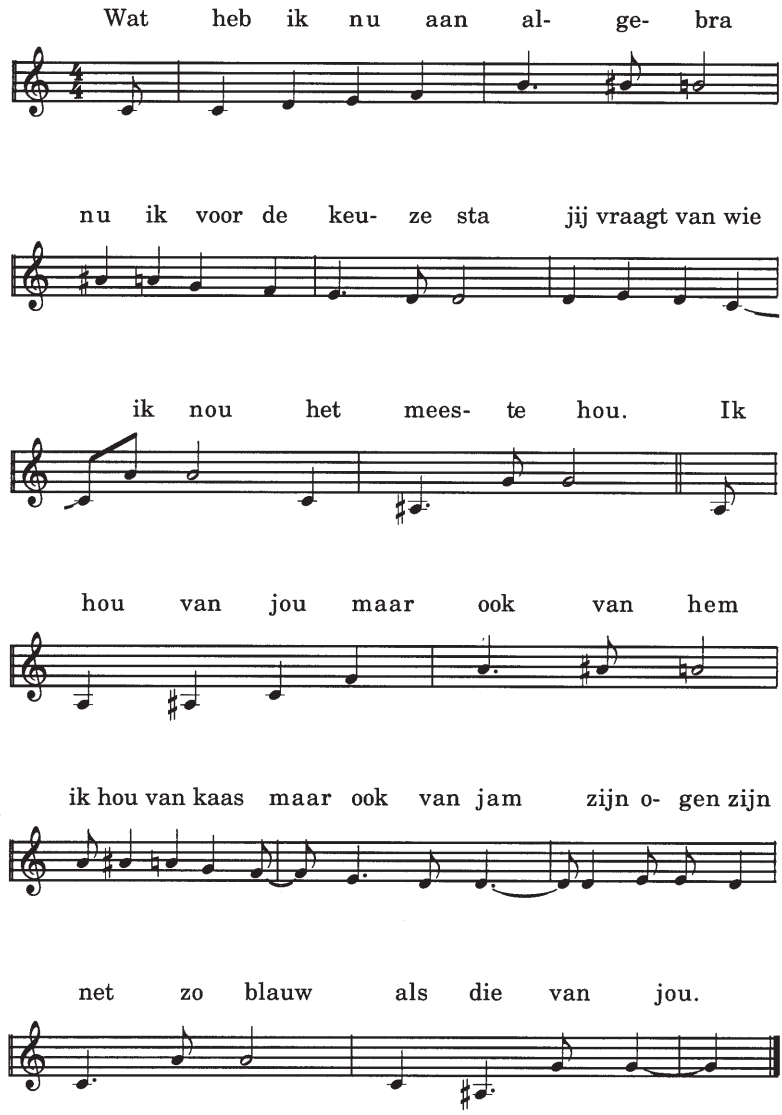
Terug naar het onderwerp, waarbij de verzameling van gegevens in de supermarkt maar één van de voorbeelden is. De huidige automatiseringstechnologie maakt het mogelijk dat er een enorme hoeveelheid gegevens kan worden opgeslagen, en dit heeft ook een nieuwe klasse van analysemethoden in het leven geroepen die met de trendy naam 'data mining' wordt aangeduid. Werd er in de traditionele mijnbouw kolen gedolven, en natuurlijk ook goud, in de huidige 'hype' worden er data ontgonnen. Soms wordt data mining als equivalent beschouwd van de zogenaamde Knowledge Discovery in Databases, afgekort met KDD, soms wordt er een onderscheid gemaakt. Het gewichtiger klinkende Knowledge Discovery in Databases zou zich bezig houden met (en ik citeer) "het niet-triviale proces van het vinden van valide, nieuwe, in potentie bruikbare, en uiteindelijk begrijpelijke patronen in data, terwijl data mining slechts een stap in het KDD proces is, die bestaat uit het toepassen van een bepaald data mining algoritme" [38]. In alle eerlijkheid kan ik hier niet echt iets nieuws in zien als ik die doelstelling vergelijk met die van de datatheorie, en de gebruikte statistische methoden. Qua het gebruik van de mogelijkheden van de computer bestaat er wél een verschil; als we in de geschiedenis van de datatheorie terugkijken naar het oorspronkelijke boek *Theory of Data* van Coombs uit 1964 zien we dat dáár de invloed van de computer eigenlijk nog helemaal niet merkbaar was. Data mining bestaat bij de gratie van de waarlijk geëxplodeerde computertechnologie, en lijkt vooral over *hele grote* data sets te gaan. Er komt een hoop macho gedoe aan te pas, waarbij de een de ander voorhoudt: "mijn dataset is veel groter dan de jouwe" [39]. Maar de grootte van de dataset is natuurlijk bijzaak. Ze kunnen wel heel groot zijn, maar vaak wordt er toch maar slechts een gedeelte van geanalyseerd. Kortom, uiteindelijk zijn data mining technieken niet veel anders dan de technieken die wij uit de hedendaagse moderne statistiek kennen en die een onderdeel vormen van de door ons voorgestane datatheorie. Data mining maakt gebruik van regressie, principale componenten analyse, correspondentie analyse, cluster analyse, neurale netwerken, beslissingsbomen, associatieregels, en 'market basket analysis': de analyse van het boodschappenmandje. De laatste term is in ieder geval wél uniek. Kort gezegd komt deze analyse erop neer, dat men probeert uit te vinden hoe groot de kans is dat, als u het ene product koopt, u ook een bepaald ander product zult kopen. De winkelier zal dan geneigd zijn om deze twee producten naast elkaar in de schappen te zetten; en voor de fabrikant van producten is dit een aanleiding voor segment-gerichte reclame binnen de consumentenpopulatie [40].

Ik zal openhartig tegen u zijn. In eerste instantie kon ik alleen maar mijn neus ophalen voor deze hele data mining manie, en er het liefst met veel dédain over spreken. Alles klinkt ook wel heel erg trendy, en de doelstellingen zijn bijna altijd ook

wel heel erg commercieel. Maar, vroeg ik mij in alle eerlijkheid af, kunnen we het hier helemaal mee afdoen? U moet weten dat één van de doelstellingen binnen onze onderzoeksgroep het ter beschikking stellen van de door ons ontwikkelde computerprogramma's voor een groot publiek is. En, dat doen wij niet gratis. Onze computerprogramma's zijn opgenomen in het wereldwijd vanuit Chicago verspreide software pakket SPSS [41]. En van de verkoop krijgen wij royalties, die op dit moment het mogelijk maken om de aanstelling van een aantal onderzoekers te financieren, om geregeld de modernste computers aan te schaffen, om met zijn allen naar congressen te gaan, ook ver weg, kortom, allerlei uitgaven te doen die in het huidige bestel van de universiteit al lang niet meer standaard gedaan kunnen worden. Dit geeft ons een enorme vrijheid. Geregeld doet zich nu de vraag voor of de door ons ontwikkelde programma's eigenlijk óók geen data mining technieken zijn. En hier zou ik dan eigenlijk geen ontkennend antwoord op moeten geven, omdat zij wel degelijk ontwikkeld zijn om patronen tussen veel individuen en veel kenmerken op te sporen door middel van visualisatie. Daarom lijkt het meer gepast om niet neerbuigend over data mining te doen, maar om de daarin gebruikte technieken aan te vullen met onze specifieke inbreng die traditioneel toch meer statistisch georiënteerd is. De standaard integratie van resampling methoden zoals de bootstrap, jackknife, en kruisvalidatie is hierbij onontbeerlijk [42]. Tegelijkertijd is het de doelstelling om nog meer aandacht te besteden aan de verdere overbrugging van de afstand tussen de traditionele statistiek en de optimal scaling technieken.

Qua data mining staat er overigens een zeer opmerkelijke passage in het nonfictie boek *The Spycatcher* van de voormalig Britse geheime dienst agent Peter Wright [43]. De Britten hadden een enorme hoeveelheid data verkregen door het onderscheppen van geheime berichten vanuit Moskou. Een jonge decodeerdeskundige paste clusteranalyse toe om gelijkenissen tussen de duizenden berichten te analyseren. Opnieuw verwondering: een van *onze* data-analyse technieken resulteerde in "een van de meest invloedrijke werktuigen van de westerse contraspionage" [44].

Wat heb ik nu aan algebra
(met dank aan zangeres Loeki Knol)



Onderwijs in de Statistiek

Ik heb het lied *Wat heb ik nu aan algebra* altijd al eens willen zingen, omdat de soepele overgang van zaken als algebra, naar liefde, broodbeleg en blauwe ogen zo absurd en tegelijkertijd vermakelijk is. Op een iets serieuzer niveau staat de tekst echter symbool voor een verontrustende gedachte, namelijk, dat als het er écht op aan komt in het leven, je niets aan exacte vakken hebt. Hoewel natuurlijk van een geheel ander kaliber, niet triviaal, maar wel verontrustend, is het recente interview in de NRC met Professor Len De Klerk bij zijn afscheid als rector van de Katholieke Universiteit Brabant. De Klerk promoveerde in Leiden in de experimentele psychologie bij Van de Geer. Hij werd daarna lector Methodenleer aan de universiteit van Amsterdam, vervolgens hoogleraar onderwijspsychologie in Tilburg, en tenslotte rector van de KUB. Voorwaar, van iemand met een dergelijk mooie carrière mag men toch een warm pleidooi voor onderwijs in de methoden en technieken verwachten. Niets blijkt minder waar.

De Klerk zegt een voorstander te zijn van het Angelsaksische onderwijssysteem - een bachelorsgraad na drie jaar en een masters degree na vier of vijf jaar. Tot het Bachelors diploma zouden de studenten grondige kennis moeten opdoen van het vakgebied waarvoor ze gekozen hebben. En ik citeer De Klerk: "Alleen dan kunnen ze zich erin gaan thuisvoelen. Pas later komen 'randvakken' als geschiedenis, wijsbegeerte, en methodologie aan de orde, die nu vaak al in de propedeuse worden gegeven. Die zijn weliswaar onmisbaar, maar behoren niet tot de *kern* van de opleiding. Eerst de stam verkennen en dan pas de takken en de bladeren"[45]. Ik heb grote moeite met deze opvatting, en ik meen dan ook dat De Klerk zich vergist: de methoden en technieken behoren niet tot de *takken* of de *bladeren* van de boom, maar vormen de *wortels*!

Ik zou dus willen pleiten voor een sterk kwantitatief onderdeel in de beginfase van de studie. Dit vormt het fundament voor een gedegen academische opleiding. Natuurlijk hoeft er van mij geen statistiek en wiskunde in abstracto onderwezen te worden. De relatie met de pedagogiek, bijvoorbeeld, moet te allen tijde aanwezig zijn. Dit maakt het vak natuurlijk nog steeds niet eenvoudig: het is en blijft moeilijk, zeker voor de beginnende student. Maar het met succes volgen van dergelijke vakken onderscheidt de betere student van de middelmatige, en daarom vallen kwantitatieve methoden en technieken naar mijn mening prima in het koersen op kwaliteit profiel van de Leidse universiteit. Met grote instemming heb ik dan ook het rapport *Vitaliteit en kritische massa: strategie voor de natuur- en technische wetenschappen* van de adviesraad voor het wetenschappelijk en technisch onderwijs gelezen [46]. Dit rapport kwam in het nieuws omdat de conclusie was dat het huidige aantal b-opleidingen gehalveerd zou moeten worden. Maar er wordt ook een versterking van de b-component in het α- en γ-onderwijs geadviseerd, waar steeds meer onderdelen van het onderzoek zich lenen voor modelmatige analyses. Daarom is het volgens de com-

missie van groot belang dat een deel van het β-talent na het vwo-examen kiest voor een α- en γ-studie. Hier is natuurlijk een grote kans weggelegd voor met name vrouwelijk β-talent, dat er in eerste instantie om allerlei redenen niet voor kiest om een puur β-vak te gaan studeren.

De automatiseringstechnologie grijpt om zich heen, en bijna niemand kan zich er aan onttrekken. Daarom is het noodzaak dat we te allen tijde kunnen uitleggen wat voor statistische methoden gebruikt worden als het grote publiek ermee te maken krijgt [47]. In de Verenigde Staten is het fenomeen 'credit scoring' bijvoorbeeld een hot issue, waarbij voor iedere vorm van kredietverlening razendsnel een creditscore wordt vastgesteld, gebaseerd op een gewogen combinatie van allerlei gegevens, zoals het financiële verleden van de consument en zijn of haar demografische gegevens. Het is onduidelijk wat een goede score is, hoe je je score te weten kunt komen, en wat te doen als je een te lage score hebt. Eigenlijk wil men precies weten hoe een dergelijke score precies berekend wordt. De geldleenbranche heeft geen echte antwoorden op deze vragen. Laatst werd het weer eens geprobeerd. Aan het woord is de vice-president van een van de drie grootste credit reporting bureaus in de Verenigde Staten. Op de vraag hoe credit scoring nu eigenlijk werkt antwoordde hij: "credit scoring is niet zo geschikt om te worden uitgelegd *want* het is een statistische methode" [48]. De uitdaging is aan ons om wél soortgelijke moeilijke materie aan een breed publiek uit te leggen.

Slot

Aan het eind van mijn rede gekomen, wil ik graag mijn dank uitspreken aan de leden van het Bestuur van het Leids Universiteits Fonds voor het vertrouwen dat zij in mij hebben gesteld door mij te benoemen tot bijzonder hoogleraar Toegepaste Datatheorie, in het bijzonder de multivariate analyse van kwalitatieve gegevens. Velen hebben aan de totstandkoming van deze leerstoel bijgedragen, in het bijzonder de Rector Magnificus, het College van Bestuur, het College van Decanen, en de huidige en vorige decaan van de Faculteit der Sociale Wetenschappen. Ook wil ik de leden van de Commissie Wetenschappelijke Bestedingen bedanken, alsmede het dagelijks bestuur van de vakgroep Pedagogische Wetenschappen.

Hooggeleerde Van de Geer, beste John,

Het is al meer dan 20 jaar geleden dat ik je student-assistent werd, maar eraan terugdenkend werden de herinneringen heel levendig. Een van de dingen die we samen deden was het adviseren van promovendi, bijvoorbeeld van buiten de faculteit. Eén daarvan staat me nog helder voor ogen; het betrof een proefschrift in de geschiedenis van de oudchristelijke kunst; het doel was om de sarcofagen met afbeeldingen van Jonas en de walvis te ordenen in de tijd. Bij het kijken in mijn oude

mapjes vond ik een door ons samen opgestelde lijst met data-analytische aantekeningen t.b.v. de promovendus. Ik citeer hieruit:

> 1. De conclusie over de orans is dubieus, mede i.v.m. de positie van de stuurman op de achtersteven, en de extreem lange mouwen van de tunica van de orans.
> 2. Variabelen *d* en *e*, *g* en *h*, *k* en *l*, *m* en *n*, en *o* en *p* zijn wederzijds uitsluitend: niet opnemen als aparte variabelen maar telkens als 1 variabele met twee categorieën.
> 3. Variabelen *i* en *j* opsplitsen in 3 variabelen: 1) plaats van stuurman en orans. 2) kleding orans 3) kleding stuurman. Is de stuurman niet naakt op 891 (wordt niets over gezegd). Heeft stuurman 747 wel mouwen? (zou uniek kenmerk zijn).

Dit is een kleine illustratie van één van de hoofdzaken die ik van je geleerd heb: n.l. dat het voor de datatheorie niet uitmaakt uit welk vakgebied de data komen, omdat de onderliggende principes voor de analyse hetzelfde zijn. Ik heb altijd geprobeerd om de door jou ingezette lijn voort te zetten om de datatheorie buiten de psychologie, en ook buiten de sociale wetenschappen, toe te passen. De datatheorie in Leiden is jouw geesteskind; in heel de wereld heeft men altijd met bewondering gekeken naar de afdeling die jij in 1969 oprichtte. Op het moment is de datatheorie opgesplitst over psychologie en pedagogiek, maar ik zie dat niet als een ongunstige ontwikkeling. Ik zal er alles aan doen om de naam van de datatheorie, neergelegd in mijn leeropdracht, hoog te houden.

Zeer gewaardeerde PIONIER medewerkers, beste Peter, Susañña, Patrick, Jacques, Anita, Bart-Jan, Elise, Halima,

Ik heb jullie in een bepaalde chronologische volgorde genoemd, waarbij de langste samenwerking al 19 jaar bestaat. Jullie inzet, enthousiasme en toegewijdheid voor het vak kennen weinig grenzen. Jullie zijn een heel bijzondere groep, en ik prijs mij gelukkig dat ik de afgelopen jaren jullie projectleider heb mogen zijn. In dit kader wil ik ook NWO bedanken. De aan mij gegunde PIONIER-subsidie heeft een van de mooiste periodes in mijn wetenschappelijke carrière ingeluid. Sommige mensen denken dat het project eind dit jaar afloopt, maar dit is een vergissing. We hebben zojuist van het College van Bestuur een stimuleringssubsidie ontvangen, en we zullen nieuwe aanvragen voor onderzoeksgeld blijven schrijven. Ik ga ervan uit dat we als groep nog vele vruchtbare jaren samen tegemoet gaan. Dit geldt uiteraard óók voor mijn nieuwe collega's en studenten in de pedagogische wetenschappen. Ik zie met veel plezier naar onze samenwerking uit.

Hooggeleerde Wagenaar, beste Willem Albert,

Hoewel niet een directe leermeester, ben je op een aantal cruciale punten in mijn carrière nadrukkelijk aanwezig geweest. Zo herinner ik mij mijn dissertatiecommissie, de nominatie voor de J.C. Ruigrokprijs, het NWO gebiedsbestuur ten tijde van mijn PIONIER-subsidie, en nu het instellen van deze bijzondere leerstoel. Ik ben je uitermate dankbaar voor het support dat ik al die jaren van je heb mogen ondervinden.

Tenslotte wil ik je nog even extra als rector bedanken voor het feit dat ik mijn oratie op deze bijzondere dag mocht houden. Voor de universiteit is deze dag bijzonder omdat het nieuwe academisch jaar pas aanstaande maandag 6 september begint, en er gewoonlijk pas daarna weer oraties plaatsvinden. Voor mij persoonlijk is 3 september al heel lang een bijzondere dag, aangezien het de trouwdag van mijn ouders is. Ik ben dankbaar dat zij beiden hier aanwezig zijn. Zonder hen had ik hier niet gestaan.

Ik heb gezegd.

[1] Hugo Claus (1962). *De Verwondering*. Amsterdam: De Bezige Bij.

[2] Guilford, J.P. (1936). *Psychometric methods*, p. 264. New York: McGraw-Hill.

[3] Ellson, D.G. & Ellson, E.C., Historical note on the rating scale, Psychological Bulletin, 50, 1953, pp. 383-384.

[4] Ibid., p. 383.

[5] Onder het kopje "Datatheorie" vinden we in: Coombs, C.H., Dawes, R.M., & Tversky, A. (1970). *Mathematical psychology: An elementary introduction*. Englewood Cliffs, NJ: Prentice-Hall: "Many measurement models in the behavioral sciences are based on geometric representations of the observed behavior. Frequently this geometric representation is a one-dimensional scale but it need not be, and multidimensional representations are becoming more common. The points on these scales or in these spaces may represent individuals or stimuli or both, and the relations among the points reflect the observations according to some rule." (o.c., p.32).

[6] De term "optimal scaling" danken we aan R. Darrell Bock van de University of Chicago. De term is populair geworden door het ALSOS (alternating least squares – optimal scaling) systeem van Young, De Leeuw, & Takane. Zie, bijvoorbeeld:
Young, F.W., De Leeuw, J. & Takane, Y. (1976), Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling features, *Psychometrika, 41,* 505-528.

[7] Horst, P. (1935). Measuring complex attitudes. *Journal of Social Psychology, 6,* 369-374.

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

[9] Guttman, L. (1941). The quantification of a class of attributes: a theory and method of scale construction. In P. Horst et al.(Eds.), *The prediction of personal adjustment*, pp. 319-348. New York: Social Science Research Council.

[10] Hayashi, C. (1952). On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statistical Mathematics, 2,* 93-96.

[11] Een onderzoek uitgevoerd in het begin van de tachtiger jaren door W.A. Wagenaar.

[12] Shepard, R.N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function I. *Psychometrika, 27,* 125-140. II. *Psychometrika, 27,* 219-246.

[13] Kruskal, J.B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29,* 1-28.

[14] Guttman, L. (1968). A general nonmetric technique for finding the smallest coordinate space for a configuration of points, *Psychometrika, 33,* 469-506.

[15] Zoals Roger Shepard in *Science* beschrijft: "At the Bell Telephone Laboratories, I began in 1960 to explore a new approach to multidimensional scaling, called "analysis of proximities," that proved capable of overcoming the limitations of the earlier approaches. I used a one-stage iterative method .... Following a few adjustments, on 17 March 1961 the iterative process with which I had been experimenting finally converged to its first stationary configuration (at just 2:33 p.m. EST, according to the computer log)." In: Roger N. Shepard (1980), Multidimensional Scaling, Tree-fitting and Clustering. *Science, 210,* p. 328.

[16] Kruskal, J.B. (1965). Analysis of factorial experiments by estimating monotone transformations of the data. *Journal of the Royal Statistical Society Series B, 27,* 251-263.

[17] Shepard, R.N. (1966). Metric structures in ordinal data. *Journal of Mathematical Psychology, 3,* 287-315.

[18] [Roskam, E.E.C.I. (1968). *Metric analysis of ordinal data in psychology.* Voorschoten: VAM.

[19] Gifi, A. (1990). *Nonlinear multivariate analysis.* Chichester: John Wiley and Sons.

[20] Breiman, L., & Friedman, J.H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association, 80,* 580-598.

[21] Ramsay, J.O. (1988). Monotone regression splines in action. *Statistical Science, 3,* 425-441.

[22] Buja, A. (1990). Remarks on functional canonical variates, alternating least squares methods and ACE. *Annals of Statistics, 18,* 1032-1069.

[23] Hastie, T., Tibshirani, R. & Buja, A. (1994). Flexible discriminant analysis. *Journal of the American Statistical Association, 89,* 1255-1270.

[24] Meulman, J.J., Hubert, L.J. & Heiser, W.J. (1998). The Data Theory Scaling System. In: A. Rizzi, M. Vichi & H.H. Bock (Eds.), *Advances in Data Science and Classification*, pp. 489-496. Berlin: Springer Verlag.

[25] Bij het gebruik van multivariate analyse als een multidimensional scaling (visualisatie) techniek denken we met name aan de enorme bijdrage van John Gower, te beginnen met: Gower, J.C. (1966). Some distance properties of latent roots and vector methods used in multivariate analysis, *Biometrika, 53,* 325-338.

[26] The gezamenlijke weergave van personen en variabelen in een laag-dimensionele ruimte wordt veelal aangeduid met de naam "biplot", voorgesteld in: Gabriel, K.R. (1971), The biplot graphic display of matrices with application to principal components analysis, *Biometrika, 58,* 453-467. Het idee is echter al terug te vinden in: Tucker, L.R (1960), Intra-individual and inter-individual multidimensionality, In: H. Gulliksen & S. Messick (Eds.), *Psychological Scaling: Theory and Applications.* New York: Wiley. Hierin staat: "I wish to outline an approach which extends the Thurstonian unidimensional paired comparison theory to inclusion of multidimensional study of individual differences. The scale for each individual is still considered as unidimensional … Multidimensionality for a

group of individuals can be introduced by using a vector for each person and allowing the vectors for many people to have various directions in the space. The stimuli may be represented by vectors in the same space … The basic model is to represent each person's scale value for a given stimulus by the scalar product between his vector and the stimulus vector." (o.c., pp. 158-159). Het vector model is veel toegepast in de analyse van preferentie data, bijvoorbeeld, al in: Carroll, J.D. (1972). Individual differences and multidimensional scaling, In: *Multidimensional scaling: Theory and applications in the behavioral sciences* (ed. R.N. Shepard, A.K. Romney & S.B. Nerlove), Vol. 1, 105-155. New York and London: Seminar Press.

[27] Hillebrand, R., & Meulman, J.J. (1992). Afstand en nabijheid: verhoudingen in de Tweede Kamer. In: J.J.A. Thomassen, M.P.C.M. Van Schendelen, and M.L. Zielonka-Goei (Eds.), *De geachte afgevaardigde ... hoe kamerleden denken over het Nederlandse parlement*, pp. 98-128. Muiderberg: Coutinho.

[28] Ibid., pp. 115-116.

[29] Ibid., pp. 114-115.

[30] Ibid., p. 118.

[31] Jansen, J. (1992), Kabinet zonder CDA komt dichterbij: kamerleden kiezen voor coalitie PvdA-VVD-D66, *Elsevier*, pp. 12-15.

[32] Suurland's Autokaart. Eindhoven: Suurland's Vademecum B.V.

[33] Van der Ham, Th., Meulman, J.J., Van Strien, D.C. & Van Engeland, H. (1997). Empirically based subgrouping of eating disorders in adolescents: a longitudinal perspective. *British Journal of Psychiatry, 170,* 363-368.

[34] "From a recent issue of *Meat and Poultry* magazine, editors quoted from *Feather,* the publication of the California poultry industry federation".

[35] Theunissen, N.C.M., Meulman, J.J., den-Ouden, A.L., Koopman, H.M., Verrips, E.G.H.W., Verloove-Vanhorick, S.P., & Wit, J.-M., Changes in quality of life can be studied when the measurement instrument is different at different time points *(Submitted).*

[36] Berry, M.J.A. & Linoff, G. (1977). Data mining techniques, for marketing, sales, and customer support, pp. 11-12. New York: John Wiley & Sons.

[37] Ibid., p. 12.

[38] Fayyad et al. (1996), Geciteerd in: Gaul, W. & Schader, M. (1999). Data Mining: A new label for an old problem, p. 5. In: W. Gaul & M. Schader (Eds.), *Mathematische Methoden der Wirtschaftswissenschaften.* Festschrift für Otto Opitz, pp. 3-14. Physica-Verlag.

[39] Friedman. J. (1997), Data Mining and Statistics: What's the connection? p. 7.

[40] Berry, M.J.A. & Linoff, G. (1977). Data mining techniques, pp. 119-120. New York: John Wiley & Sons.

[41] Meulman, J.J., Heiser, W.J. & SPSS (1999). *SPSS Categories 10.0.* Chicago: SPSS Inc.

[42] Een statische innovatie, de bootstrap, haalt de New York Times van 8 November,

1988: “A new technique that involves powerful computer calculations is greatly enhancing the statistical analysis of problems in virtually all fields of science.” … “Dr. Efron’s idea allows statisticians to use unorthodox methods to see patterns in data because now, for the first time, statisticians can assess how accurate those methods are.”

[43] Wright, P. (1988). *Spycatcher*. New York: Dell Pub Co.

[44] “I explained that the main weakness was that despite the breakthrough offered by my classification of illegal broadcasts from Moscow, GCHQ had insufficient coverage of the traffic ... we still had only twelve and fifteen radio positions intercepting these signals ...Tordella was much taken by the possibilities, and agreed to guarantee a worldwide take of 100 percent for at least two years. He was as good as his word, and soon the intelligence was flooding back to the Counterclan Committee. A young GCHQ cryptanalyst named Peter Marychurch (now the director of GCHQ) transformed my laborious handwritten classifications by processing the thousands of broadcasts on computer and applying ‘cluster analysis’ to isolate similarities in the traffic, which made the classifications infinitely more precise. Within a few years this work has become one of the most important tools in Western counterespionage. Peter Wright, Spycatcher, p. 153.

[45] Kamerman, S. (1999), Eerst de stam, dan de takken en de bladeren: Len de Klerk over het universitair bestel, *NRC Handelsblad, Wetenschap en Onderwijs*, p. 3.

[46] Adviesraad voor het Wetenschaps- en Technologiebeleid (1999). Vitaliteit en kritische massa: Strategie voor de natuur- en technische wetenschappen, p. 31.

[47] From the Press, *Chance, 12*, p. 7.

[48] “It doesn’t lend itself to explanation because it’s a statistical method,” said Chris Fehring, vice president of the Vincinnati office of Trans Union Credit Bureau, one of the three largest credit reporting bureaus in the country. (Des Moines Register, October 11, 1998.)