

V85

TNO-rapport

**CANONISCHE ANALYSE,
MET BEHULP VAN SPLINE FUNKTIES**

TNO Preventie en Gezondheid
Gorter Bibliotheek

02 DEC 2004

Postbus 2215
2301 CE LEIDEN

NIPG-publikatienummer
92.024

April 1992

Alle rechten voorbehouden.
Niets uit deze uitgave mag worden
vermenigvuldigd en/of openbaar gemaakt
door middel van druk, fotokopie, microfilm
of op welke andere wijze dan ook, zonder
voorafgaande toestemming van TNO.

Indien dit rapport in opdracht werd
uitgebracht, wordt voor de rechten en
verplichtingen van opdrachtgever en
opdrachtnemer verwezen naar de
'Algemene Voorwaarden voor Onderzoeks-
opdrachten aan TNO', dan wel de
betreffende terzake tussen partijen
gesloten overeenkomst.

Het ter inzage geven van het TNO-rapport
aan direct belanghebbenden is toegestaan.

© TNO

Stamboeknummer

19.316

H. Vos

Deze uitgave is te bestellen door het overmaken van *f* 27,85 (incl. BTW) op postrekening 99.889 van het NIPG-TNO te Leiden onder vermelding van bestelnummer 92.024

INHOUD

pagina

SAMENVATTING

i

1. INLEIDING

1

1.1 Twee sets variabelen

1

1.2 Het probleemgebied

1

2. DE SCALING-BENADERING VOOR 2 SETS VARIABELEN

4

2.1 CANALS

4

2.2 OVERALS voor k sets variabelen

4

2.3 Verschillen tussen CANALS en OVERALS

5

2.4 Indicatorcodering

5

2.5 Metrische variabelen in CANALS en OVERALS

6

3. DE TRANSFORMATIE-BENADERING

8

3.1 Lineaire funkties

8

3.2 Polynome funkties

9

3.2.1 Eigenschappen van polynome funkties

9

3.3 Spline funkties

10

3.3.1 Knots

10

3.3.2 Continuïteit

11

3.4 Basisfunkties en dimensionaliteit

12

3.5 Lineaire combinaties van basisfunkties

14

3.6 Restricties op de coëfficiënten

14

3.7 Berekening van B-splines

15

3.8 Bepaling van knotposities en graad van de spline

16

3.9 Strategieën voor plaatsing

17

INHOUD (vervolg)	pagina
3.10 Toepassingsmogelijkheden	18
3.10.1 Regressie splines	18
3.10.2 Interpolerende splines	18
3.10.3 Verschillen	19
4. DE SCALING-TRANSFORMATIE BENADERING	20
4.1 Inleiding	20
4.2 De EQSTRA-benadering	20
4.3 Restricties	23
4.3.1 Additiviteitsrestricties	23
4.3.2 Vormrestricties	23
4.4 Lineaire Ongelijkheids Restricties	25
4.5 Gelijkheidsrestricties	27
4.6 Beperkingen van Lineaire Ongelijkheids Restricties	28
4.7 De EQSTRA-procedure	28
5. VOORBEELD MET EMPIRISCHE GEGEVENS	31
5.1 Relatie tussen Hinder en Geurconcentratie voor Industriële bronnen	31
5.1.1 Dosisvariabelen	31
5.1.2 Response variabele	32
5.2 Vraagstellingen	32
5.3 Werkwijze	33
5.4 Covariaten	37
5.5 Positie ten opzichte van BRON	41
5.6 Bespreking van de resultaten	43
5.7 Discussie	45
5.8 Conclusie	46

INHOUD (vervolg)

Pagina

LITERATUUR

47

BIJLAGE : TABELLEN

49

Dit rapport is het resultaat van een afstudeerproject, waarbij de dagelijkse begeleiding in handen was van drs. H.M.E. Miedema, verbonden als onderzoeker aan het Nederlands Instituut voor Praeventieve Gezondheidszorg TNO te Leiden. Ik wil hem bedanken voor zijn motiverende begeleiding en het beschikbaar stellen van gegevens.

Begeleiding vanuit de vakgroep Methoden en Technieken werd verzorgd door dr. A. Mooijaart. Voor zijn advisering wil ik hem eveneens bedanken.

SAMENVATTING

In dit rapport wordt een methode voorgesteld voor het analyseren van twee sets variabelen voor de situatie waarin het meetniveau van de sets gemengd is en in ieder geval metrische variabelen bevat. Centraal staat de *vorm* van de transformaties van *metrische* variabelen.

Voor de analyse van 2 sets zijn methoden beschikbaar, als bijvoorbeeld canonische correlatie analyse (CA) met het algemene lineaire model als uitgangspunt. Variabelen worden behandeld als metrische variabelen. Van alle variabelen in een set worden lineaire combinaties samengesteld, zodanig dat de samenhang tussen de samengestelde variabelen maximaal is. Transformaties van variabelen zijn hier lineair.

In niet-lineaire generalisaties van CA wordt echter hetzelfde aangetroffen: ook hier worden metrische variabelen lineair behandeld. In dit rapport wordt het uitgangspunt gehanteerd, dat een transformatie van een metrische variabele niet noodzakelijk een rechte lijn is, maar er ook anders uit kan zien. Het kan dan in sommige gevallen noodzakelijk zijn het domein van de variabele onder te verdelen in segmenten en vervolgens een basis voor transformatie te construeren. Een manier om dit te realiseren is met behulp van *B-splines*. Wordt dit gedaan voor beide sets, dan kan een criterium geformuleerd worden voor maximale gelijkheid in termen van de splines voor elk van de sets. Conclusies over de relaties tussen de sets worden getrokken door inspectie van de vorm van elke getransformeerde variabele.

1. INLEIDING

1.1 Twee sets variabelen

Bij veel toepassingen van Lineaire Multivariate Analyse (LMVA) is het zinvol een aantal variabelen samen te beschouwen als een set. Onderzocht wordt, hoe een lineaire combinatie van een aantal variabelen samenhangt met een (lineaire combinatie van een aantal) andere variabele(n). Bij multiple regressie analyse bijvoorbeeld wordt een aantal onafhankelijke variabelen gerelateerd aan een afhankelijke variabele. In dit rapport ligt het accent op dit soort meer gecompliceerde modellen, in het bijzonder niet-lineaire modellen, waar sprake is van twee sets variabelen. Onder een set kan worden verstaan: een combinatie van variabelen waarvan interesse bestaat in hun relatie met een andere (combinatie van) variabele(n). De indeling van bepaalde variabelen in een van de sets berust op a priori criteria.

1.2 Het probleemgebied

Stel: er zijn gegevens verzameld met de volgende structuur:

- de gegevens zijn te beschouwen als twee sets;
- de ene set bevat zowel metrische variabelen, betrekking hebbend op een of andere blootstellingsgrootte (*exposure*) als nominale variabelen;
- de andere set bevat ordinale variabelen, betrekking hebbend op de reactie (*response*) van respondenten.

interesse gaat uit naar:

"de vorm van de relatie tussen de twee sets variabelen, rekening houdend met het meetniveau van elke afzonderlijke variabele".

De gegevens zouden bijvoorbeeld verzameld kunnen zijn in het kader van een onderzoek naar geluidsoverlast en zouden de volgende kenmerken kunnen hebben:

SET1 (*exposure set*): Geluidsniveau gemeten in deciBel
 Soort geluidsbron (vliegveld, fabriek, station)

SET2 (*response set*): Gerapporteerde overlast, gemeten op een 5 puntsschaal

Een set bestaat niet noodzakelijkerwijs uit meerdere variabelen. Ook als er sprake is van 1 variabele kan deze als een set opgevat worden. Het hier omschreven probleem heeft kenmerken die specifiek zijn voor twee benaderingen binnen data-analytische opvattingen:

- bij de eerste benadering ligt de nadruk op *optimale schaling* van *niet-metrische variabelen*. Ordinale of nominale variabelen worden geschaald, zodanig dat een of andere verliesfunctie zo klein mogelijk is. De optimale schaling kan worden onderworpen aan restricties met betrekking tot meetniveau, rang en orthogonaliteit van de kwantificatiematrix. Na het vinden van de optimale transformaties worden LMVA technieken toegepast op de getransformeerde variabelen. De Alternating Least Squares (ALS) modellen passen binnen deze benadering.
- bij de tweede benadering ligt het accent op *niet-lineaire transformaties* van *metrische variabelen*. De transformaties hebben een specifieke polynome vorm, of zijn spline functies. De multivariate transformatie van een set variabelen is de gewogen som van de transformaties van de variabelen in deze set. Ook hier kunnen restricties worden opgelegd aan de toegestane transformaties van variabelen: deze kunnen monotoon of lineair zijn, stijgend of dalend.

Beide benaderingen hebben gemeen, dat optimaliteit van schaling resp. transformatie wordt gedefinieerd in termen van maximale samenhang.

De hoofdvraagstelling in dit rapport luidt: Kan er een procedure worden geprogrammeerd, waarin kenmerken uit de twee benaderingen verenigd worden. Zo ja, levert toepassing hiervan zinvolle resultaten op?

De opzet is als volgt:

Een aantal voor de hoofdvraagstelling relevante eigenschappen van de twee genoemde benaderingen worden beschouwd in hoofdstukken 2 en 3. Het accent zal liggen op de tweede benadering. In hoofdstuk 4 wordt een methode voorgesteld, die delen uit beide benaderingen combineert. De geprogrammeerde procedure komt eveneens in hoofdstuk 4 aan de orde. In hoofdstuk 5 tenslotte wordt een voorbeeld gepresenteerd met empirische gegevens.

2. DE SCALING-BENADERING VOOR 2 SETS VARIABELEN

2.1 CANALS (Van der Burg & De Leeuw, 1983)

Deze techniek is ontwikkeld voor het berekenen van de optimale transformaties die de canonische correlatie tussen twee sets variabelen maximaliseert indien de sets variabelen bevatten die van verschillend meetniveau zijn en kan worden opgevat als een generalisatie van Lineaire Canonische Analyse. CANALS kent twee stappen: een stap waarin de canonische correlatie gemaximaliseerd wordt volgens de klassieke Canonische Analyse en een stap waarin de gevonden gewichten optimaal getransformeerd worden, rekening houdend met de per variabele geldende restricties. Naast de klassieke metrische transformatie bestaat nu ook de nominale of ordinale transformatie. In deze stap komt de non-lineariteit tot uitdrukking. De oplossing voor de subsets wordt bij CANALS berekend door verschillende typen regressie van de niet-gerestricteerde oplossingsvector op de oorspronkelijke variabele uit te voeren. Het meetniveau van een variabele (ordinaal of metrisch) bepaalt het type regressie (monotoon of lineair).

2.2 OVERALS voor k sets variabelen (Van der Burg, De Leeuw & Verdegaa, 1988)

Wordt HOMALS (homogeniteitsanalyse voor k variabelen van nominaal meetniveau, Gifi 1990) als uitgangspunt genomen, dan kan OVERALS omschreven worden als een combinatie van HOMALS en optimal scaling met additiviteitsrestricties. Deze restricties zijn van belang om groepen variabelen te definiëren. In het algemeen kan OVERALS beschouwd worden als een generalisatie van een groot aantal lineaire en niet-lineaire technieken: door variatie van opties met betrekking tot meetniveau en groepssamenstelling kunnen zowel lineaire als niet-lineaire problemen in combinatie met optimal scaling met dit programma opgelost worden.

2.3 Verschillen tussen CANALS en OVERALS

Hoewel OVERALS met $k=2$ theoretisch CANALS generaliseert, zijn er toch wat subtiele verschillen: CANALS berekent een oplossing zodanig dat de som van de kwadraten van de canonische correlaties (in p dimensies) zo groot mogelijk wordt. OVERALS met 2 sets echter berekent een oplossing, waarbij de som van de canonische correlaties maximaal is. Voor de 1-dimensionale oplossing maakt dit verschil niets uit, maar bij meerdere dimensies kunnen er verschillen optreden. De CANALS-oplossing wordt derhalve wel equivalent met, maar niet identiek aan, de OVERALS-oplossing genoemd indien: $k=2$ en $p > 1$. Een beperking van CANALS is voorts, dat het programma, evenals bij lineaire CA, rang 1 restricties toepast. Dit heeft tot gevolg, dat de kwantificaties voor de verschillende dimensies proportioneel zijn.

2.4 Indicatorcodering

Een manier om categorische variabelen te representeren is met behulp van indicatormatrices: 1 wordt gescoord indien een kolomcategorie van toepassing is op het rij-object, 0 indien dit niet het geval is. Doordat categorieën elkaar wederzijds uitsluiten kan er in elke rij van indicatormatrix G_j , behorende bij variabele j , slechts éénmaal een 1 voorkomen. Variabelen binnen een set kunnen worden opgevat als subvariabelen van 1 setvariabele met evenveel categorieën als het aantal combinaties van de oorspronkelijke variabelen. In plaats van op de oorspronkelijke variabelen kan de analyse nu uitgevoerd worden op deze interactieve coderingen. Het op deze wijze groeperen van variabelen kan nuttig zijn, indien verondersteld mag worden dat de variabelen interactie vertonen. Het aantal categorieën echter kan onaanvaardbaar hoog worden: voor bijvoorbeeld 4 variabelen met 5 categorieën zijn dit er al 625. Veel cellen zullen echter leeg zijn omdat veel combinaties niet voor zullen komen. Aan dit probleem kan tegemoet gekomen worden door oplegging van *additiviteitsrestricties* aan de variabelen. In variantie-analytische termen betekent dit dat slechts hoofdeffecten onderzocht worden. Indicatormatrices G_j kunnen nu worden gecateneerd tot super-indicatormatrix G .

Tabel 2.1 Scores van 4 personen op 3 variabelen en de bijbehorende super-indicatormatrix G

1	2	3	a	b	c	p	q	r	v	w
a	p	v	1	0	0	1	0	0	1	0
c	q	v	0	0	1	0	1	0	1	0
b	p	w	0	1	0	1	0	0	0	1
e	r	v	0	0	1	0	0	1	1	0

Eén kolom van de afzonderlijke elementen G_j is overbodig: zijn er twee bekend dan is daarmee de laatste gedetermineerd (bij 3 categorieën). Zonder verlies van informatie kunnen deze uit G weggelaten worden.

2.5 Metrische variabelen in CANALS en OVERALS

Gegeven datavector x en vector q van transformaties van x . Toegestane metrische transformaties van x kunnen nu worden geformuleerd als:

$$q = ax + b, \text{ waarbij } a \text{ en } b \text{ constanten.}$$

Ofwel: de vector van de getransformeerde datavector is proportioneel met de oorspronkelijke datavector. De enige reden voor transformatie is, dat deze variabele optimaal wordt met betrekking tot de gespecificeerde verliesfunctie.

Beide analysemethoden zijn expliciet ontworpen voor het analyseren van categorische variabelen. Indien meer informatie gewenst is met betrekking tot metrische variabelen, bijvoorbeeld over de optimale *vorm*, dan moeten naast lineaire transformaties ook niet-lineaire transformaties voor metrische variabelen toegestaan worden. Een interessante mogelijkheid is discretiseren, dat wil zeggen het hercoderen van de metrische variabele in een beperkt aantal klassen. Worden vervolgens de gediscretiseerde variabelen gespecificeerd als nominale of ordinale variabelen in OVERALS, dan zou uit de optimale transformaties mogelijk iets kunnen worden afgeleid over de vorm van de relatie met andere variabelen. Dit is in het

algemeen een vrij ruwe manier van te werk gaan, bovendien blijven de gevonden transformaties beperkt tot stapsgewijze functies.

Een methode die in het geval van een metrische variabele gebruik maakt van meer informatie, namelijk de *lokatie* van een datapunt binnen dat interval, kan worden gevonden in de toepassing van eerste of hogere graads *B-splines*.

3. DE TRANSFORMATIE-BENADERING

3.1 Lineaire functies

Om in een tweedimensionale ruimte met een orthogonaal assenstelsel (X, Y) n puntenparen (x, y) te beschrijven wordt meestal gebruik gemaakt van een lijn die de relatie tussen de assen X en Y uitdrukt. In het algemeen liggen deze punten niet op een rechte lijn. Aangenomen wordt echter, dat de punten toch op een rechte lijn behoren te liggen: afwijking van deze lijn in Y -richting worden opgevat als meetfouten. Anders gezegd: Y is lineair afhankelijk van X . Getracht wordt de puntenparen zo goed mogelijk te benaderen door te zoeken naar coëfficiënten a_0 en a_1 , die $\sum_i (a_0 + a_1 x_i - y_i)^2, i=1, \dots, n$ ofwel $\|ax-y\|$ minimaliseert.

Deze methode is in de approximatie-theorie bekend als de Methode der Kleinste Kwadraten (Least Squares, LS) en wordt ook wel de lineaire regressie van Y op X genoemd. De coëfficiënten a worden gevonden als: $a=(x'x)^{-1}x'y$.

De afstanden in Y -richting (residuen) kunnen door middel van een grafiek zichtbaar gemaakt worden. Indien het verband tussen X en Y een lineair verband is, is er geen duidelijk patroon in de residuen te ontdekken. Heeft het patroon van de residuen bijvoorbeeld een hoefijzerstructuur, dan is er reden te veronderstellen dat het lineaire model de gegevens niet adequaat weergeeft. In sommige gevallen kan door middel van een transformatie van de data lineariteit toch verondersteld worden, waarmee alle aantrekkelijke eigenschappen van het lineaire model behouden kunnen worden. Niet-lineaire relaties kunnen onder meer met behulp van polynomen beschreven worden.

3.2. Polynome functies

Een polynoom is een functie met vergelijking:

$$f(x) = a_0 + a_1x + \dots + a_{N-1}x^{N-1} + a_Nx^N.$$

waarbij: $N \in \mathbb{N}$;

als $a_N \neq 0$ heeft $f(x)$ graad N , of orde $N+1$.

Als $\{x_0, x_1, \dots, x_N\}$ een verzameling van $N+1$ waarden is, dan bestaat er voor elke verzameling van waarden $\{y_0, y_1, \dots, y_N\}$ een unieke polynoom, die puntenparen $\{x_i, y_i\}$, $i=1, 2, \dots, N+1$ verbindt. Deze polynoom wordt de interpolerende polynoom genoemd en is van graad N of lager. Eenvoudig is te zien, dat ook een rechte als een polynoom geschreven kan worden.

De polynoom heeft dan slechts waarden $\neq 0$ voor a_0 en a_1 , een eerste graads interpolerende polynoom. Is het niet strikt noodzakelijk dat de polynoom alle punten verbindt, dan kan de graad van de polynoom sterk worden gereduceerd. Gezocht wordt naar een meer eenvoudige functie van graad $m < N$, die de punten in het XY-vlak zo goed mogelijk benadert. Afwijkingen bestaan uit afstanden in Y-richting. Er dient nu een of andere verliesfunctie gespecificeerd te worden, die vervolgens geminimaliseerd wordt. Ook hier kan de beste benadering in LS zin gebruikt worden. Worden data x als onafhankelijke of eplanatory variabele en y als afhankelijke of response variabele opgevat, dan kunnen de onbekende coëfficiënten a_i geschat worden door middel van (polynome) regressie. Indien regressie wordt opgevat als een transformatie probleem, dan kunnen coëfficiënten a worden opgevat als schattingen voor onbekende transformatie-parameters.

3.2.1 Eigenschappen van polynome functies

Polynomen bezitten twee aantrekkelijke eigenschappen:

1. lineair in de parameters a_i , d.w.z. een lineaire combinatie van de functies $1, x, x^2, \dots, x^N$
2. gemakkelijk te manipuleren in termen van differentiëren, integreren.

Een belangrijke beperking van polynomen echter is hun gebrek aan stabiliteit, in die zin, dat een lokale verandering in het ene gedeelte van het interval een ongewenste verandering in een ander gedeelte van het interval kan bewerkstelligen, zonder dat in dat laatste interval iets is veranderd. Hogere graads polynomen vertonen bovendien een sterk oscillerend gedrag bij de eindpunten van het interval. Deze eigenschappen beperken hun gebruiksmogelijkheden voor statistische doeleinden.

3.3 Spline funkties

Vanwege de nadelen van polynomen kan het zinvol zijn het domein van x onder te verdelen in subintervallen. Voor elk van deze intervallen wordt de funktie als een polynoom van een bepaalde graad gedefinieerd, waarbij de graad van de polynoom in elk interval dezelfde is. Het voordeel is nu, dat deze polynomen van lage graad kunnen zijn. De Spline funktie bestaat nu uit een aaneenschakeling van de polynome segmenten van gelijke graad, die geforceerd worden op de intervalgrenzen zo vloeiend mogelijk op elkaar aan te sluiten. Deze Spline funkties kunnen verschillende eigenschappen hebben, die tot uitdrukking komen in de benaming in de literatuur: M -splines werden voor het eerst gedefinieerd door Curry en Schoenberg (1947). Een uitgebreide behandeling van de nauw verwante B -splines komt van de Boor (1978). I -splines werden gedefinieerd door Ramsay (1982). De term Spline vindt zijn oorsprong in een tekeninstrument, gebruikt om een vloeiende curve door een aantal punten te trekken. De herkomst van de verschillende benamingen voor de splines is minder duidelijk: alleen Ramsay vermeldt dat I -spline staat voor 'Integrated Spline'.

3.3.1 Knots

De eerste stap bij bepaling van de spline is het definiëren van een aantal subintervallen in domein $[L,U]$ van de gegevens. Hun aantal en breedte worden bepaald door specificatie van een reeks knooppunten, of knotsequentie:

$$L = t_0 < t_1, \dots, t_{q-1} < t_q = U$$

De punten $L=t_0$ en $U=t_q$ begrenzen het domein. Deze punten worden *exterior knots*, t_1, \dots, t_{q-1} de *interior knots* genoemd.

In elk subinterval $[t_j, t_{j+1}]$, $j=0, \dots, q-1$ wordt een polynoom P_j van orde k ofwel graad $k-1$ gedefinieerd. De gezochte functie bestaat uit een aaneenschakeling van polynomen P_j op het interval $[L, U]$. Indien deze polynomen op de intervalgrenzen vloeiend op elkaar aansluiten betekent dit, dat behalve de functiewaarden ook de afgeleiden tot een bepaalde graad aan elkaar gelijk moeten zijn. Bij *maximale vloeiendheid* geldt voor de afgeleiden :

$$(D^{m-1}P_j)(t_{j+1}) = (D^{m-1}P_{j+1})(t_{j+1}), \quad m=1, \dots, k-1; \quad j=1, \dots, q-1$$

waarbij : $(D^i P_j)(t_k)$ de i^e afgeleide van Polynoom j in knot k .

3.3.2 Continuïteit

Continuïteitseigenschappen zijn van belang voor het vloeiend in elkaar overlopen van de spline van het ene naar het andere interval. De eis dat de Spline continu moet zijn is een noodzakelijke, maar niet voldoende voorwaarde voor een vloeiende overgang. Hiervoor geldt als extra eis, dat de afgeleiden van beide segmenten in het punt van overgang van een interval i naar interval $i+1$ aan elkaar gelijk moeten zijn en continu. Gelijkheid van de eerste afgeleide is een voorwaarde voor *visuele vloeiendheid* en is dus alleen voor tweede graads splines een voldoende voorwaarde voor maximale vloeiendheid. Voor derde en hogere graads splines geldt, dat er respectievelijk een tweede en hogere afgeleide kan bestaan, waardoor deze functies vloeiender kunnen zijn dan nulde of eerste graads Splines. Omdat polynomen tot de klasse oneindig differentieerbare functies behoren, zijn de continuïteitseisen alleen van belang in de knots. In elkaar overlappende polynomen hebben in de knots afgeleiden tot maximaal orde $k-2$. Zijn er bijvoorbeeld geen knots, dan zijn de functie en $k-1$ afgeleiden continu. Als er 1 knot is, dan zijn de functie en maximaal $k-2$ afgeleiden continu. Zijn er k knots

op een punt, dan is de functie discontinu. Bij polynomen van lage orde, dus bijvoorbeeld als $k=1$ bestaat de spline uit een discontinuë stapfunctie. Als $k=2$ bestaat de spline uit rechte lijnen, die op elkaar aansluiten in de knots (piecewise linears). De afgeleide is in dit punt niet gedefinieerd, dus de overgang van in de knots verloopt niet noodzakelijk visueel vloeiend.

3.4 Basisfuncties en dimensionaliteit

De constructie van een uit segmenten bestaande polynoom wordt gerealiseerd door definitie van een set basisfuncties $M_i, i=1, \dots, n$ zodanig dat de spline geschreven kan worden als een gewogen som $\sum a_i M_i$ van deze basisfuncties. Curry en Schoenberg (1947) definieerden hun M -splines zodanig, dat M_i positief is in interval (t_i, t_{i+k}) en 0 elders. Verder is M_i zodanig genormaliseerd, dat het oppervlak onder $M_i=1$. Het minimum aantal basisfuncties benodigd voor unieke bepaling van een spline wordt de *dimensie* van de basis genoemd.

Aangetoond kan worden, dat de dimensie van een minimum basis afhangt van de gekozen graad en het aantal knots voor splinefunctie S volgens:

$$\text{Dim}(S) = q-1+k$$

ofwel het aantal interior knots + de orde van de spline.

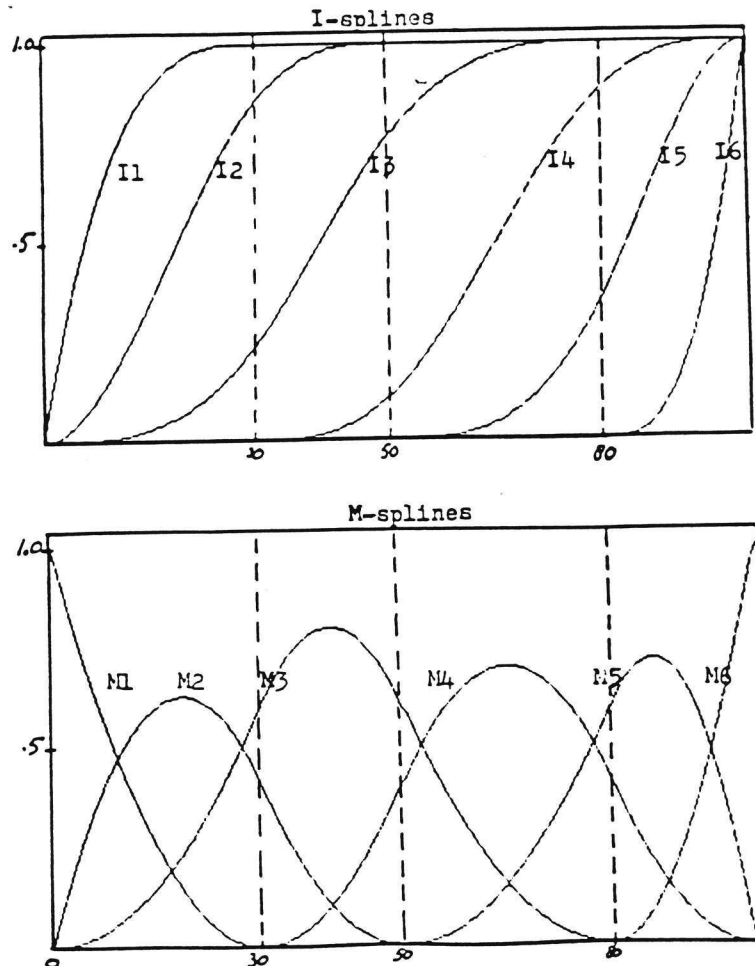
Door per M_i te integreren worden de geïntegreerde, ofwel I -splines (Ramsay, 1982) geconstrueerd;

$$I_i(x) = \int_L^x M_i(u) du$$

Deze functies zijn opnieuw basisfuncties, nu echter van *graad* k . Omdat M_i nergens negatief is, zijn dit strikt stijgende functies. Gecombineerd met niet-negatieve waarden voor de coëfficiënten a_i wordt $\sum a_i I_i$ een monotoon stijgende funk-

tie. Wordt bijvoorbeeld de laatste coëfficiënt op 0 gesteld, dan is de funktiewaarde in t_q gelijk aan 1. Andersoortige restricties op de a_i 's zijn eveneens mogelijk, evenals verdere restricties op de afgeleiden in de knots. Laatstgenoemde vallen echter buiten het bestek van dit rapport. Nadeel van I -splines is, dat niet precies bekend is waarvan ze de basis zijn.

Figuur 3.1: Basis I -splines en basis M splines



B -splines zijn transformaties van M -splines, in de zin dat de normalisatie zodanig is, dat $\sum_j b_{ij} = 1, i=1, \dots, n; j=1, \dots, q-1+k$. Dit is vooral aantrekkelijk voor het modelleren van categorische data: immers elke b_{ij} kan worden opgevat als de kans

dat datapunt i in categorie j valt. Vanwege deze eigenschap worden verder alleen B -splines behandeld: een gewogen som van basisfuncties wordt verder aangeduid met B .

3.5 Lineaire combinaties van basisfuncties

Elke B_i is zelf een spline en dus een uit delen opgebouwde polynoom. De aantrekkelijke eigenschappen van deze polynomen zoals lineariteit in de coëfficiënten en eenvoudige manipuleerbaarheid, zijn ook van toepassing op een gewogen som van deze basisfuncties. De som van twee functies met graad N levert een nieuwe functie, ook weer van graad N . Met een basis van lineaire splines zal B dus lineair zijn in alle intervallen en is verbonden in de knots. Zijn de basisfuncties kubisch, dan is B kubisch in elk interval en is maximaal twee keer differentieerbaar in de knots.

3.6 Restricties op de coëfficiënten

Omdat B_i zodanig is gedefinieerd, dat deze positief is in het interval $[t_i, t_{i+k}]$ en 0 elders zal de coëfficiënt a_i B alleen beïnvloeden in dit interval en kan op deze manier dus eenvoudig een lokale invloed uitoefenen.

Aan B kunnen verschillende restricties worden opgelegd door middel van restricties op de coëfficiënten a_i van de basisfuncties. Beperkingen zijn mogelijk, zodanig dat B een bepaalde karakteristieke vorm krijgt. Worden bijvoorbeeld de eerste coëfficiënt van een lineaire B -spline op 0 en de tweede op een waarde > 0 geresliceerd, dan zal B bij $(t_0, 0)$ beginnen en in het eerste interval stijgen. In § 4.3 wordt uitgebreider op deze problematiek ingegaan.

3.7 Berekening van B-splines

B-splines van orde $k=1$ worden gedefinieerd als:

$$B_1(x|k=1,t) = \begin{cases} 1 & , t_i \leq x < t_{i+1} \\ 0 & , \text{ elders} \end{cases}$$

B-splines van orde k zijn recursief gerelateerd aan die van orde $k-1$ volgens: (de Boor 1978)

$$B_i(x|k,t) = \left[\frac{x-t_i}{t_{i+k-1}-t_i} B_{i,k-1}(x) + \frac{t_{i+k}-x}{t_{i+k}-t_{i+1}} B_{i+1,k-1}(x) \right] \quad , k > 1$$

Voor unieke bepaling van de functie moeten er naast de $q-1$ interior knots $2k$ extra knots worden gespecificeerd, die niet in het interval liggen. Dit mogen alle punten zijn die voldoen aan:

$$t_{-(k-1)} \leq \dots \leq t_{-1} \leq t_0 = L \quad \text{en} \quad U = t_q \leq t_{q+1} \leq \dots \leq t_{q+k-1} \quad , k > 1$$

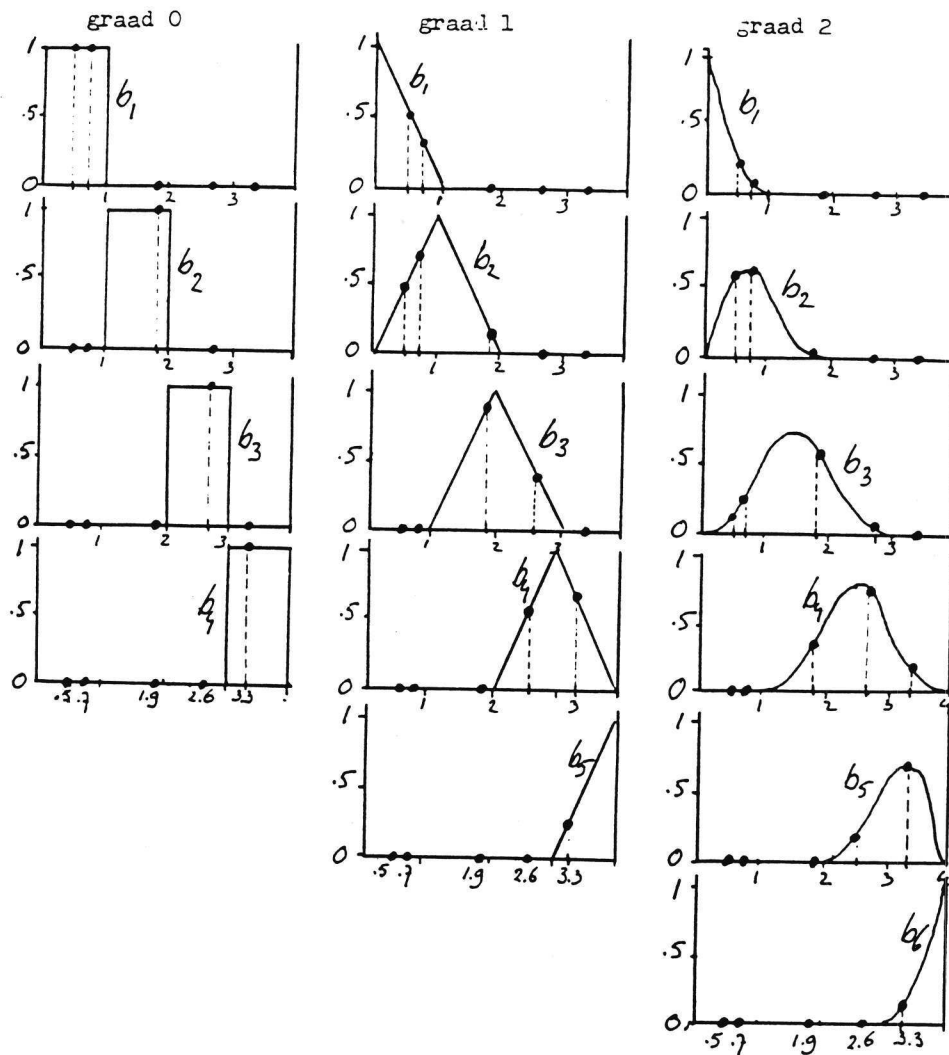
In de praktijk is het gebruikelijk deze knots zo te kiezen, dat $t_{-(k-1)}, \dots, t_0 = L$ en $t_q, \dots, t_{q+k-1} = U$. Op deze manier wordt toegelaten, dat de functie aan de eindpunten van interval $[L, U]$ discontinu is.

Getallenvoorbeeld

Tabel 3.1 $0^e, 1^e$ en 2^e graads basis B-splines, geëvalueerd voor 5 waarden x , op knotsequentie 1,2,3.

	graad van de Spline														
	0				1					2					
	b1	b2	b3	b4	b1	b2	b3	b4	b5	b1	b2	b3	b4	b5	b6
.5	1	0	0	0	.5	.5	0	0	0	.31	.60	.09	0	0	0
.7	1	0	0	0	.3	.7	0	0	0	.11	.68	.21	0	0	0
1.9	0	1	0	0	0	.1	.9	0	0	0	.01	.59	.40	0	0
2.6	0	0	1	0	0	0	.4	.6	0	0	0	.08	.73	.19	0
3.3	0	0	0	1	0	0	0	.7	.3	0	0	0	.21	.68	.11

Figuur 3.2: $0^e, 1^e$ en 2^e graads basis B-splines, geëvalueerd voor de 5 datapunten uit tabel 3.1



3.8 Bepaling van knotposities en graad van de spline

De keuze van aantal en plaats van de interior knots kan zowel a priori als a posteriori gemaakt worden en is van belang, omdat deze keuze in grote mate de eigenschappen van de Spline functie bepaalt. Overwegingen die een rol kunnen spelen in de keuze zijn:

- a. hoe meer knots in een bepaald gebied, hoe beter de datapunten in dat gebied benaderd kunnen worden;
- b. hoe meer datapunten tussen twee knots, hoe beter de 'ware' functie in dat interval geschat kan worden.

Laatstgenoemde overweging is van meer belang dan de eerste, immers het heeft weinig zin om veel knots te plaatsen in een gebied waar een gering aantal datapunten een slechte benadering door B zou impliceren. Op grond van statistische overwegingen wordt een minimum van 4 à 5 datapunten tussen de knots wenselijk geacht (Winsberg en Ramsay, 1981). In het algemeen verdient het aanbeveling zowel de graad als het aantal knots zo laag mogelijk te houden. Deze bepalen het aantal parameters in het model: het doel van data analyse is immers de verzamelde gegevens zo goed mogelijk te beschrijven met zo weinig mogelijk parameters.

3.9 Strategieën voor plaatsing

Gegeven bovenstaande overwegingen kunnen een viertal verschillende strategieën met betrekking tot plaatsing worden onderscheiden :

1. *Vrije plaatsing*

Op grond van eerdere analyses of op basis van a priori veronderstellingen omtrent de vorm van het verband en/of de gewenste mate van detaillering worden knots geplaatst. Dit kan lokaal verschillen. Hierbij geldt, dat door vergroting van een interval oneffenheden en afwijkingen van monotoniciteit gereduceerd kunnen worden. Is lokaal veel detaillering gewenst, dan kunnen lokaal meerdere knots dicht bij elkaar worden geplaatst. Indien de nieuwe resultaten hiertoe aanleiding geven kunnen achteraf nog knots verwijderd of bijgeplaatst worden.

2. *Gelijke afstand van elkaar (equidistant)*

Met het bepalen van het aantal knots ligt nu tevens de plaats vast. Deze manier komt in aanmerking indien er geen reden tot veronderstelling van lokale eigenaardigheden is. Bij variabelen van nominaal en ordinaal meetniveau

kan deze methode gebruikt worden, indien categorienummers opvolgend zijn.

3. *Evenveel datapunten tussen de knots (n-tiles)*

De knotposities worden nu bepaald door de verdeling van de datapunten over het interval. Alleen het aantal dient gespecificeerd te worden. Een zinvolle eerste plaatsing is bijvoorbeeld een enkele knot ter hoogte van de mediaan, daarna op de kwartielen etc. Deze methode garandeert een gelijk aantal observaties binnen een interval, hetgeen betekent dat in ieder interval de functie even goed benaderd kan worden.

4. *Optimalisatie van aantal en plaats*

Hiervoor zijn technieken ontwikkeld, maar de bruikbaarheid hiervan is nog niet aangetoond.

3.10 Toepassingsmogelijkheden

3.10.1 Regressie Splines

B-splines worden soms onderscheiden naar toepassing. Is het doel het benaderen door middel van *B* van een functie *Y*, bijvoorbeeld de waarden die de data volgens een bepaald gespecificeerd model moeten aannemen, dan worden deze functies ook wel regressie splines genoemd (Hastie & Tibshirani,1990; Ramsay,1988). De regressiecoëfficiënten kunnen dan worden beschouwd als schattingen voor transformatieparameters van de onafhankelijke variabele *X*, zodanig dat *B* de functiewaarden *Y* zo goed mogelijk benadert in LS zin.

3.10.2 Interpolerende Splines

Andere toepassingen zijn de zgn. interpolatie- of smoothing splines, waarbij het doel bestaat uit het fitten van een lijn in de data, die hier bestaan uit bivariate observaties (x_i, y_i) (Hastie & Tibshirani,1990; Helzer,1983; Lancaster,1986;

Ramsay, 1988). Een dergelijke functie wordt om deze reden ook wel een scatterplot-smoother genoemd . Het uitgangspunt hierbij is, dat er een relatie tussen de assen bestaat. De verliesfunctie is echter een andere dan bij regressie splines : deze wordt gecontroleerd door één enkele parameter, waarmee de verhouding tussen verlies en gladheid naar believen kan worden ingesteld.

3.10.3 Verschillen

Het verschil tussen regressie en interpolerende of smoothing splines komt het duidelijkst tot uiting in de plaatsing van de knots. Bij regressie splines worden de knots op a priori gekozen waarden geplaatst en zijn in het algemeen minder talrijk dan bij interpolatie splines. Regressie splines lijken geschikter voor statistische toepassingen omdat ze in doorgaans minder parameters hebben.

Splines vinden steeds meer toepassingen, zowel binnen de sociale als de exacte wetenschappen. Hun grote mate van flexibiliteit is hiervoor de belangrijkste reden. Soms hebben gegevens in sociaalwetenschappelijk onderzoek een grillig karakter, of geeft een inspectie van hun verdeling aanleiding tot veronderstelling van niet-lineariteit. Vooral indien deze niet-lineariteit een sterk *lokaal* karakter heeft lijkt er een belangrijke rol voor splines te zijn weggelegd.

Ramsay (1988) geeft een aantal voorbeelden van de succesvolle toepassing van (monotone) spline functies voor de volgende analysetechnieken:

- Monotone Spline Regressie
- Monotone Spline Principale Componenten Analyse
- Monotone Spline Canonische Correlatie Analyse
- Item Response modellen met Spline functies
- Multidimensional Scaling

In het volgende hoofdstuk zal een methode worden geïntroduceerd, die gebaseerd is op Monotone Spline Canonische Correlatie Analyse.

4. DE SCALING-TRANSFORMATIE BENADERING

4.1 Inleiding

In voorgaand hoofdstuk werd behandeld, hoe met behulp van *B*-splines een basis voor *transformatie* van een metrische variabele kon worden geconstrueerd. Was het doel het benaderen van een modelvector *Y*, dan werd *B* de regressie spline genoemd. Dat een dergelijke modelvector zou kunnen bestaan uit een *optimale schaling* van een set categorische variabelen kwam in hoofdstuk 2 aan de orde.

In dit hoofdstuk wordt een combinatie van beide methoden voorgesteld, waarbij de gebruikelijke lineaire en monotone restricties worden uitgebreid met *concave* en *convexe* restricties voor metrische variabelen (§ 4.3.2). In § 4.2 zal het model, waarin beide benaderingen worden gecombineerd, worden geïntroduceerd.

In § 4.4 zal nader op gerestricteerde optimalisatie worden ingegaan. De geprogrammeerde procedure wordt in § 4.7 behandeld. In hoofdstuk 5 tenslotte zal een praktische toepassing worden gepresenteerd.

4.2 De EQSTRA-benadering

Bij de hier voorgestelde aanpak is de samenhang tussen exposure set en response set het optimaliteitscriterium. Nominale en ordinale variabelen worden geschaald volgens de Optimal Scaling-Benadering (Gifi,1981), metrische variabelen worden getransformeerd volgens de Transformatie Benadering (Hastie & Tibshirani,1990; Lancaster,1986; Ramsay,1988), zodanig dat hun samenhang of gelijkheid maximaal is. Een simultane schaling en transformatie wordt voorgesteld. Het probleem kan geformuleerd worden als: het vinden van een Transformatie, die omschreven kan worden als een combinatie van optimal scaling en transformatie volgens:

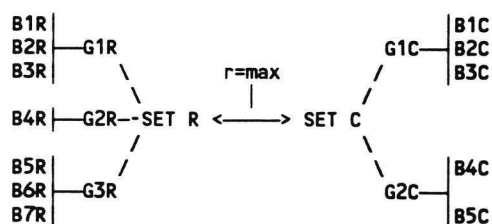
$$\text{Transformatie (X)} = \text{transformatie [schaling(X)]},$$

waarbij X een geobserveerde variabele. De sets echter bestaan uit variabelen van gemengd meetniveau, de exposure set bevat metrische en nominale variabelen, de response set ordinale variabelen. Is X metrisch, dan is het schalings-gedeelte van de Transformatie triviaal, is X nominaal dan kan het transformatie-gedeelte triviaal zijn. Het uiteindelijke doel is het afleiden van de vorm van de relaties tussen de twee sets middels het vinden van Transformaties.

Deze benadering kenmerkt zich door de volgende eigenschappen:

1. De procedure bestaat uit twee stappen: in de eerste stap wordt de schaling samen met de transformatie geoptimaliseerd. In de tweede stap worden deze gescheiden. Deze methode werd eerder gebruikt door Kruskal & Shepard (1974) bij hun component analyse voor ordinale data. Ook Gifi (1990) vermeldt zo'n 2 stappen benadering voor metrische gegevens: vergelijking van optimale met a priori kwantificaties van een in klassen verdeelde metrische variabele kan uitwijzen hoe a priori kwantificaties moeten worden getransformeerd om optimale transformaties te vinden. Gifi noemt dit 'HOMALS als eerste stap'. In § 2.5 werd een vergelijkbaar gebruik van OVERALS overwogen.
2. De optimalisatie van Transformaties van sets variabelen met betrekking tot additieve combinaties van die variabelen geschiedt met behulp van B -splines. Elke afzonderlijke variabele wordt gerepresenteerd door een splinematrice, waarna deze worden geconcateneerd. De additiviteits-restrictie kan (gedeeltelijk) vervallen, indien gebruik gemaakt wordt van productfuncties (interactieve variabelen).
3. A priori kennis komt tot uitdrukking in de keuze van de splines en de keuze van de vorm, waarop deze eventueel gerestricteerd worden. Met behulp van B -splines kan de afwijking van Transformaties van monotoniteit worden weggewerkt. Aanpassing kan worden verkregen door het veranderen van aantal of plaats van de knots.
4. Het criterium voor selectie van Transformaties is maximale gelijkheid van getransformeerde sets van variabelen.

Figuur 4.1 Model EQSTRA (Equalization by Scaling and Transformation)
 voorbeeld met $I_r=7$ variabelen in $J_r=3$ groepen in set R
 en: $I_c=5$ variabelen in $J_c=2$ groepen in set C.



(Combinaties van) splines B_{i_r} en B_{i_c} ($i_r=1, \dots, I_r, i_c=1, \dots, I_c$) vormen groepen G_{j_r} en G_{j_c} ($j_r=1, \dots, J_r, j_c=1, \dots, J_c$) in resp. SET R en SET C. Groepen variabelen worden door de gebruiker gespecificeerd. De objectscores worden per R-groep genormaliseerd op $1/J_r$, per C-groep op $1/J_c$. Dit impliceert, dat de bijdrage van de groepen in elke set gelijk zijn. Coëfficiënten voor de splinegroepen worden geschat, rekening houdend met het meetniveau van de variabelen, zodanig dat de sets maximaal correleren.

Het EQSTRA probleem luidt:

maximaliseer $\mathbf{pR}'\mathbf{Cq}$

onder beperkende voorwaarden: $\mathbf{r}_j'\mathbf{r}_j=1/J_r, j=1, \dots, J_r$

$\mathbf{c}_j'\mathbf{c}_j=1/J_c, j=1, \dots, J_c$

$\mathbf{e}'\mathbf{R} = \mathbf{e}'\mathbf{C} = 0$

en: er kunnen restricties zijn voor sommige variabelen. (1)

waarbij: $\mathbf{R} : J_r$ groepen B -splines voor de R-set (2)

$\mathbf{C} : J_c$ groepen B -splines voor de C-set (2)

$\mathbf{p} : \text{een gewichtenvector voor R}$ (3)

$\mathbf{q} : \text{een gewichtenvector voor C}$ (3)

$\mathbf{r} : \text{objectvector } \mathbf{Rp} / \|\mathbf{Rp}\|$

$\mathbf{c} : \text{objectvector } \mathbf{Cq} / \|\mathbf{Cq}\|$

$\mathbf{e} : \text{een n-vector met enen.}$

Probleem (1) wordt besproken in § 4.4. De wijze waarop de groepen samengesteld worden (2) komt aan de orde in § 4.3. Probleem (3), het vinden van de gewichtenvectoren komt in § 4.4 aan de orde.

4.3 Restricties

4.3.1 Additiviteitsrestricties

Elke variabele wordt gerepresenteerd door een B -spline. Vervolgens worden groepen samengesteld. Een manier om groepen (discrete) variabelen samen te stellen is door specificatie van de super-indicatrix. Dit kwam in § 2.4 aan de orde. Spline matrices kunnen worden opgevat als generalisaties van indicator-matrices : waren bij indicator-matrices slechts 0 en 1 toegestaan, bij spline matrices kunnen dit ook waarden tussen 0 en 1 zijn. Indien de spline matrices op eenzelfde manier behandeld worden, dus door middel van concatenatie, dan betekent dat ook hier oplegging van additiviteitsrestricties: alleen de hoofdeffecten worden beschouwd. Bestaat er interesse in interacties tussen variabelen, dan kunnen productsplines in het model opgenomen worden. Hierbij worden alle mogelijke combinaties van categorieën van twee variabelen in een product uitgedrukt. Voor deze klasse van variabelen wordt een ongerestricteerde oplossing gezocht. Een uitgebreide behandeling van het additieve model kan gevonden worden in Hastie en Tibshirani (1990).

4.3.2 Vormrestricties

Monotone Restricties

Variabelen worden getransformeerd omwille van interpretatieve argumenten. Stel er is a priori informatie met betrekking tot een bepaalde variabele beschikbaar : deze is op nominaal niveau gemeten. Categorienummers zijn in dit geval slechts toegekend om de categorieën van elkaar te kunnen onderscheiden. Zonder verlies

van informatie kan elk ander getal aan de categorie toegekend worden, zolang de categorieën maar gescheiden blijven. Dit wordt Scaling genoemd. Wordt bovendien een verliesfunctie L gespecificeerd en worden de categorieën geschaald (volgens bepaalde regels) zodanig dat L minimaal is, dan is dit Optimal Scaling. Is bekend dat een variabele op ordinaal niveau gemeten is, of zijn er redenen om een variabele als zodanig behandelen, dan zijn slechts schalingen toegestaan die de ordening op eenzelfde manier representeren zoals die in de oorspronkelijke variabele aangetroffen worden. Worden datavector h en vector q , de transformaties van h , beschouwd dan moet gelden :

$$h_i < h_j \rightarrow q_i \leq q_j$$

Een dergelijke transformatie is een monotone transformatie, de hierbij horende functie een monotone functie.

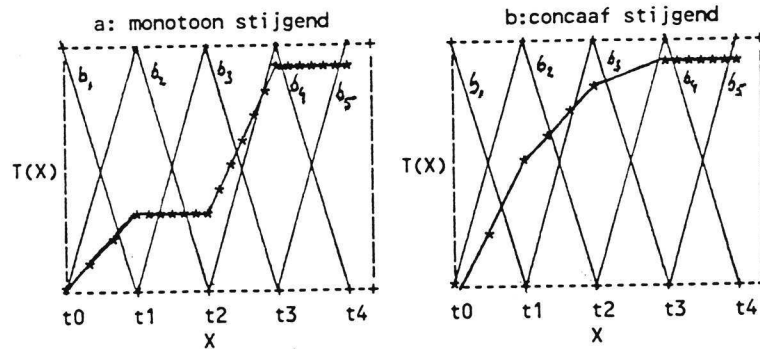
Concave en Convexe Restricties

Bijzondere gevallen van een monotoon stijgende transformaties zijn concaaf en convex stijgende transformaties. Een convexe (*concaaf*) functie kan worden omschreven als: $f(x)$ is convex (*concaaf*) indien voor elk puntenpaar (a,b) op de grafiek van $f(x)$ geldt, dat de koorde die deze twee punten verbindt nergens onder (*boven*) de grafiek van $f(x)$ ligt.

Een bekende concaaf stijgende functie is de wortelfunctie, een bekende convex stijgende functie is de exponentiële functie. Convexe en concaaf restricties kunnen zinvol zijn voor metrische variabelen, monotone restricties kunnen zinvol zijn voor zowel metrische als ordinale variabelen.

Deze hierboven beschreven transformaties kunnen worden afgedwongen middels restricties op de *verschillen* tussen B -spline coëfficiënten. Deze verschillen kunnen worden geformuleerd in termen van ongelijkheden. Worden transformaties $T(X)$ uitgezet als functie van X met knotsequentie t dan zouden volgende grafieken kunnen ontstaan:

Figuur 4.2



4.4 Lineaire Ongelijkheids Restricties

De voorwaarden:

$$\{ a_5 \geq \dots \geq a_2 \geq a_1 \}$$

zijn, indien B van graad 1 is, voldoende voorwaarden om B monotoon te laten stijgen. Immers, elke $a_i B_i + a_{i+1} B_{i+1}$ mag in een bepaald interval geen lagere waarde hebben dan in het voorgaande. De restricties gelden voor het verschil tussen maximaal twee opeenvolgende coëfficiënten. Moet B behalve monotoon stijgend ook nog concaaf zijn, dan zijn de eisen:

$$\begin{cases} a_2 & \geq a_1 \\ a_{i+2} - a_{i+1} & \leq a_{i+1} - a_i \\ a_n & \geq a_{n-1} \end{cases}, i=1, \dots, n-2 \quad (4.1)$$

voldoende voorwaarden om de functie tot concaaf stijgend te dwingen. De voorwaarden gelden nu voor verschillen van maximaal drie opeenvolgende coëfficiënten in twee opvolgende paren. Schattingen voor gewichten a_i worden gezocht, die gegeven de restricties zo goed mogelijk een andere functie benaderen, waarbij 'zo goed mogelijk' een zo klein mogelijk LS verlies betekent.

Een dergelijk probleem, waarbij de te minimaliseren functie een kwadratische functie (LS) is en alle restricties lineair zijn, wordt in de optimalisatie literatuur

aangeduid als een *kwadratisch* programmeerprobleem. Een overzicht van deze problematiek wordt gegeven door ondermeer Fletcher (1981).

Hier zal nader worden beschreven hoe dit probleem geformuleerd kan worden in termen van restricties en doelfuncties. Dit wordt op een intuïtieve manier gedaan: voor een meer uitgebreide en diepgaande behandeling van de optimalisatieproblematiek wordt verwezen naar meer wiskundig georiënteerde literatuur.

Ter illustratie wordt de constraint set voor het concaaf stijgend gerestricteerde lineaire splines probleem uit figuur 4.2.b uitgeschreven. Er zijn 4 intervallen, dus het aantal lineaire basisfuncties bedraagt 5. Gezocht worden 5 gewichten a_1, \dots, a_5 voor de basisfuncties, zodanig dat B in overeenstemming is met de eisen met betrekking tot de concave vorm.

De ongelijkheden uit (4.1) kunnen voor 5 coëfficiënten worden herschreven als:

interval	coëfficiënten a_i	≥	0
1	$-a_1 + a_2$	≥	0
1,2	$a_1 - 2a_2 + a_3$	≥	0
2,3	$a_2 - 2a_3 + a_4$	≥	0
3,4	$a_3 - 2a_4 + a_5$	≥	0
4	$a_4 + a_5$	≥	0

(4.2)

G, h worden:

interval	G	h
1	$-1 \ 1 \ 0 \ 0 \ 0$	0
1,2	$1 \ -2 \ 1 \ 0 \ 0$	0
2,3	$0 \ 1 \ -2 \ 1 \ 0$	≥ 0
3,4	$0 \ 0 \ 1 \ -2 \ 1$	0
4	$0 \ 0 \ 0 \ 1 \ -1$	0

(4.3)

Het probleem van het vinden van de coëfficiënten a kan nu geformuleerd worden als het volgende (gerestricteerde) LS minimalisatie probleem:

Gegeven spline matrix B en modelvector f :

$$\begin{array}{ll}
 \text{Minimaliseer} & \|Ba - f\| \\
 \text{gegeven restricties} & Ga \geq h
 \end{array}
 \tag{4.4}$$

Voor de oplossing van dit probleem is gebruik gemaakt van een algoritme, ontwikkeld door Lawson en Hanson (1974). Zij beschrijven, hoe probleem (4.4) kan worden getransformeerd in een equivalent afstandsprobleem, dat vervolgens met behulp van niet-negativiteitseisen wordt opgelost. Bij laatstgenoemde procedure worden twee sets indices onderscheiden: een set voor de parameters waarvan de restricties geschonden zijn en een complementaire set voor de ongeschonden restricties of ongerestricteerde parameters. Tijdens loop van het algoritme worden de sets telkens aangepast. Soortgelijke algoritmen worden in Fletcher (1982) beschreven als de *active set*-methode. Thisted (1988) noemt deze methode de *manual* methode en merkt tevens op dat deze algoritmen niet noodzakelijk een optimale oplossing geven. Welk effect dit zou kunnen hebben op de maximale correlatiecoëfficiënt is niet onderzocht. Voor een meer mathematische beschrijving van deze methode wordt naar Lawson en Hanson verwezen.

4.5 Gelijkheidsrestricties

Behalve vormrestricties kunnen ook andersoortig lineair gerestricteerde LS problemen opgelost worden door specificatie van de gewenste parametercoëfficiënten in termen van (G, h) .

Zo is het middels ongelijkheidsrestricties eveneens mogelijk gelijkheidsrestricties af te dwingen. Wordt bijvoorbeeld de volgende constraint set beschouwd:

$$\begin{array}{ccccccc}
 & & G & & h & & \\
 1 & 0 & 0 & 0 & \dots & 0 & c & a_1 \geq c \\
 -1 & 0 & 0 & 0 & \dots & 0 & -c & c \geq a_1 \\
 \cdot & \cdot & \cdot & \cdot & \dots & \cdot & \geq & \cdot \geq \cdot \\
 0 & 1 & -1 & 0 & \dots & 0 & 0 & a_2 \geq a_3 \\
 0 & -1 & 1 & 0 & \dots & 0 & 0 & a_3 \geq a_2
 \end{array} \tag{4.5}$$

Hiervoor zal slechts een oplossing gevonden kunnen worden, indien $a_1=c$ en $a_2=a_3$. Voor fixatie op constanten zou gedacht kunnen worden aan gelijkstelling van de toegestane transformatie aan een andere transformatie.

4.6 Beperkingen van Lineaire Ongelijkheids Restricties

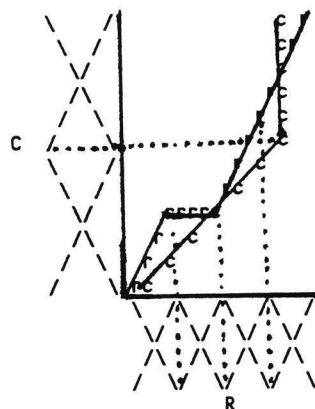
Benadrukt wordt, dat de hierboven gepresenteerde constraint sets (4.3) en (4.5) slechts gelden voor *eerste graads* splines op intervallen van *gelijke lengte*. Voor tweede en hogere graads splines is de restrictiematrix niet meer in termen van lineaire ongelijkheids restricties te specificeren: de constraints worden niet-lineair in hun coëfficiënten. Hiervoor zijn wel oplossingen te vinden, die echter buiten het bestek van dit rapport vallen. Voor intervallen van ongelijke lengte wordt de constraint set geschaald door elke constraint functie te wegen met de lengte van het interval waarin de constraint functie actief is.

4.7 De EQSTRA-Procedure

EQSTRA veronderstelt twee sets variabelen, SET R (Row-set) en SET C (Column-set). Ook indien er per set sprake is van slechts één variabele is er sprake van een set. De sets bestaan uit groepen en de groepen bestaan uit (combinaties van) variabelen. Per variabele kan worden vastgelegd welke transformaties zijn toegestaan. De standaard keuzemogelijkheden hierbij zijn:

- *ongerestricteerd*
- *monotoon*
- *convex*
- *concaaf*
- *lineair*

Figuur 4.3 1 variabele per groep, 1 groep per set, lineaire splines:
R-set: 3 interior knots, monotoon; C-set: 1 interior knot, concaaf.



Sets R en C worden maximaal aan elkaar gerelateerd door het optimaliseren van de correlatie tussen de lineaire combinaties van de b_i 's, rekening houdend met de restricties die per variabele gelden. In de bijzondere situatie van 1 groep per set kan in het ongerestricteerde geval het probleem geformuleerd worden als een eigenwaarde probleem en kan de oplossing analytisch gevonden worden. In elke andere situatie dient de oplossing iteratief berekend te worden.

Een iteratie bestaat uit 2 gedeelten: in het eerste gedeelte worden de (eventueel gerestricteerde) coëfficiënten van R bepaald die zo goed mogelijk de op dat moment bekende genormaliseerde lineaire combinaties van C benaderen, in het tweede gedeelte is dit juist andersom.

De EQSTRA-procedure onderscheidt de volgende stappen:

voorbereiding

1. Kies voor variabele i de knotsequentie en graad : bereken B_i ;
standaard keuzemogelijkheden voor knots zijn: *n-tiles* , *equidistant* of *vrij*.
splines voor nominale en ordinale variabelen krijgen graad 0, splines voor metrische variabelen zijn tenminste van graad 1 (§ 3.9) ;
2. Kies per variabele de gewenste vormrestrictie ;

standaard keuzemogelijkheden zijn: *ongerestricteerd*, *monotoon*, *convex*, *concaaf* of *lineair*; splines voor nominale variabelen, productsplines en splines voor metrische variabelen met graad hoger dan 1 worden ongerestricteerd gefit (§ 4.3.2) ;

3. Stel groepen variabelen samen. Indien een productspline deel uitmaakt van een groep kunnen indien gewenst categorieën samengevoegd worden ;
4. Stel sets R en C samen uit de beschikbare groepen ;
5. Op basis van de geselecteerde groepen en de voor de variabelen in deze groepen geldende restricties worden restrictiematrices G_r, h_r en G_c, h_c voor resp. Rij- en Kolomset berekend. Deze zijn gecorrigeerd voor evt. ongelijke intervallengte (§ 4.4 en 4.6) ;
6. Bereken F_r en F_c als R en C in afwijking van het gemiddelde ;

optimalisatie

7. Initialiseer rijsetvector p_0 met evenveel elementen als R dimensies heeft ;
8. Bereken objectvector $r = F_r p / \|F_r p\|$;
9. Minimaliseer $\|F_c q - r\|$ met betrekking tot $G_c q \geq h_c$ (§ 4.4) ;
10. Bereken objectvector $c = F_c q / \|F_c q\|$;
11. Bereken correlatie rc ;
12. Minimaliseer $\|F_r p - c\|$ met betrekking tot $G_r p \geq h_r$ (§ 4.4) ;
13. Indien rc gestabiliseerd is, of een bepaald aantal iteraties bereikt is ga naar stap 15;
14. Ga terug naar stap 8 ;

uitvoer

15. Druk rc af en lever transformatieplots van de variabelen.

De hier omschreven procedure is omwille van de duidelijkheid beschreven voor 1 groep per set. Zijn er n groepen in een set, dan worden stap 9 en/of stap 12 n keer uitgevoerd.

5. VOORBEELD MET EMPIRISCHE GEGEVENS

5.1 Relatie tussen Hinder en Geurconcentratie voor Industriële Bronnen

Gegevens werden verzameld als vervolg op eerder in Nederland uitgevoerd onderzoek naar de relatie hinder-geurconcentratie (Miedema,1990). Hiertoe zijn drie industriële geurbronnen geselecteerd, waarbij de jaarlijkse geuremissie van de installatie werd bepaald. Voorts zijn er bij de bewoners van geselecteerde omringende wijken vragenlijsten afgenomen die betrekking hadden op de kwaliteit van de woonomgeving. Een vraag over de ervaren hinder van de onderzochte bron was hierin opgenomen, eveneens zijn een aantal achtergrondvariabelen gemeten.

De in dit voorbeeld betrokken bedrijven zijn 3 industriële bronnen (BRON), die elk een verschillende, als onaangenaam te classificeren geur verspreiden. De bronnen zullen worden aangeduid met A, B en C.

5.1.1 Dosisvariabelen

Geurconcentratie, Afstand en Hoek

Om in de woonomgeving de geurbelasting te bepalen zijn op basis van de gemeten emissiewaarden verspreidingsberekeningen uitgevoerd om de in de omgeving voorkomende geurbelasting te bepalen. Hierbij werd ondermeer rekening gehouden met de windrichting, windsterkte en de positie van de respondent ten opzichte van de geurbron. Deze berekeningen werden uitgevoerd met het Lange Termijn Frequentie Distributie Model (LTFDM).

Als klimatologie werd de verdeling van uurlijkse waarnemingen van een nabijgelegen weerstation over een periode van 12 maanden voorafgaand aan de afname van de vragenlijsten gebruikt. Per respondent werd het 99.99^e percentiel voor de geurconcentratieverdeling bepaald (C99.99). Deze maat geeft derhalve de geur-

concentratie aan, die gedurende 0.1 % van het jaar overschreden werd. De logaritme van deze maat (LIMM) wordt als dosismaat gebruikt in de analyse.

Afstand en Hoek werden berekend uit de X en Y coördinaten die voor elke respondent bepaald waren. Voor positie ten opzichte van de bron werden de logaritme uit Afstand (LAFS) en Hoek gehercodeerd in 4 kwadranten (HOEK) beschouwd.

5.1.2 Response variabele

Hinder

De ondervonden hinder (HIND) is bepaald door het antwoord op de vraag naar de hinder van industriële bronnen. Dit betreft een meting op een 5 puntsschaal. Hierop is een correctie uitgevoerd, die betrekking heeft op de mogelijkheid, dat de gerapporteerde hinder niet veroorzaakt werd door de bron, waarvoor de geurconcentraties waren bepaald. Hiertoe was een vraag opgenomen met betrekking tot de oorzaak van de hinder. Was dit een andere dan de bron waarvoor de geurconcentraties waren bepaald, dan werd respondent beschouwd als niet gehinderd door de onderzochte bron.

5.2 Vraagstellingen

De centrale vraag die getracht zal worden te beantwoorden luidt: "Hoe ziet de relatie tussen geurimmissie en hinder (geurhinder) eruit en welke factoren zijn mogelijkterwijs van invloed op deze relatie". Om deze vraag te kunnen beantwoorden is een selectie gemaakt uit het totaal van de beschikbare variabelen. Een volledig beeld kan daarom niet verwacht worden, het doel is hier het praktisch uittesten van het programma en het kijken in hoeverre de resultaten aansluiten bij de verwachtingen.

Eerst wordt bepaald, hoe het verband tussen LIMM en HIND eruit ziet. Daarna wordt bekeken of dit verband voor de drie onderzochte lokaties verschilt. Vervolgens zal van een aantal covariaten worden bepaald, of deze het verband tussen

LIMM en HIND belangrijk beïnvloeden. Tenslotte wordt de relatie tussen LAFS, HOEK en HIND onderzocht. In § 5.3 wordt de gevolgde werkwijze toegelicht, resultaten worden besproken in § 5.6.

De covariaten zijn:

- ZUIV : oordeel zuiverheid van de atmosfeer, 5 puntsschaal
- GEZO : oordeel gezondheid van de atmosfeer, 7 puntsschaal
- RISI : risico perceptie, 4 puntsschaal
- LFTD : leeftijd van respondent in jaren
- SEXE : geslacht
- ROOK : wordt er gerookt in huis, dichotoom

Van de variabelen kan de betekenis van de categorienummers worden gevonden in de bijlage.

De gegevens uit rij- en kolomset kunnen gezamenlijk worden weergegeven in een tabel. Het aantal categorieën per variabele is gelijk aan het aantal parameters (orde + aantal interior knots) dat de spline voor die variabele bepaalt. Indien minstens een der variabelen in de tabel metrisch is wordt deze tabel 'smooth' genoemd : per basisspline in de rijset worden gemiddelden berekend voor de basissplines in de kolomset. Tabellen zijn opgenomen in de bijlage. De gepresenteerde figuren stellen steeds de relatie tussen de variabele (x-as) en HIND (y-as) voor.

5.3 Werkwijze

Allereerst zal voor de metrische variabele LIMM de (Pearson) correlatie met ordinale variabele HIND worden bepaald. Indien beide variabelen metrisch opgevat worden, is dit met 4 parameters het meest restrictieve model. Vervolgens zullen transformaties toegestaan worden, waarbij steeds meer parameters in het model worden opgenomen in de vorm van extra knots, waardoor het model steeds minder restrictief wordt. De verwachting is, dat door steeds meer vrijheid in het model te brengen, de correlatie zal toenemen. Vervolgens wordt BRON opgenomen

in de dosis-set. Indien de transformatieplots interpreteerbaar zijn en de correlatie een aanvaardbare waarde aanneemt, zullen de covariaten worden opgenomen.

(LIMM) versus (HIND)

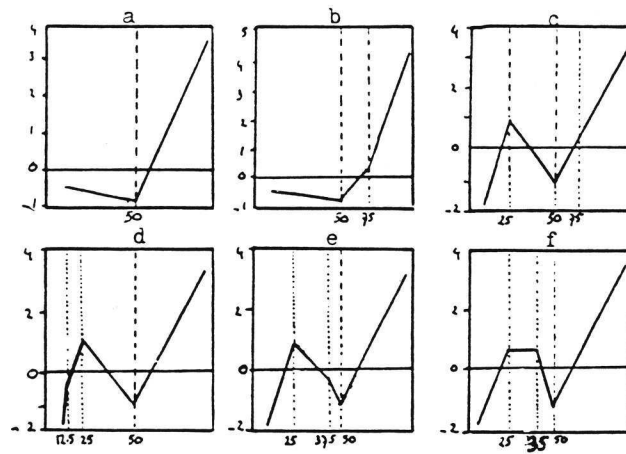
De (Pearson) correlatie (R) tussen LIMM en HIND bedroeg .11. HIND behandeld als een monotoon stijgende stapfunctie verhoogde de correlatie tot .14. Vervolgens werden transformaties voor beide variabelen toegestaan, waarbij steeds op exploratieve wijze knots werden geplaatst of verwijderd bij LIMM. De toegestane transformaties waren een piecewise linear functie voor LIMM en een monotoon stijgende stapfunctie voor HIND. De optimale correlatie wordt aangeduid met R^+ , knotposities worden aangeduid met percentages. 50 komt dus overeen met het 50^e percentiel ofwel de mediaan. Resultaten bij verschillende knotsequenties staan in tabel 5.1.

Tabel 5.1 (LIMM) versus (HIND)

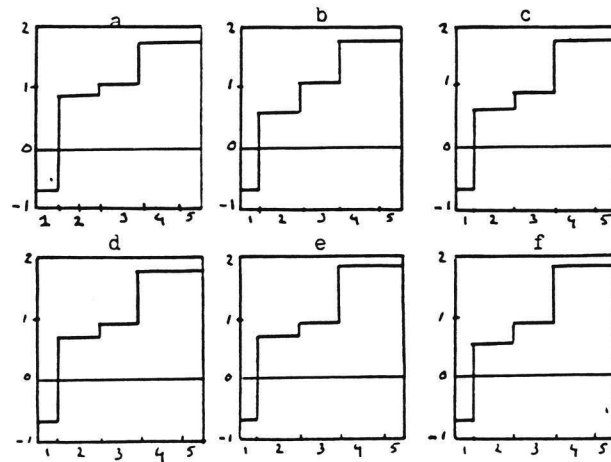
knots (LIMM)			R^+
a.		50	.19
b.		50 75	.19
c.	25	50 75	.24
d.	12.5 25	50	.24
e.	25 37.5	50	.24
f.	25 35	50	.24

Fig 5.1a (a tm. f) zijn de transformaties van LIMM, 5.1b (a tm. f) de bijbehorende schalingen van HIND voor de verschillende knotsequenties uit tabel 5.1(a tm. f).

Figuur 5.1.a 6 transformaties LIMM voor de 6 verschillende knotsequenties uit tabel 5.1.



Figuur 5.1.b 6 transformaties HIND voor de 6 knotsequenties uit tabel 5.1.



De transformatie van LIMM lijkt niet-lineair te zijn. Interpreteerbaar zijn deze resultaten allerminst, de verwachting is dat hinder toeneemt indien de geurimmissie toeneemt. Figuren 5.1.c tm 5.1.f laten zien, dat tussen de knots op het 25^e en het 50^e percentiel van stijging geen sprake is. Met LIMM lijkt iets bijzonders aan de hand te zijn. Uit een nadere beschouwing van de gezamenlijke verdeling van LIMM en BRON blijkt dat de waarden voor LIMM lager dan het 35^e percentiel

uitsluitend bron A betreffen. Plaatsing van een knot op 35 contrasteert dus immisiewaarden in bron A versus die in bronnen B en C. De mediaan bleek in dit geval een ongelukkige eerste keuze voor knotplaatsing.

(LIMM+BRON) versus (HIND).

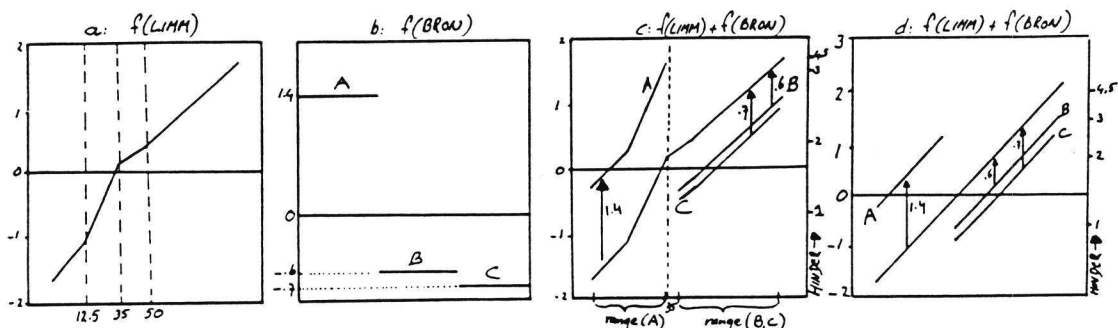
Opname van BRON in het model geeft de volgende resultaten:

Tabel 5.2: (LIMM+BRON) versus (HIND)

knots (LIMM) R^+	
25	35
50	.270

Hoewel toevoeging van BRON op het eerste gezicht slechts een geringe toename van R^+ tot gevolg heeft, is deze variabele uit interpretatieve overweging blijkbaar van belang: toevoeging van BRON beïnvloedt transformatie van LIMM sterk. Deze wijkt nu weinig af van lineariteit (fig 5.2.a).

Figuur 5.2.



De schalingen van BRON (fig 5.2.b) laten zien, dat lokatie A wordt gecontrasteerd versus lokaties B en C, die vrijwel gelijk behandeld worden. Beide transformatieplots gecombineerd wordt geïllustreerd door fig 5.2.c. Deze figuur is geconstrueerd door de gevonden constanten voor de categorieën van BRON op te tellen

bij de transformatie van LMM (fig 5.2.a). Uit deze figuur kan afgeleid worden, dat de vorm van de hinder voor de drie lokaties weliswaar vergelijkbaar is, maar dat de hinder in A optreedt bij aanzienlijk lagere geurconcentratiewaarden. Hindercategorieën staan vermeld aan de rechterzijde van figuren 5.4.c en 5.4.d. Omdat de transformatie van LMM niet al te veel van lineariteit afwijkt, wordt LMM in volgende analyses lineair behandeld. Voorwaarde is dan wel dat LMM steeds additief gecombineerd wordt met BRON. De correlatie met HIND bedraagt dan .26. In dit licht beschouwd was de correlatieverhoging door toevoeging van BRON toch aanzienlijk, nl. van .14 naar .27: het meeste effect van BRON werd reeds verklaard door plaatsing van een knot op het 35^e percentiel van LMM. De gevonden schalingen voor de categorieën van BRON zullen als uitgangspunt worden gehanteerd voor vergelijking tussen bronnen, indien covariaten in het model worden opgenomen.

5.4 Covariaten

Om na te gaan of HIND naast (LMM+BRON) beter voorspeld kan worden, wordt telkens een der covariaten toegevoegd. Per toegevoegde covariaat zal worden gekeken, of de correlatie beduidend toeneemt, ofwel dat deze covariaat de vorm van de relatie tussen LMM en HIND verandert. De eerder gevonden correlatie van .26 zal hierbij als uitgangspunt gehanteerd worden. Het model luidt steeds:

$$(LMM+BRON)+(covariaat) \textit{ versus } (HIND).$$

Variabelen gemeten op een N-puntsschaal worden steeds behandeld als een monotoon stijgende stapfunctie. Figuur 5.4 is een grafiek-matrix, waarbij rijen de set-samenstellingen voorstellen en kolommen de variabelen. De eerste drie kolommen zijn steeds de rijset, waarbij de derde kolom steeds de covariaat is. De laatste kolom is de kolomset.

Waargenomen zuiverheid van de lucht

Toevoeging van oordeel over de zuiverheid van de lucht verhoogt de correlatie van .26 tot .31. De eerder gevonden schaling van BRON blijft vrijwel gelijk. De

toename van de correlatie wordt hoofdzakelijk veroorzaakt door de correlatie tussen ZUIV en HIND. De (erge) hindercategorieën worden iets beter onderscheiden van de middencategorie. Bij ZUIV worden de laagste categorieën samengenomen, ervaren (erg) zuivere lucht gaat samen met geringe hinder, terwijl ervaren onzuivere lucht gepaard gaat met (erge) hinder (fig 5.4.a). De hoogste categorie van ZUIV was leeg. (tabel A.3)

Oordeel gezondheid buitenlucht

Opname van GEZO verhoogt de correlatie tot .31. De schaling van BRON verandert niet wezenlijk (fig 5.4.b), ook hier lijkt de toename van de correlatie toe te schrijven aan de correlatie tussen GEZO en HIND. De hoogste categorie van GEZO was leeg (tabel A.4).

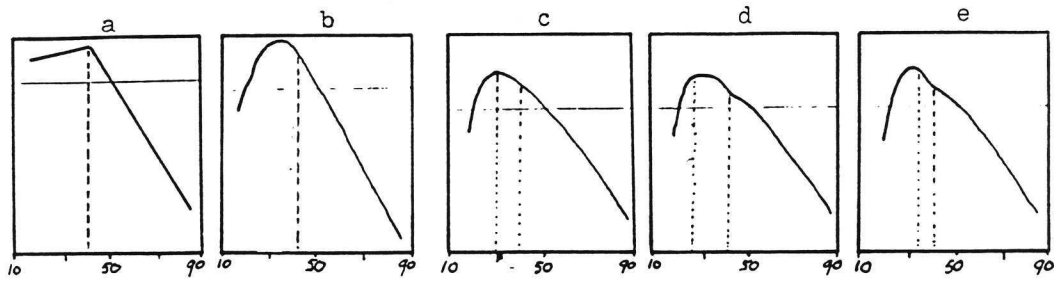
Risico

Toevoeging van de mate van risicoperceptie verhoogt de correlatie aanzienlijk, tot .43. De schaling van BRON verandert niet wezenlijk. De toename lijkt toe te schrijven aan de correlatie tussen risicoperceptie en hinder. Bij de transformatie van RISI worden de middencategorieën samengenomen versus de extremen. Hoog risicoperceptie gaat samen met erge hinder. De hoge hindercategorieën worden beter onderscheiden (fig 5.4.c).

Leeftijd

LFTD lijkt enigszins van belang voor geurhinder. De correlatie neemt toe tot .30, indien LFTD (lineair) wordt toegevoegd. BRON wordt zodanig herschaald, dat het geringe onderscheid tussen B en C vrijwel geheel verdwijnt. De *globale* transformatie van LFTD is dalend. Een piecewise linear transformatie van LFTD met een interior knot ter hoogte van de mediaan deed de correlatie gering stijgen tot .31. De transformatie van LFTD suggereerde een niet-lineair verband met hinder. (fig 5.3.a). Een tweede graads functie voor LFTD deed de correlatie weliswaar nauwelijks stijgen (naar .32), maar laat een lokaal niet-lineair verband zien bij de jongere helft van de respondenten (fig 5.3.b). De functie is aanvankelijk stijgend, maar begint rond het 35^e levensjaar te dalen. Een extra knot ter hoogte van het eerste kwartiel werd bijgeplaatst. (fig 5.3.c)

Figuur 5.3 Graad- en knotvarianten voor LFTD.



Vervolgens werd de knot op het eerste kwartiel verplaatst. Wordt deze iets naar links verplaatst (fig 5.3.d), dan blijkt het maximum iets rechts van de knot te liggen en wordt de functie iets afgeplat. Wordt de knot iets naar rechts verplaatst (fig 5.3.e), dan verliest de functie zijn concave karakter, hetgeen enigszins afbreuk doet aan de interpretatiemogelijkheden. Plaatsing op het 25^e percentiel lijkt een goede keuze. Overwogen kan worden de knot op de mediaan te verwijderen.

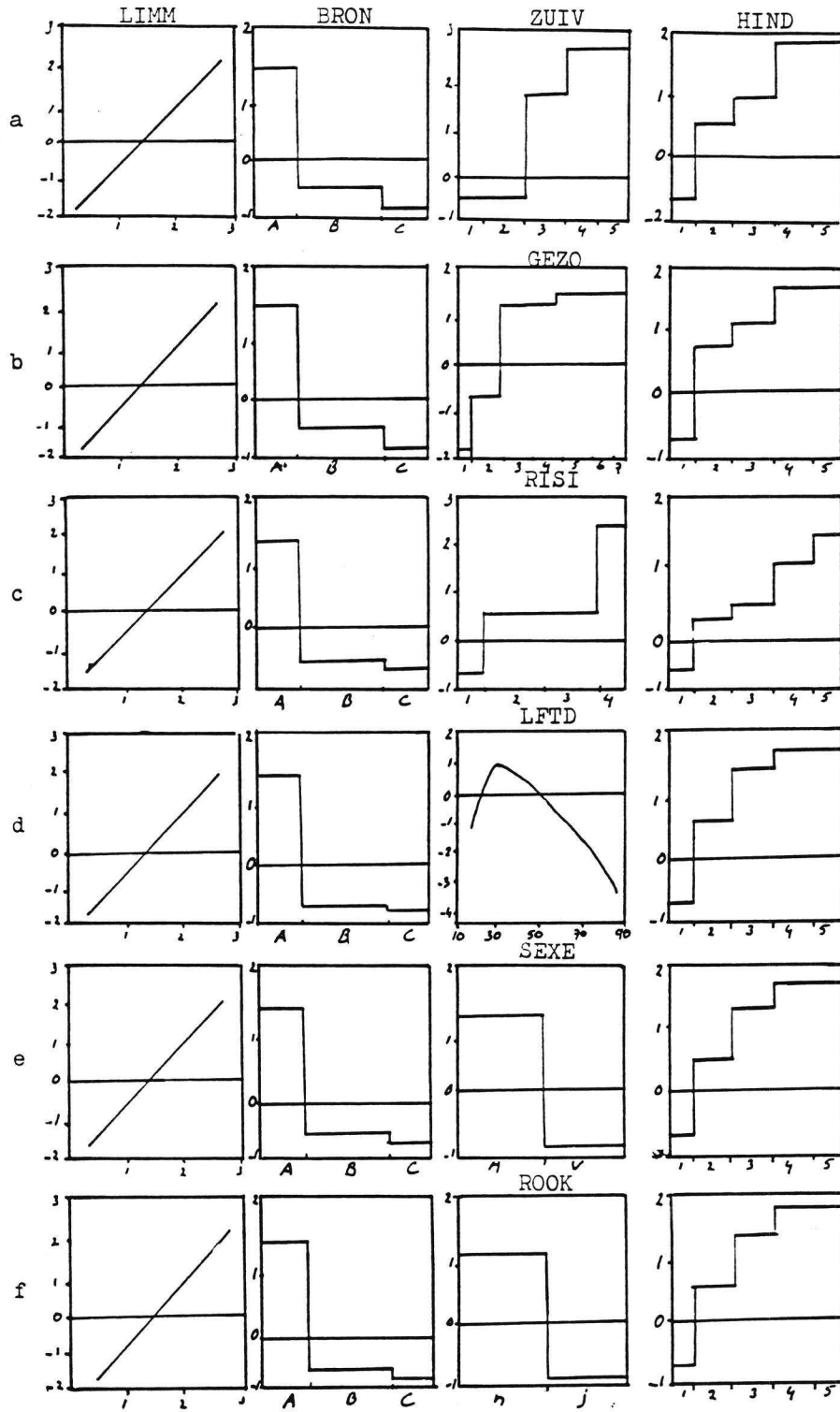
Geslacht

Toevoegen van `SEXE` draagt niet bij tot verhoging van de correlatie. Evenmin wordt de vorm van relatie `LIMM-HIND` beïnvloed: geurhinder lijkt geen verband te houden met geslacht.

Wordt er gerookt in huis

Toevoegen van `ROOK` leverde eveneens geen toename van de correlatie op. Evenmin wordt de vorm van de relatie `LIMM-HIND` beïnvloed: geurhinder lijkt geen verband te houden met rookgedrag in huis.

Figuur 5.4



Tabel 5.3 Overzicht bijdrage van covariaten aan de correlatie tussen (LIMM+BRON) en HIND

Variabele	R	toename R
ZUIV	.31	.05
GEZO	.31	.05
RISI	.43	.17
LFTD	.30	.04
SEXE	.26	0
ROOK	.26	0

5.5 Positie ten opzichte van BRON

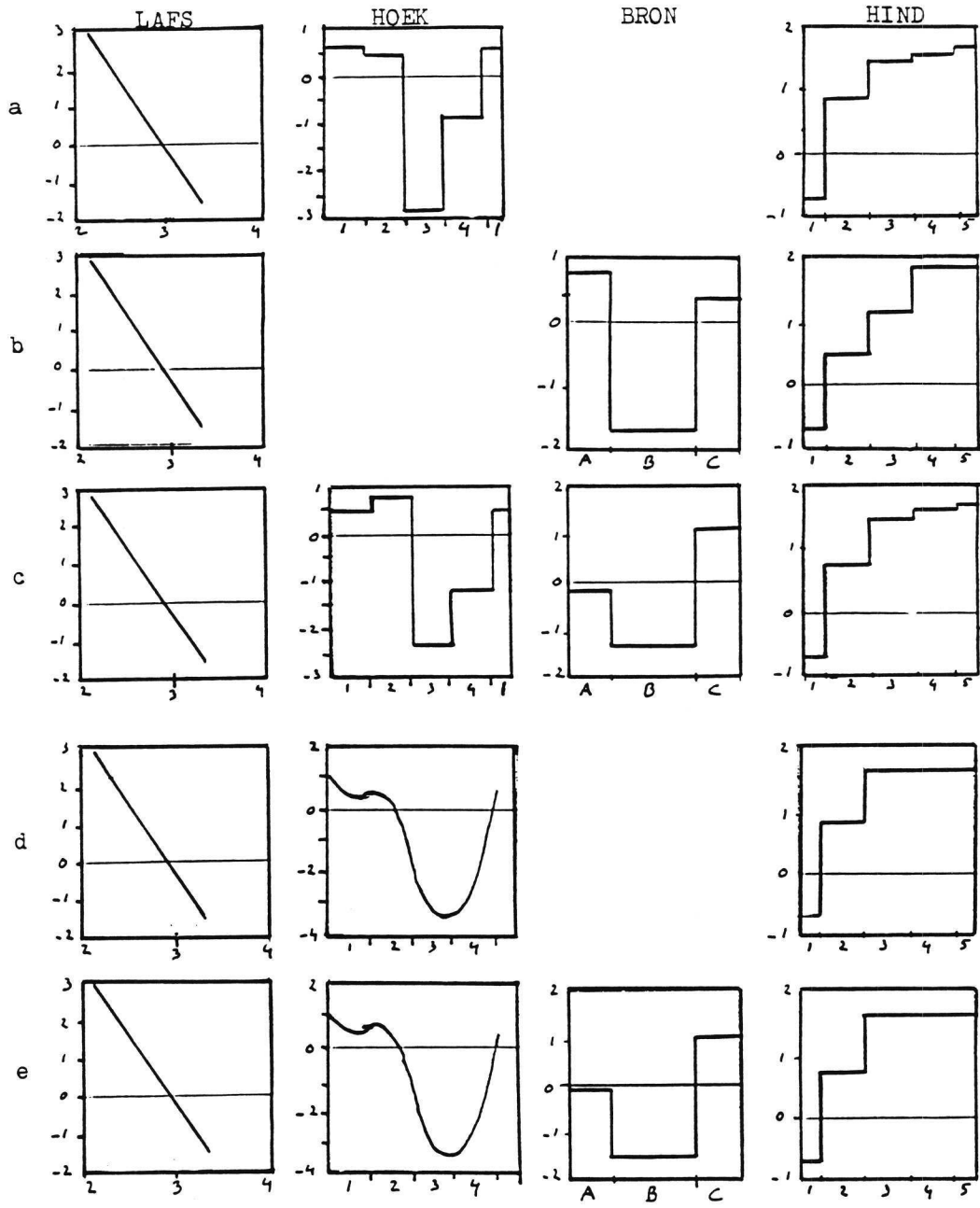
Om dit effect te onderzoeken, kon geen gebruik gemaakt worden van LIMM, omdat positie in het berekenen van LIMM verwerkt was. Afstand tot de bron (LAFS) evenals de hoek waaronder respondent woonde (HOEK) werden beschouwd in relatie met ondervonden hinder. Tussen LAFS en HIND werd een (Pearson) correlatie van .24 gevonden. Toestaan van een monotoon stijgende stapfunctie voor HIND verhoogde de correlatie tot .25. Plaatsing van 1 interior knot ter hoogte van de mediaan bij LAFS gaf geen aanleiding tot het vermoeden van een niet-lineaire relatie tussen LAFS en HIND : de vorm van de relatie, evenals de correlatie bleven gelijk. Vervolgens werden HOEK en BRON toegevoegd.

Opname van HOEK in het model veroorzaakt een toename van de correlatie tot .27 (fig 5.5.a). De herschaling van HOEK suggereert dat de minste hinder vooral in het derde kwadrant wordt gerapporteerd. Toevoegen van BRON verhoogde de correlatie tot .26. Uit de schaling van BRON valt op te maken, dat bij vergelijkbare afstand respondenten bij bronnen A en C iets meer hinder dan gemiddeld rapporteren, terwijl respondenten bij bron B onder het gemiddelde blijven (fig 5.5.b).

BRON en HOEK samen opgenomen verhoogt de correlatie tot .28. De eerder gevonden positieve constante voor bron A verandert nu in een geringe negatieve. De constante voor lokatie C neemt verder toe, terwijl deze voor B vrijwel gelijk blijft (fig 5.5.c).

Getracht werd een tweede graads spline voor HOEK te fitten voor beide modellen. Het resultaat bleek lastig te interpreteren, de figuren zijn ter illustratie opgenomen (fig 5.5.d,e).

Figuur 5.5



5.6 Bespreking van de resultaten

Van de drie onderzochte lokaties is er één, waar relatief weinig geurconcentratie is gemeten: de hoogst gemeten waarde in lokatie A is nog altijd lager dan de laagste gemeten waarde in lokaties B en C. Niettemin is er in lokatie A sprake van een gemiddelde hinder, die vergelijkbaar is met die in lokaties B en C. In een eerdere analyse van dezelfde gegevens werd eveneens een afwijkend patroon in lokatie A aangetroffen (Miedema, 1990). De conclusie, dat geurconcentratie niet de enige factor is die bepalend is voor gerapporteerde hinder lijkt gerechtvaardigd. Een stijgend verband werd gevonden tussen geurconcentratie en hinder, dat weinig van lineariteit afweek. Deze lineariteit kon verklaard worden door rekening te houden met de aard van de geur. De correlatie bedroeg dan .26. Toevoeging van covariaten leverde in geen enkel geval een beduidende verandering op in de *vorm* van de relatie tussen immissie en hinder. De vorm van deze relatie lijkt stabiel. De toename van de correlatie was in de meeste gevallen gering, met uitzondering van de factor risico.

ZUIV,GEZO

Waargenomen onzuiverheid en opvattingen over gezondheid van de lucht veranderden de vorm van het verband niet wezenlijk. De correlaties namen in beide gevallen toe tot .31. Het verschil zou mogelijk kunnen worden verklaard door een verschil in response style, zoals 'extreme' versus 'gematigde' antwoordpatronen. Er werd geen lokatie effect aangetroffen.

RISI

In eerder onderzoek werd het afwijkende hinderpatroon in lokatie A in verband gebracht met historische feiten met betrekking tot de externe veiligheid, die lokaal de nodige publiciteit kregen. Deze interpretatie werd echter niet bevestigd door de hier gevonden resultaten: opname van risico in het model veroorzaakte weliswaar een aanzienlijke stijging van de correlatie, maar aan de vorm van het verband tussen immissie en hinder veranderde weinig. Evenmin veranderden de schaling van BRON, waaruit opgemaakt kan worden dat opname van risico geen verhoogde hinder in lokatie A verklaart. De sterke toename lijkt te kunnen wor-

den toegeschreven aan het verband tussen risicoperceptie en hinderbeleving. Het lijkt gezien de omvang van de stijging niet aannemelijk, dat er slechts sprake is van response bias. Een voor de hand liggende conclusie is dat perceptie van risico's de geurhinder verhoogt.

LFTD

Hoewel de toename van de correlatie bij toevoeging van leeftijd van dezelfde grootte was als bij toevoeging van resp. ZUIV en GEZO kan deze hier niet veroorzaakt worden door response bias. De opvallende lokale niet-lineaire transformatie van leeftijd geeft reden tot nadere beschouwing. Tot ongeveer het 30^e levensjaar werd een stijgend verband tussen hinder en leeftijd gevonden, terwijl dit verband boven het 30^e levensjaar dalend werd. De daling zou enerzijds in verband gebracht kunnen worden met somatische factoren, in die zin dat naarmate men ouder wordt men geleidelijk aan minder goed geuren kan waarnemen. Ook gewenning zou hier een rol kunnen spelen. De stijging anderzijds lijkt niet te verklaren door somatische factoren: het is niet aannemelijk, dat men tot het 30^e levensjaar steeds beter geuren kan waarnemen. Andere factoren dan geurconcentratie lijken de oorzaak van deze stijging. Te denken valt aan sociale of psychologische factoren. Nader onderzoek zou uit moeten wijzen, of dit inderdaad het geval is.

LAFS en HOEK

Het verband tussen afstand en hinder is zoals te verwachten was: naarmate men verder verwijderd is van de bron neemt de hinder af. De hoogste 2 hindercategorieën werden gelijk behandeld. Positie ten opzichte van de bron bleek in lokaties A en C relatief de meeste hinder te verklaren: bij vergelijkbare afstand was de ervaren hinder in lokaties A en C groter dan die in B. Werd rekening gehouden met de windrichting, dan verdween dit effect voor A, terwijl het voor lokatie C toenam. Blijkbaar heeft windrichting in C een toename en in A een afname van de hinder tot gevolg. In B is er vrijwel geen effect.

De ervaren hinder bleek het grootst in het tweede kwadrant op gemiddeld 133 graden, het noordwesten. De minste hinder werd ervaren in het derde kwadrant, gemiddeld op 192 graden, grofweg het west-zuidwesten. Te verwachten zou zijn,

dat de hinder het sterkst zou zijn in het noordoosten, indien men aanneemt dat zuidwestelijke windrichting relatief vaak voorkomt en windrichting een belangrijke rol zou spelen. In figuur 5.5a lijkt dit ook wel enigszins het geval, echter in figuur 5.5c verdwijnt dit effect. Om op basis van de drie in het model opgenomen variabelen zinvolle uitspraken te doen over de relatie tussen aard van de windrichting en hinder lijkt riskant. Gezien de geringe verhoging van de correlatie dient voorzichtigheid bij interpretatie in acht genomen te worden.

5.7 Discussie

De gevonden correlaties zijn voor sociaalwetenschappelijke maatstaven laag te noemen. Bij dosis-response onderzoek daarentegen worden correlaties van .30 en hoger als wezenlijke verbanden aangemerkt en correlaties hoger dan .50 zijn eerder uitzondering dan regel. Bij dosis-response onderzoek is de sterkte van het verband vaak van secundair belang, meer aandacht wordt besteed aan de interpretatie van de vorm. De sterkte van de (toename van de) associatie kan als indicatie dienen voor de voorzichtigheid die in acht genomen moet worden bij de interpretatie van de gevonden vorm.

De in dit hoofdstuk behandelde problematiek bleek op een spaarzame wijze te beschrijven: in de meeste gevallen waren hiervoor weinig parameters nodig in de vorm van knots en graad. Van de drie variabelen, die in aanmerking kwamen voor splines van graad 2 of hoger (LIMM, LAFS en LFTD) bleek een tweede graads-functie alleen zinvol voor LFTD. De mediaan bleek bij LAFS en LFTD een bevredigende keuze voor plaatsing van de eerste knot. Bij LIMM bleek de mediaan een ongelukkige keuze. Voor dergelijke belangrijke verklarende variabelen is het noodzakelijk vooraf de gezamenlijke verdeling met andere belangrijke verklarende variabelen te onderzoeken. Mocht de verdeling daartoe aanleiding geven dan kan, mits de ranges van de categorieën niet te veel overlap vertonen, knot-plaatsing op 'conditionele medianen' worden overwogen. Is er wel veel overlap, dan zal een compromis gezocht moeten worden.

5.8 Conclusie

EQSTRA lijkt een bruikbaar programma te zijn, dat zinvolle resultaten op kan leveren. Door de verschillende transformatie- en restrictie opties wordt het aantal mogelijkheden om op adequate wijze metrische gecombineerd met categorische gegevens te modelleren uitgebreid.

De bij de procedure doorgevoerde menustructuur draagt bij tot beperking van invoerfouten, tot de mogelijkheid snel parameters te kunnen veranderen en resultaten te vergelijken. Tevens wordt op deze manier het APL2 gedeelte vrijwel geheel afgeschermd. Echter: de gebruiker moet over een krachtige en uitgebreide PC configuratie kunnen beschikken, waarop APL2 is geïnstalleerd. (EQSTRA werd ontwikkeld op een 80386 machine, voorzien van mathematische coprocessor, 8 MB direct toegankelijk geheugen en voorzien van een VGA kleurenmonitor). De procedure is nog wel voor uitbreiding vatbaar. Zo is het momenteel nog niet mogelijk 2^e en hogere graads functies te restricteren. Hiervoor is een meer geavanceerde optimalisatie-routine vereist. Ook valt te denken aan uitbreiding van de soorten vormen waarop de functies gerestricteerd kunnen worden. Een uitbreiding voor adequate behandeling van meerdere groepen per set is in een vergevorderd stadium. Een uitbreiding naar meerdere dimensies behoort eveneens tot de mogelijkheden.

LITERATUUR

- BOOR, C.A. DE A Practical Guide to Splines. New York, Springer Verlag, 1978.
- BURG, E. VAN DER & J. DE LEEUW. Non-Linear canonical correlation. Br. J. Math. Stat. Psychol. 34 (1983) 54-80
- BURG, E. VAN DER, J. DE LEEUW & R. VERDEGAAL. Homogeneity Analysis With k Sets of Variables: an Alternating Least Squares Method with Optimal Scaling Features. Psychometrika 33 (1988) 177-197
- CURRY, H.B. & I.J. SCHOENBERG. On spline distribution functions and their limits: the Polya distribution functions. Bull. Am. Math. Soc. 53 (1947) Abstract 380t, 1114
- EDWARDS, A.L. Multiple Regression and the Analysis of Variance and Covariance. San Francisco, Freeman and Company, 1979.
- FLETCHER, R. Practical Methods of Optimization: vol.2. New York, Wiley, 1982.
- GEER, J.P. VAN DE. Analyse van kategorische gegevens. Deventer, Van Loghum Slaterus, 1988.
- GIFI, A. Nonlinear Multivariate Analysis. Chichester (etc.), Wiley, 1990.
- HASTIE, T.J. & R.J. TIBSHIRANI. Generalized Additive Models. London (etc.), Chapman and Hall, 1990.
- HELZER, G. Applied Linear Algebra with APL. Boston (etc.), Little, Brown and Company, 1983.
- HOEK, G. VAN DER & A.W.J. KOLEN. Inleiding Lineaire en Niet-Lineaire programmering. Leiden (etc.), Stenfert Kroese, 1986.
- KRUSKAL, J.B. & R.N. SHEPARD. A nonmetric variety of linear factor analysis. Psychometrika 39 (1974) 123-157
- LANCASTER, P. Curve and Surface Fitting. London (etc.), Academic Press, 1986.
- LAWSON, C.L. & R.J. HANSON. Solving Least Squares Problems. London, Prentice-Hall, 1974.
- MIEDEMA, H.M.E. Equalization by Scaling and Transformation. Leiden, Nederlands Instituut voor Praeventieve Gezondheidszorg-TNO, 1989.
- MIEDEMA, H.M.E. Relatie tussen Hinder en Geurconcentratie voor (bio)industriële bronnen. Leiden, Nederlands Instituut voor Praeventieve Gezondheidszorg-TNO rapport, 1990. (concept)
- RAMSAY, J.O. Some Statistical approaches to multidimensional scaling data. J. Roy. Stat. Soc. 145 (1982) series A: p. 285-312
- RAMSAY, J.O. Monotone Regression Splines in Action. Statistical Science 3 (1988) 4: p. 425-461
- RIJCKEVORSEL, L.L.A. VAN. Horseshoes in Multiple Correspondence Analysis. Leiden, DSWO Press, 1987.
- THISTED, R.A. Elements of Statistical Computing. London, (etc.), Chapman and Hall, 1988.
- WINSBERG, S. & J.O. RAMSAY. Analysis of Pairwise Preference Data Using Integrated Splines. Psychometrika 46 (1981) 171-186
- WINSBERG, S. & J.O. RAMSAY. Monotone spline transformations for dimension reduction. Psychometrika 48 (1983) 575-595

BIJLAGE

Hindercategorieën zijn:

- 1: niet
- 2: net niet
- 3: net
- 4: hinderlijk
- 5: erg hinderlijk

Tabel A.1 (Smoothed)

LIMM	HIND					N
	1	2	3	4	5	
1.0	69.5	8.4	5.4	9.8	6.9	471
1.8	62.2	9.9	7.7	13.4	6.8	388
totaal	66.2	9.1	6.4	11.4	6.9	859

Tabel A.2:

BRON		HIND					N
		1	2	3	4	5	
1	A	68.6	7.7	4.7	10.0	9.0	299
2	B	58.2	13.6	7.7	13.6	6.8	220
3	C	69.4	7.4	7.1	11.2	5.0	340
totaal		66.2	9.1	6.4	11.4	6.9	859

Tabel A.3

ZUIV	HIND					N
	1	2	3	4	5	
1 heel zuiver	74.1	8.6	5.2	8.6	3.4	58
2 zuiver	69.4	8.7	6.3	9.2	6.5	601
3 niet helemaal	54.4	12.4	8.3	16.6	8.3	169
4 niet	54.8	0.0	0.0	32.3	12.9	31
totaal	66.2	9.1	6.4	11.4	6.9	859

Tabel A.4

GEZO	HIND					N
	1	2	3	4	5	
1 erg	83.3	4.2	0.0	4.2	8.3	24
2 gezond	71.2	8.3	6.3	8.9	5.3	507
3 tamelijk	57.5	10.6	7.2	16.4	8.2	207
4 niet	57.8	8.4	9.6	14.5	9.9	83
5 tamelijk ong	47.1	17.6	0.0	23.5	11.8	17
6 ongezond	61.9	14.3	0.0	9.5	14.3	21
totaal	66.2	9.1	6.4	11.4	6.9	859

Tabel A.5

RISI	HIND					N
	1	2	3	4	5	
1 nooit	76.6	7.5	5.2	7.3	3.4	534
2 zelden	51.9	14.3	11.7	15.6	6.5	77
3 soms	57.4	11.6	7.1	14.2	9.7	155
4 vaak	33.3	9.7	7.5	26.9	22.6	93
totaal	66.2	9.1	6.4	11.4	6.9	859

Tabel A.6. (Smoothed)

LFTD	HIND					N
	1	2	3	4	5	
22	67.2	12.1	2.7	12.9	5.1	54
29	59.6	10.8	7.0	15.6	7.0	143
43	62.6	9.0	7.7	13.0	7.7	372
60	72.0	8.1	5.6	7.6	6.6	198
71	78.3	6.9	4.1	5.7	4.9	92
totaal	66.2	9.1	6.4	11.4	6.9	859

Tabel A.7

SEXE	HIND					N
	1	2	3	4	5	
1 man	66.0	7.8	6.9	13.0	6.3	347
2 vrouw	66.4	10.0	6.1	10.4	7.2	512
totaal	66.2	9.1	6.4	11.4	6.9	859

Tabel A.8

ROOK	HIND					N
	1	2	3	4	5	
0 nee	64.4	9.3	6.7	12.6	7.0	388
1 ja	67.7	8.9	6.2	10.4	6.8	471
totaal	66.2	9.1	6.4	11.4	6.9	859

Tabel A.9. (Smoothed)

LAFS	HIND					N
	1	2	3	4	5	
2.7	58.8	10.2	7.6	14.9	8.5	292
3.0	70.1	8.5	5.8	9.6	6.0	567
totaal	66.2	9.1	6.4	11.4	6.9	859

Tabel A.10.

HOEK	gem	HIND					N
		1	2	3	4	5	
1(0-90)	30	67.8	7.5	8.1	10.3	6.4	360
2(90-180)	133	57.2	13.2	6.6	12.8	10.1	257
3(180-270)	192	73.9	5.8	2.9	14.5	2.9	69
4(270-360)	336	73.4	7.5	4.0	10.4	4.6	173
totaal		66.2	9.1	6.4	11.4	6.9	859

Reprografie NIPG-TNO
Projectnummer: 3550