RR 87-14

HOMALS VOOR BEGINNERS

Gerda M. van den Berg

Rijksuniversiteit Leiden
Vakgroep Datatheorie en
Vakgroep Sociaal-Wetenschappelijke Informatica
Middelstegracht 4
2312 TW Leiden

December 1986

INHOUDSOPGAVE:

1. Doel en globale werkwijze van HOMALS

HOMALS is de naam van een computerprogramma dat geschreven is om nominale variabelen te analyseren. HOMALS staat voor "homogeneity analysis by means of alternating least squares". "Alternating least squares" verwijst naar het algoritme waarop het computerprogramma is gebaseerd; "homogeneity" verwijst naar het doel van de analyse.

HOMALS kan gebruikt worden wanneer men wil nagaan of er relaties bestaan tussen variabelen die op nominaal niveau gemeten zijn. Nominaal betekent dat men van de categorieën van elke afzonderlijke variabele slechts weet dat de ene categorie de andere niet is, maar dat men niet weet hoe de categorieën zich onderling verhouden. Een voorbeeld van zo'n variabele is 'godsdienst', waarbij als categorieën onderscheiden kunnen zijn: gereformeerd, hervormd, katholiek, niet godsdienstig. Een ander voorbeeld is 'burgerlijke staat' met als categorieën: ongehuwd, gehuwd, gescheiden, weduwe/-weduwnaar.

Of een variabele als nominaal wordt beschouwd, is afhankelijk van de keuze die de onderzoeker zelf maakt. Men kan schooltype in het secundair onderwijs beschouwen als een nominale variabele, maar ook als een ordinale variabele en dan de categorieën ordenen van laag naar hoog. HOMALS gaat er echter van uit dat alle variabelen nominaal zijn. Wil men een variabele beslist als ordinale of numerieke variabele analyseren, dan moet men geen HOMALS toepassen.

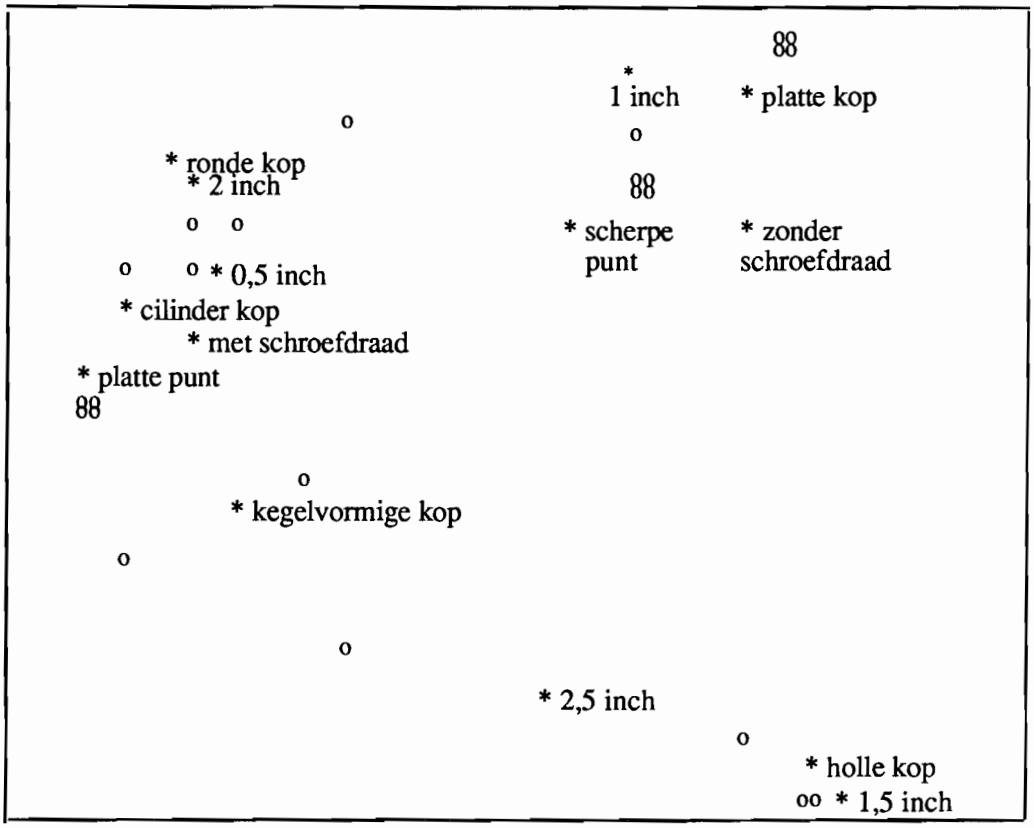
Dat een variabele als nominaal wordt beschouwd, houdt in dat elke variabele een beperkt aantal discrete categorieën moet bevatten. Continue variabelen zoals lengte, leeftijd en vaak ook testscores moeten eerst teruggebracht worden tot een beperkter aantal categorieën, voordat zij met behulp van HOMALS kunnen worden geanalyseerd.

Voert men HOMALS uit op een set nominale variabelen, dan levert dat als oplossing een ruimtelijke afbeelding, een plaatje, op. Men kan zelf bepalen of HOMALS zo'n afbeelding voor één, twee of meer dimensies moet uitrekenen. In deze afbeelding worden zowel de geobserveerde eenheden (individuen, cases, in het algemeen: objecten) weergegeven als de categorieën. De variabelen zelf zijn als zodanig niet terug te vinden in de afbeelding, maar wel de categorieën. Men komt dus bijvoorbeeld schoolvak op geen enkele manier als variabele in de oplossing tegen, maar wel de categorieën rekenen, taal, wereldoriëntatie, gymnastiek, enz. In de afbeelding komen objecten dichter bij elkaar te liggen naarmate zij meer op elkaar lijken in termen van de categorieën waartoe zij behoren. Het kan zijn dat er tien variabelen zijn. Als twee objecten bij negen van die tien variabelen in dezelfde categorie worden ingedeeld, dan liggen zij dichter bij

elkaar in de afbeelding dan twee objecten die bij slechts één variabele in dezelfde categorie vallen.
Het is daarbij op geen enkele manier noodzakelijk dat elke variabele dezelfde categorieën bevat. Een object kan tot de volgende categorieën behoren: vrouw, verkoopster, stemt CPN, ongehuwd. Een tweede object kan gekenmerkt worden door: vrouw, verkoopster, stemt CPN, gehuwd. En een derde object: man, boekhouder, stemt VVD, gehuwd. Het eerste object ligt in de oplossing veel dichter bij het tweede object dan bij het derde object. De eerste twee objecten vallen immers op drie van de vier variabelen in dezelfde categorie, terwijl het eerste en het derde object bij geen enkele variabele tot dezelfde categorie behoren. In het algemeen geldt dat in de afbeelding de afstand tussen objectpunten kleiner is naarmate hun patronen van categorieën (ook wel: antwoordpatronen) meer op elkaar lijken. Objecten met identieke patronen vallen samen.
Hierdoor is het mogelijk groepen objecten op te sporen die qua antwoordpatronen veel overeenkomen. Zo'n groep noemen we homogeen. De homogeniteit waarnaar we op zoek zijn, betreft in dit geval dus de mate van overeenkomst tussen objecten. Groepen objecten die weinig met elkaar overeenkomen, liggen zoveel mogelijk gescheiden van elkaar in de afbeelding. Men kan HOMALS dus gebruiken om homogene clusters objecten te vinden.
Met evenveel recht kan men echter spreken over de homogeniteit van de variabelen. Als de categorieën van de ene variabele sterk samenhangen met die van een andere variabele, worden de objecten door beide variabelen in ongeveer dezelfde groepen verdeeld. Dergelijke variabelen noemt men homogeen, omdat zij op dezelfde manier onderscheid maken tussen de objecten. Daarmee is overigens niet gezegd dat zij ook hetzelfde meten.
Behalve objectpunten bevat een HOMALS-oplossing ook categoriepunten. Elke categorie wordt afgebeeld in het zwaartepunt van de objecten die tot de betreffende categorie behoren. Het zwaartepunt is het punt dat de kleinste totale afstand heeft tot de betreffende groep objecten.
Categoriepunten liggen derhalve dichter bij de objecten die tot die categorie behoren dan bij de objecten die niet tot die categorie behoren. Is er een homogene groep objecten die voor het merendeel tot dezelfde categorie behoren, dan ligt deze categorie dicht bij deze objecten. Behoort een tweede groep objecten tot een andere categorie, dan ligt deze andere categorie in het zwaartepunt van die tweede groep. Wanneer beide groepen onderling weinig overeenkomen, dan liggen zij duidelijk gescheiden van elkaar in de oplossing. Ook de beide categorieën liggen dan gescheiden van elkaar in de oplossing.

Figuur 1 is een voorbeeld van een HOMALS-oplossing in twee dimensies. De gegevens die geanalyseerd zijn, hadden betrekking op een verzameling spijkers en

schroeven. Van deze "objecten" waren vier uiterlijke kenmerken gecodeerd, n.l. de aan- of afwezigheid van een schroefdraad, de vorm van de kop, de lengte en de vorm van de punt. In de afbeelding zijn de objecten weergegeven als cirkeltjes en de verschillende categorieën als sterretjes. Bij iedere categorie is de betekenis aangegeven. De catego - rieën die bij dezelfde variabelen behoren, moeten echter nog wel bij elkaar gezocht worden. Zo horen de categorieën "platte punt" en "scherpe punt" bij dezelfde variabele. In de afbeelding is dat niet weergegeven.



Figuur 1 Voorbeeld van een HOMALS-oplossing. 0 = object; * = categorie

Op basis van een afbeelding als Figuur 1 kunnen twee soorten conclusies worden getrokken. Allereerst kan worden nagegaan of er meer of minder homogene groepen objecten te onderscheiden zijn. Men kan bijvoorbeeld op grond van deze HOMALS-oplossing besluiten de objecten in vier groepen te verdelen overeenkomstig het kwa - drant waarin zij liggen. Aldus worden schroeven, spijkers, bouten en nagels onder - scheiden. Men creëert daarmee met andere woorden een typologie. Zo'n typologie kan als uitgangspunt voor andere analyses dienen.

Het tweede type conlusies betreft de samenhang tussen variabelen. Nauwkeuriger geformuleerd: de samenhang tussen categorieën van variabelen. In Figuur 1 zijn de categorieën "holle kop" en "1.5 inch" dicht bij elkaar rechtsonder in de afbeelding te vinden. Dat betekent dat beide categorieën vooral in combinatie met elkaar voorkomen. Weet je eenmaal dat bij een bepaald object de ene categorie is waargenomen, dan kun je met behoorlijke zekerheid voorspellen dat ook de andere categorie zich zal voordoen bij het betreffende object. Daarmee is verder nog niets gezegd over een mogelijke relatie tussen de variabelen "vorm van de kop" en "lengte".

HOMALS is vooral een beschrijvende techniek. Dit houdt in dat men meestal verschil - lende oplossingen bekijkt voordat men besluit aan welke beschrijving van de dataset men de voorkeur geeft. Soms wordt een oplossing gedomineerd door één of twee bijzondere objecten die op veel punten van alle andere objecten afwijken. Dan kan het beter zijn deze excentriekelingen te verwijderen en HOMALS opnieuw toe te passen op de resterende objecten van de dataset. Bij de beschrijving van de resultaten moeten dan enerzijds de excentriekelingen besproken worden en anderzijds de HOMALS-oplossing met de overige objecten.
Vergelijkbare overwegingen gelden voor afzonderlijke categorieën en variabelen. Wan - neer een enkele categorie of variabele het beeld onduidelijk maakt, is het verstandiger deze niet op te nemen in de analyses met HOMALS. Steeds geldt dat men er dan wel aan moet denken de verwijderde informatie afzonderlijk te rapporteren.
Het zal inmiddels waarschijnlijk opgevallen zijn, dat er nog helemaal niet gesproken is over parameters of grootheden waaruit kan worden afgelezen of een HOMALS-oplossing al dan niet significant is. Omdat HOMALS een beschrijvende techniek is en niet een toetsende techniek, ontbreken dergelijke maten. HOMALS wordt gebruikt om na te gaan welke relaties in de dataset aanwezig zijn. Er kan niet uit worden afgeleid hoe groot de kans is dat die relaties ook in werkelijkheid voorkomen.

In het navolgende wordt in enigszins vereenvoudigde vorm het rekenalgoritme waarvan HOMALS gebruik maakt aan de orde gesteld. Dat leidt tot een nauwkeuriger beschrijving van een aantal kenmerken van de oplossingen en van een aantal grootheden die in het programma berekend worden. Vervolgens worden verschillende verwante problemen behandeld en deze inleiding besluit met een voorbeeld van het gebruik van HOMALS.

2. Het algoritme

De invoer voor het programma HOMALS bestaat uit een dataset met een aantal objecten die op verschillende variabelen zijn gecodeerd. Het linkerdeel van Tabel 1 toont zo'n datamatrix met in dit geval 10 objecten en 3 variabelen. De categorieën van de variabelen zijn met letters aangeduid om eens te meer de nadruk te leggen op het feit dat het om nominale variabelen gaat. Als eerste stap kent HOMALS aan ieder object tamelijk willekeurig een score ofwel een kwantificatie toe. Als enige restrictie geldt daarbij dat het gemiddelde van deze scores gelijk aan 0 moet zijn. Vervolgens worden de objectscores gestandaardiseerd zodat de variantie van de objectscores gelijk aan 1 is. In Tabel 1 zijn de scores toegekend op grond van de waarde van ieder object op de eerste variabele. Objecten met een gelijke waarde op deze variabele hebben dezelfde score gekregen.

Tabel 1 Voorbeeld van een initiële berekening van objectscores

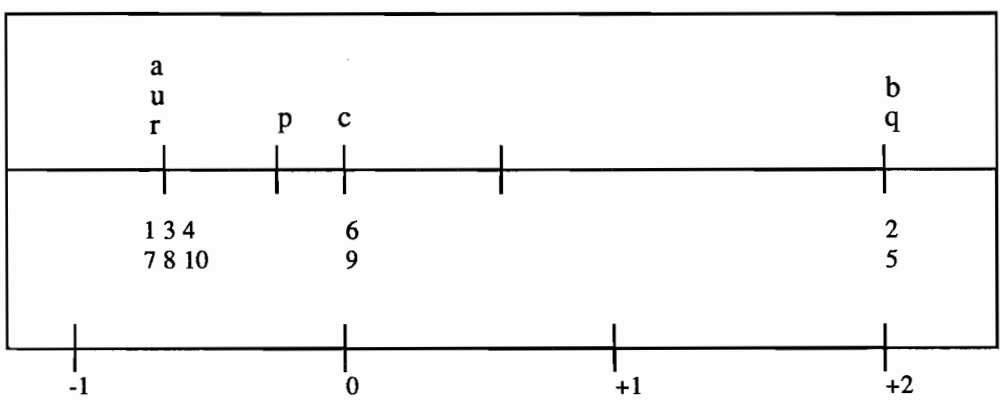
oorspronkelijke datamatrix				objectscores gekozen o.g.v. variabele 1	gestandaardiseerde objectscores
	variabelen				
object	1	2	3		
1	a	p	u	-1	-0.65
2	b	q	v	3	1.94
3	a	r	v	-1	-.065
4	a	p	u	-1	-.065
5	b	p	v	3	1.94
6	c	p	v	0	0
7	a	p	u	-1	-0.65
8	a	p	v	-1	-0.65
9	c	p	v	0	0
10	a	p	v	-1	-0.65

In de volgende stap rekent het programma aan de hand van deze objectscores uit welke kwantificatie toegekend moet worden aan de categorieën van de variabelen. Een categorie wordt gekwantificeerd als het gemiddelde van de objectscores van de objecten die tot die categorie behoren.

Tabel 2 Voorbeeld van de berekening van categoriekwantificaties door HOMALS

object	gestandaardiseerde objectscores	categorie kwantificaties
1	-0.65	a = -0.65
2	1.94	b = 1.94
3	-0.65	c = 0
4	-0.65	
5	1.94	p = 0.16
6	0	q = 1.94
7	-0.65	r = -0.65
8	-0.65	
9	0	u = -0.65
10	-0.65	v = 0.28

In Tabel 2 zijn deze kwantificaties berekend op grond van de eerder toegekende gestandaardiseerde objectscores. Categorie v heeft de kwantificatie 0.28 gekregen. Dit is het gemiddelde van de score van de objecten 2, 3, 5, 6, 8, 9 en 10 die allemaal tot deze categorie behoren.
In dit voorbeeld zijn de objectscores in eerste instantie toegekend op grond van de waarde op de eerste variabele. Dat vinden we terug in de kwantificatie van de categorieën van deze variabele.



Figuur 2: Voorbeeld van een initiële oplossing in één dimensie

Figuur 2 geeft een afbeelding van deze oplossing. Elk object en elke categorie heeft slechts één kwantificatie en daarom is de oplossing een afbeelding in één dimensie ofwel een rechte lijn. Voor de duidelijkheid zijn de categorieën boven deze lijn geschreven en de objecten eronder. Het programma HOMALS is met deze oplossing niet

tevreden. Dat valt intuïtief te begrijpen wanneer je je realiseert dat de objecten 1, 3, 4, 7, 8 en 10 allemaal dezelfde score hebben. Uit de datamatrix blijkt echter dat deze objecten niet op alle variabelen dezelfde waarde hebben. Rekentechnisch vinden we dit terug in de variantie en het verlies ("loss") van de gekwantificeerde variabelen.

Tabel 3 Voorbeeld van variantie en verlies van gekwantificeerde variabelen

categorie kwantificaties	variantie van de gekwantificeerde variabelen	verlies (1-variantie)
a = -0.65 b = 1.94 c = 0	1.000	0
p = 0.16 q = 1.94 r = -0.65	0.437	0.563
u = -0.65 v = 0.28	0.179	0.821
gemiddelde variantie	0.539	gemiddeld verlies 0.461

In Tabel 3 staan de varianties van de variabelen. De variantie van de eerste variabele is gelijk aan de variantie van de objectscores, n.l. 1, omdat de objectscores gekozen zijn op grond van deze variabele. Het verlies is gedefinieerd als de variantie van de afstand van de objectscores tot de categorieën waartoe zij behoren. Dit is hetzelfde als de varian - tie van de objectscores (die gelijk aan 1 is) minus de variantie van de gekwantificeerde variabele. De eerste variabele heeft geen verlies aangezien alle objectscores samenvallen met de bijbehorende categorieën van de betreffende variabele. Voor de beide andere variabelen geldt dat echter niet. Het totale gemiddelde verlies van deze oplossing bedraagt derhalve 0.461. Naarmate het verlies lager is, is de oplossing beter. Een klein verlies betekent namelijk dat de objecten dicht liggen bij de categoriepunten die op hen van toepassing zijn. De meest eenvoudige manier voor HOMALS om een prima oplossing te bewerkstelligen in termen van de afstanden tussen objecten en categorieën komt erop neer dat alle objecten en alle categorieën in hetzelfde punt worden afgebeeld. Zo'n oplossing levert echter geen enkele informatie. Ter voorkoming van zulke gede - genereerde oplossingen is dan ook de restrictie aan het programma toegevoegd dat de variantie van de objectscores gelijk aan 1 moet zijn.

De oplossing die in Tabel 2 en Figuur 2 is afgebeeld, leidt zoals gezegd tot een gemid - deld verlies van 0.461. HOMALS gaat nu proberen deze oplossing te verbeteren. Daar - toe worden nieuwe objectscores uitgerekend op grond van de in eerste instantie verkre - gen categoriekwantificaties. Elk object krijgt als nieuwe score het gemiddelde van de kwantificaties van de categorieën waartoe het behoort. Object 1 heeft bijvoorbeeld de categorieën a, p en u en krijgt als nieuwe score: (-0.65 -0.16 -0.65) / 3 = -0.48. Voor ieder object kan aldus een nieuwe score worden berekend. Tabel 4 geeft daarvan de uitwerking.

Tabel 4 Voorbeeld van de iteratieve fase van HOMALS

initiële categorie kwantificaties	nieuwe objectscores	gestandaardiseerde objectscores	nieuwe categorie kwantificaties
a = -0.65	-0.48	-0.85	a = -0.62
b = 1.94	1.38	2.42	b = 1.81
c = 0	-0.34	-0.59	c = 0.07
	-0.48	-0.85	
p = 0.16	0.68	1.20	p = 0.23
q = 1.94	0.04	0.07	q = 2.42
r = 0.65	-0.48	-0.85	r = 0.59
	-0.18	-0.31	
u = -0.65	0.04	0.07	u = 0.85
v = 0.28	-0.18	-0.31	v = 0.36

De nieuwe objectscores worden gestandaardiseerd zodat hun variantie wederom 1 is. De iteratie is compleet wanneer op basis van deze nieuwe gestandaardiseerde object - scores ook nieuwe categoriekwantificaties zijn berekend.

HOMALS rekent dan voor deze nieuwe gegevens het verlies uit. Is dit verlies merkbaar minder dan het verlies van de vorige oplossing, dan zal HOMALS een nieuwe iteratie in gang zetten. Dit proces van het afwisselend herberekenen van de objectscores en van de categoriekwantificaties ("alternating least squares") wordt zolang voortgezet totdat geen vooruitgang meer wordt geboekt in termen van vermindering van verlies.

Een HOMALS-oplossing kan voor meer dan één dimensie worden berekend. Nadat objectscores en categoriekwantificaties in één dimensie zijn berekend, kiest HOMALS - indien dat gewenst wordt - een tweede reeks objectscores. Het programma doet dat zodanig dat de objectscores in deze tweede dimensie ongecorreleerd zijn met de object - scores in de eerste dimensie. Is de oplossing voor de tweede dimensie berekend, dan kan het programma verder gaan met de derde dimensie. Enzovoort.

3. Kenmerken van een HOMALS-oplossing

HOMALS levert een afbeelding van objecten en categorieën op. Het gemiddelde van de objectscores is 0 en de variantie is 1. Objecten met hetzelfde antwoordpatroon vallen samen in de oplossing. Zij krijgen met andere woorden dezelfde score.
Een categoriekwantificatie is het gemiddelde van de scores van de objecten die tot die categorie behoren. Is er slechts één object in een bepaalde categorie, dan valt deze categorie samen met het betreffende objectpunt. De variantie van een gekwantificeerde variabele wordt de discriminatiewaarde van deze variabele genoemd. Naarmate de variantie groter is, discrimineren de categorieën van een variabele beter tussen de objecten, want dan is de gemiddelde afstand van de objecten tot hun bijbehorende categorie geringer. Hoe groter de discriminatiewaarde van een variabele, des te beter is de gevonden oplossing voor deze variabele. In de afbeelding zien we in zo'n geval dat iedere categorie van de variabele dicht bij de objecten ligt die tot die categorie behoren en ver van de overige objecten vandaan.
De gemiddelde discriminatiewaarde van alle variabelen is de eigenwaarde. Het verlies van een oplossing is gelijk aan 1-eigenwaarde.
Naar keuze kan een HOMALS-oplossing voor één of meer dimensíes worden berekend. Het maximale aantal dimensies bedraagt het totale aantal categorieën minus het aantal variabelen. De oplossingen in de verschillende dimensies zijn genest. Dit betekent dat de eerste dimensie altijd hetzelfde is, hoeveel dimensies men ook opvraagt. Hetzelfde geldt voor alle volgende dimensies. Lagere dimensies hebben altijd een betere gemiddelde discriminatiewaarde dan hogere dimensies. Dat neemt niet weg dat een afzonderlijke variabele soms beter discrimineert in een hogere dan in een lagere dimensie.
HOMALS is bedoeld als techniek om een dataset te beschrijven. Zo'n beschrijving wordt ingewikkelder naarmate een oplossing meer dimensies telt. Daarom is het verstandig niet meer dan twee of drie dimensies te hanteren. Naarmate men meer dimensies hanteert, vervaagt het verschil met een beschrijving aan de hand van de oorspronkelijke variabelen steeds meer en is de winst die het gebruik van HOMALS oplevert, steeds geringer.
Soms valt het zelfs aan te bevelen slechts één dimensie te kiezen. Dat is vooral het geval wanneer men een aantal variabelen heeft die in grote lijnen allemaal hetzelfde meten. Men kan dan de objectscores gebruiken om de informatie uit de totale groep variabelen samen te vatten. In zulke gevallen doet men er echter verstandig aan in plaats van HOMALS het programma PRIMALS te kiezen. Dat programma is namelijk speciaal ontwikkeld voor ééndimensionale oplossingen en het verricht verschillende berekenin -

gen die in zulke situaties nuttig zijn en die door HOMALS niet worden uitgevoerd. Een toepassing daarvan is te vinden in De Leeuw & Kreft (1985).
Tenslotte dient nog opgemerkt te worden dat er verschillende mogelijkheden zijn voor de behandeling van missing values. Objecten waarvan de codering op een bepaalde variabele ontbreekt, doen gewoon mee in de analyse. Zij krijgen derhalve een object score toegekend op basis van hun categorieën bij de overige variabelen. Een onderzoe - ker kan echter van mening zijn dat de objecten die op een bepaalde variabele 'missing' zijn, in zekere zin op elkaar lijken. Denkt men aan een vraag als die naar het inkomen, dan is het niet onwaarschijnlijk dat alle respondenten die deze vraag overgeslagen heb - ben, iets gemeen hebben. In zulke gevallen kan de onderzoeker ervoor kiezen de cate - gorie 'missing' toe te voegen aan de variabele. Deze categorie wordt vervolgens op dezelfde wijze door HOMALS behandeld als alle andere categorieën in de analyse. Het is ook mogelijk elk object met een missing value als een aparte categorie van de betref - fende variabele te beschouwen. Categorieën met een lage frequentie van voorkomen komen echter over het algemeen nogal aan de buitenkant van een HOMALS-afbeelding te liggen en domineren daardoor de oplossing. Daarom verdient zo'n benadering van missing values lang niet in alle gevallen aanbeveling.

4. Beschrijving van het programma

De uitvoer van het programma HOMALS levert allerlei informatie standaard of naar keuze op. In het kort worden hieronder de verschillende mogelijkheden besproken en wordt hun betekenis aangegeven. Omdat in de uitvoer van het programma gebruik wordt gemaakt van engelstalige begrippen, worden ook hier de engelse begrippen ge - hanteerd.
History of iterations: Afwisselend berekent HOMALS objectscores en categorie - kwantificaties. Gezamenlijk vormen beide berekeningen een iteratie van het programma. Het programma itereert net zo lang totdat het convergentiecriterium is bereikt. Standaard gebruikt HOMALS als convergentiecriterium dat de winst tussen twee opeenvolgende iteraties minstens 0.00015 moet bedragen. Is de winst minder, dan is het programma klaar met rekenen.
In het overzicht van de iteraties is te vinden hoeveel iteraties zijn uitgevoerd, hoe groot per iteratie de totale fit van de oplossing was en hoe groot het verschil in totale fit was tussen twee opeenvolgende iteraties.
Eigenvalues: Per dimensie berekent HOMALS de eigenwaarden. De eigenwaarde is de gemiddelde discriminatiewaarde van de variabelen op de betreffende dimensie.

Naarmate een discriminatiewaarde dichter bij 1 ligt, is de gemiddelde afstand van de objecten tot hun categorieën van die variabele geringer.
Telt men de eigenwaarden van alle dimensies die men heeft laten berekenen op, dan verkrijgt men de totale fit van de oplossing. Deze totale fit staat ook vermeld in het overzicht van de iteraties.
Discrimination measures per variable per dimension: Voor elke dimensie en iedere variabele apart wordt de discriminatiewaarde gegeven. Hoe hoger de discriminatiewaarde, des te beter "doet die variabele het" op de betreffende dimensie.
Plot of discrimination measures: Voor de eerste twee dimensies kunnen de discriminatiewaarden ook geplot worden. Daarmee wordt in feite dezelfde informatie verstrekt als in de hierboven genoemde tabel.
Object scores: De objectscores worden weergegeven in een tabel met net zoveel kolommen als er dimensies in de oplossing zijn. De objectscores in de verschillende dimensies zijn ongecorreleerd.
Category quantifications: Net als de kwantificatie van de objecten wordt ook een overzicht gegeven van de kwantificatie van de categorieën van de variabelen. De categoriekwantificaties in de verschillende dimensies zijn niet per sé ongecorreleerd.
Plot of object scores: Behalve een tabel met objectscores levert HOMALS ook plots van objectscores op. Dergelijke plots kunnen alleen voor de eerste twee dimensies verkregen worden.
Naar keuze kunnen plots van objectscores gemaakt worden zonder of met nadere aanduidingen. Vraagt men een "unlabeled" plot, dan wordt een afbeelding gemaakt waarin ieder object voorgesteld wordt door het cijfer 1. Vallen twee objecten samen, dan worden zij gezamenlijk voorgesteld door het cijfer 2. Grotere groepen identieke objecten worden evenzo voorgesteld door het cijfer dat de groepsgrootte aangeeft.
Daarnaast is het mogelijk een "labeled" plot van objectscores te laten maken. Men moet dan opgeven welke variabele als label gebruikt moet worden. In zo'n plot wordt ieder object voorgesteld met het cijfer van de categorie waartoe het betreffende object behoort. Een 1 in de afbeelding betekent dus dat op die plaats een object uit categorie 1 staat. Samenvallende objecten worden aangegeven met het plus-teken. Onder de plot is te vinden tot welke categorieën de objecten behoren die met zo'n teken worden bedoeld.
Op de derde plaats is het mogelijk objectscores af te beelden gelabeled met een passieve variabele. Een passieve variabele doet op geen enkele manier mee in de berekeningen van de oplossing. Hij wordt alleen gebruikt om de objectscores in een plot te benoemen.
Plot of category quantifications: Net zoals bij de objectscores is het bij de categoriekwantificaties mogelijk deze in een plot af te beelden. Ook hier geldt dat een dergelijke plot slechts voor de eerste twee dimensies gemaakt kan worden.

Plots met categoriekwantificaties zijn altijd gelabeled. Naar keuze kan men de originele categorieaanduidingen als label gebruiken of het volgnummer van de variabele waartoe elke afzonderlijke categorie behoort. Beeldt men alle categoriekwantificaties in dezelfde plot af, dan betekent in het eerste geval bijvoorbeeld het cijfer 2, dat het om de tweede categorie van een variabele gaat. Op grond van de eerder genoemde tabel met categorie - kwantificaties moet men zelf nog opzoeken welke variabele bedoeld wordt. In het tweede geval betekent het cijfer 2 dat het om een categorie van de tweede variabele in de analyse gaat. Welke categorie bedoeld wordt, moet men eveneens zelf opzoeken in de genoemde tabel.
Men kan er ook voor kiezen van iedere variabele een afzonderlijke plot te maken. In dat geval worden de afgebeelde categoriekwantificaties altijd gelabeled met hun categorie - nummer.

De beschreven informatiemogelijkheden van het programma HOMALS zullen lang niet altijd allemaal even zinvol zijn. Een onderzoeker die geïnteresseerd is in het onder - scheiden van verschillende groepen objecten, doet er verstandig aan vooral plots van objectscores op te vragen. Gaat de aandacht echter vooral uit naar samenhang tussen categorieën van variabelen, dan zijn plots van categoriekwantificaties het meest relevant. In de volgende paragraaf wordt uitgebreid een voorbeeld besproken van een dataset die mede met behulp van HOMALS is geanalyseerd. Het betreft woningkenmerken van 151 respondenten. Dit voorbeeld is overgenomen uit Van de Geer (1985).

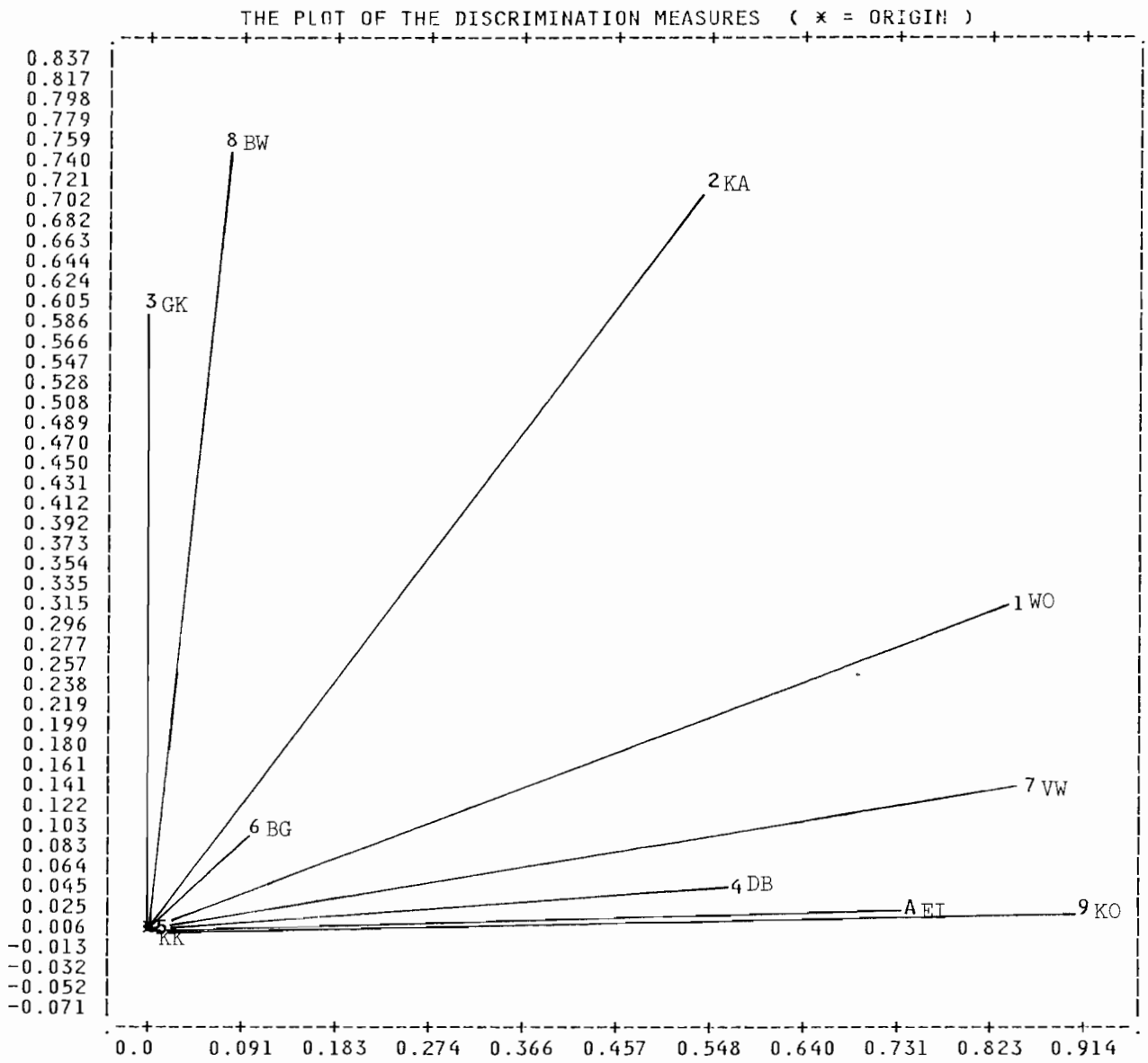
5. Voorbeeld: woningkenmerken

Beschrijving van de data

De analyse die hier als voorbeeld wordt gepresenteerd, is uitgevoerd op gegevens over woningkenmerken. 151 respondenten hebben vragen beantwoord over hun woon - ruimte. De respondenten zijn afkomstig uit verschillende wijken in Rotterdam en het betreffende onderzoek is in 1975 uitgevoerd door het Economisch Geografisch Instituut van de Erasmus Universiteit te Rotterdam. De analyse heeft betrekking op onderstaande variabelen en categorieën.

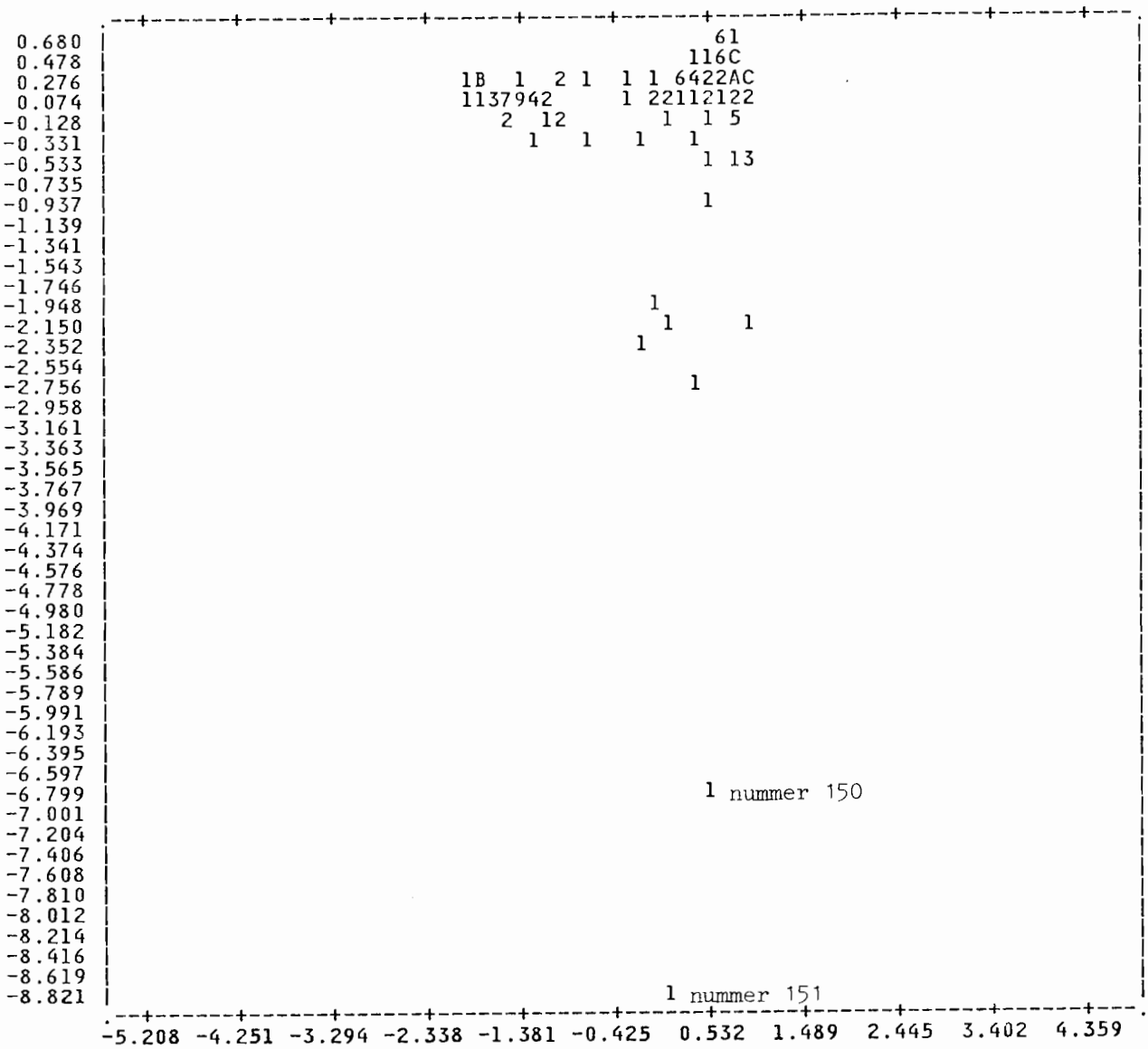
1	Woning (WO)	1 = kamer
	Omschrijving die het	2 = etage
	best bij uw woning past	3 = portiekwoning
		4 = (galerij)flat
		5 = eengezinswoning
		6 = anders
		7 = geen antwoord

2	Kamers (KA) Aantal kamers (inclusief keuken, exclusief toilet en badkamer).	1 = 1 kamer 2 = 2 kamers 3 = 3 kamers 4 = 4 kamers 5 = 5 kamers 6 = 6 kamers of meer 7 = geen antwoord
3	Gemeenschappelijke kamers (GK) Aantal woonkamers gemeenschappelijk met personen die niet tot de eigen huishouding behoren.	1 = geen gemeenschappelijke kamers 2 = 1 gemeenschappelijke kamer 3 = 2 gemeenschappelijke kamers 4 = 3 gemeenschappelijke kamers 5 = 4 gemeenschappelijke kamers 6 = geen antwoord
4	Douche/bad (DB) Heeft u een douche of bad?	1 = ja 2 = nee
5	Keuken (KK) Heeft u een keuken?	1 = ja 2 = nee
6	Berging (BG) Heeft u een aparte berging?	1 = ja 2 = nee
7	Verwarming (VW) Hoe wordt de woning verwarmd?	1 = centrale verwarming 2 = kachel met uitgang via een schoorsteen of de gevel 3 = anders (geen schoorsteen, geen centrale verwarming)
8	Bewoner (BW) Wat voor type bewoner bent u?	1 = huurder 2 = onderhuurder 3 = eigenaar 4 = geen antwoord
9	Kosten (KO) Hoeveel bedragen uw woonlasten per maand?	1 = *f* 100,- of minder 2 = *f* 101,- - *f* 200,- 3 = *f* 201,- - *f* 300,- 4 = meer dan *f* 300,-
10	Eigenaar (EI) In geval het een huurwoning betreft: wie is de eigenaar van de woning?	1 = particulier 2 = onroerend-goed maatschappij 3 = woningbouwvereniging 4 = overheid 5 = anders 6 = onbekend



De plot van discriminatiewaarden, 151 respondenten

In de plot zien we een groep variabelen die een hoge discriminatiewaarde hebben op de eerste (horizontale) dimensie. Het gaat om de variabelen "woning", "verwarming", "douche/bad", "kosten" en "eigenaar". De variabele "kamers" is zowel op de eerste als op de tweede dimensie belangrijk. De variabelen "bewoner", "gemeenschappelijke ka - mers" en "kamers" bepalen de tweede (verticale) dimensie. De variabelen "berging" en "keuken" lijken van geen enkele belang te zijn, aangezien zij dicht bij de oorsprong in de plot liggen. Om te ontdekken hoe de verschillende categorieën van de variabelen sa - menhangen, moeten we de plot van de categoriekwantificaties bekijken. Eerst zullen we echter de plot van de objectscores inspecteren om na te gaan hoe de configuratie van respondenten eruit ziet.



De plot van objectscores, 151 respondenten

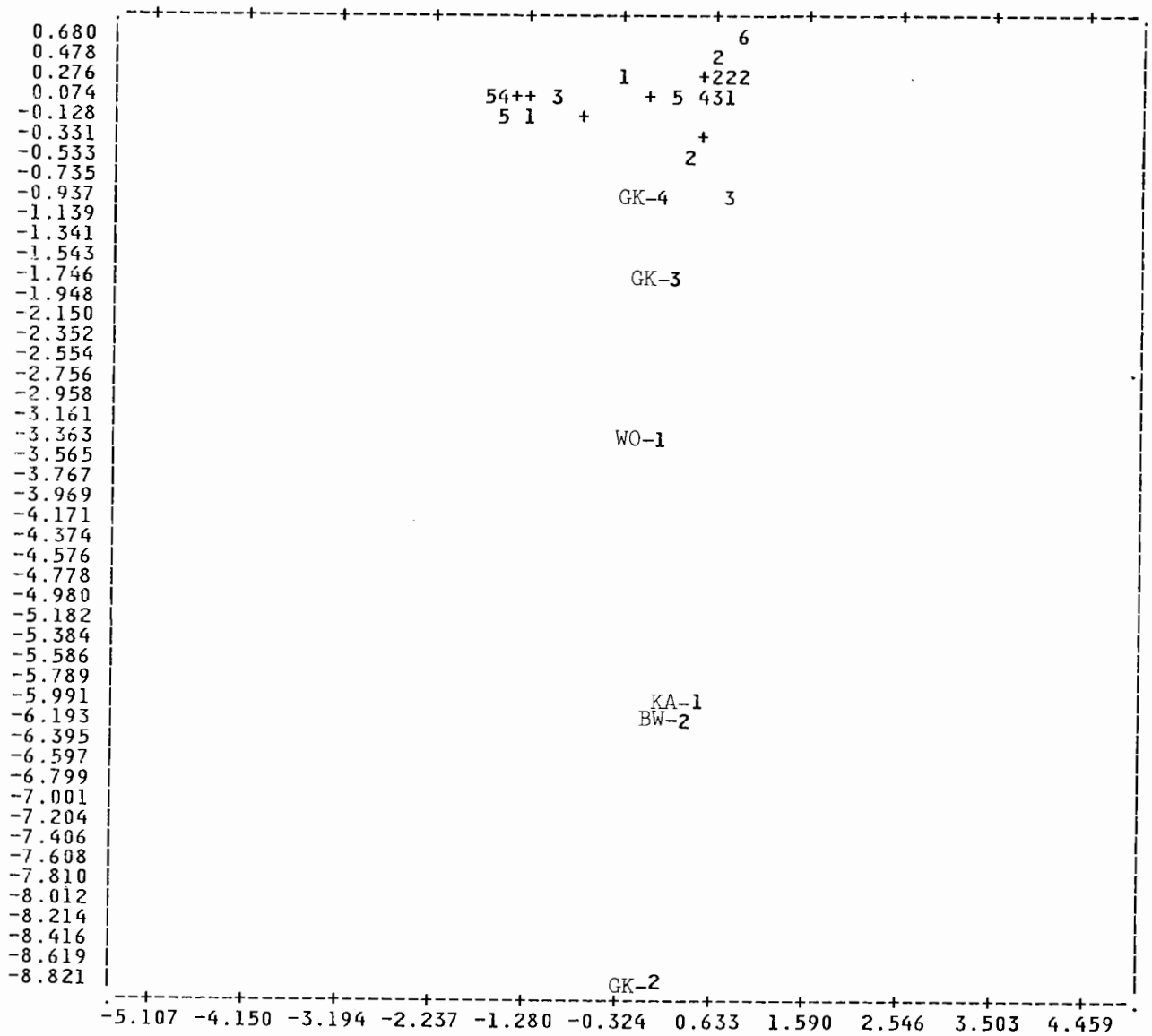
In de plot van objectscores springen de twee uitbijters aan de onderkant het meest in het oog. We kunnen hen identificeren door in de tabel met objectscores individuen te zoeken met een hoge negatieve score op de tweede dimensie. De twee uitbijters krijgen de nummers 150 en 151. We bekijken vervolgens de ruwe data om na te gaan welke combinatie van kenmerken deze individuen zo uniek maakt.

De gegevens van de uitbijters zijn:

	WO	KA	GK	DB	KK	BG	VW	BW	KO	EI
150	1^4	1^3	1	1	1	2	3	2^3	1	6
151	3	1^3	2^1	1	1	2	1	2^3	2	0

De bovenindexen in de profielen zijn de marginale frequenties van de betreffende cate - gorieën. We zien dat zowel respondent 150 als 151 KA-1 (één kamer) hebben en BW-2 (onderhuurder) hetgeen een unieke combinatie is. Bovendien heeft respondent 151 als enige score GK-2 (één gemeenschappelijke kamer).

Nu we weten waarom deze twee respondenten uitbijters zijn, is het interssant om de data opnieuw te analyseren zonder deze personen, zodat we beter zicht krijgen op de rest van de configuratie. Eerst bekijken we echter nog de plot van categoriekwanti - ficaties.



De plot van categoriekwantificaties, 151 respondenten, gelabeled met oorspronkelijke categorienummers

Voor elke afzonderlijke dimensie is de kwantificatie van een categorie het gemiddelde van de objectscores van de personen die tot de betreffende categorie behoren. Omdat individu 151 een unieke score heeft op "gemeenschappelijke kamers", namelijk categorie 2, zijn de coördinaten van GK-2 gelijk aan de objectscores van individu 151. We vinden GK-2 onderaan in de plot van de categoriekwantificaties.

Categorie BW-2 is het gemiddelde van de objectscores van respondenten 62, 150 en 151. Deze drie personen liggen aan de buitenkant in de plot van de objectscores en dus ligt deze categorie ook aan de buitenkant.

Als we respondent 151 uit de analyse verwijderen, verwachten we dat de discriminatiewaarde van de derde variabele (GK) op de tweede dimensie zal verminderen. Aangezien echter de categorieën GK-3 en GK-4 ook nogal aan de buitenkant liggen, zal hun invloed in de volgende analyse toenemen zodat de variabele "gemeenschappelijke kamers" op de tweede dimensie belangrijk kan blijven.

De invloed van variabele 2 (kamers) en 8 (bewoner) kan een stuk minder worden als we de twee uitbijters verwijderen uit de analyse. Omdat er vier respondenten zijn met categorie WO-1, van wie er slechts één verwijderd wordt, kan het zijn dat de variabele "woning" belangrijker wordt op de tweede dimensie.

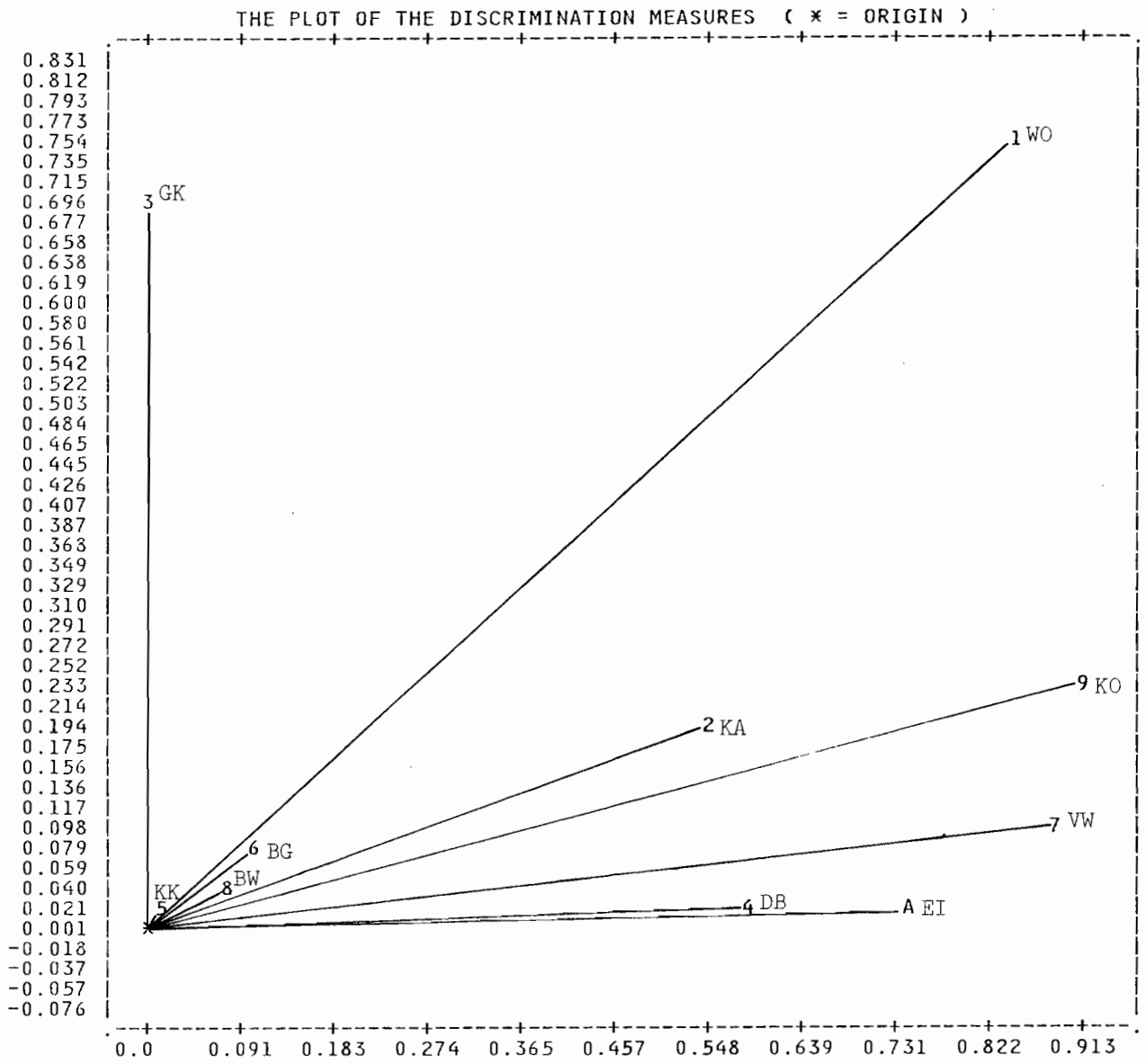
De plot van discriminatiewaarden, 149 respondenten (blz. 18)

Dezelfde variabelen als in de vorige analyse hebben hoge ladingen op de eerste dimensie. Zelfs de volgorde van de discriminatiewaarden op de eerste dimensie is hetzelfde als in de vorige analyse. De tweede dimensie wordt bepaald door variabele 1 (woning) en 3 (gemeenschappelijke kamers). De invloed van variabele 2 (kamers) en 8 (bewoner) is verdwenen.

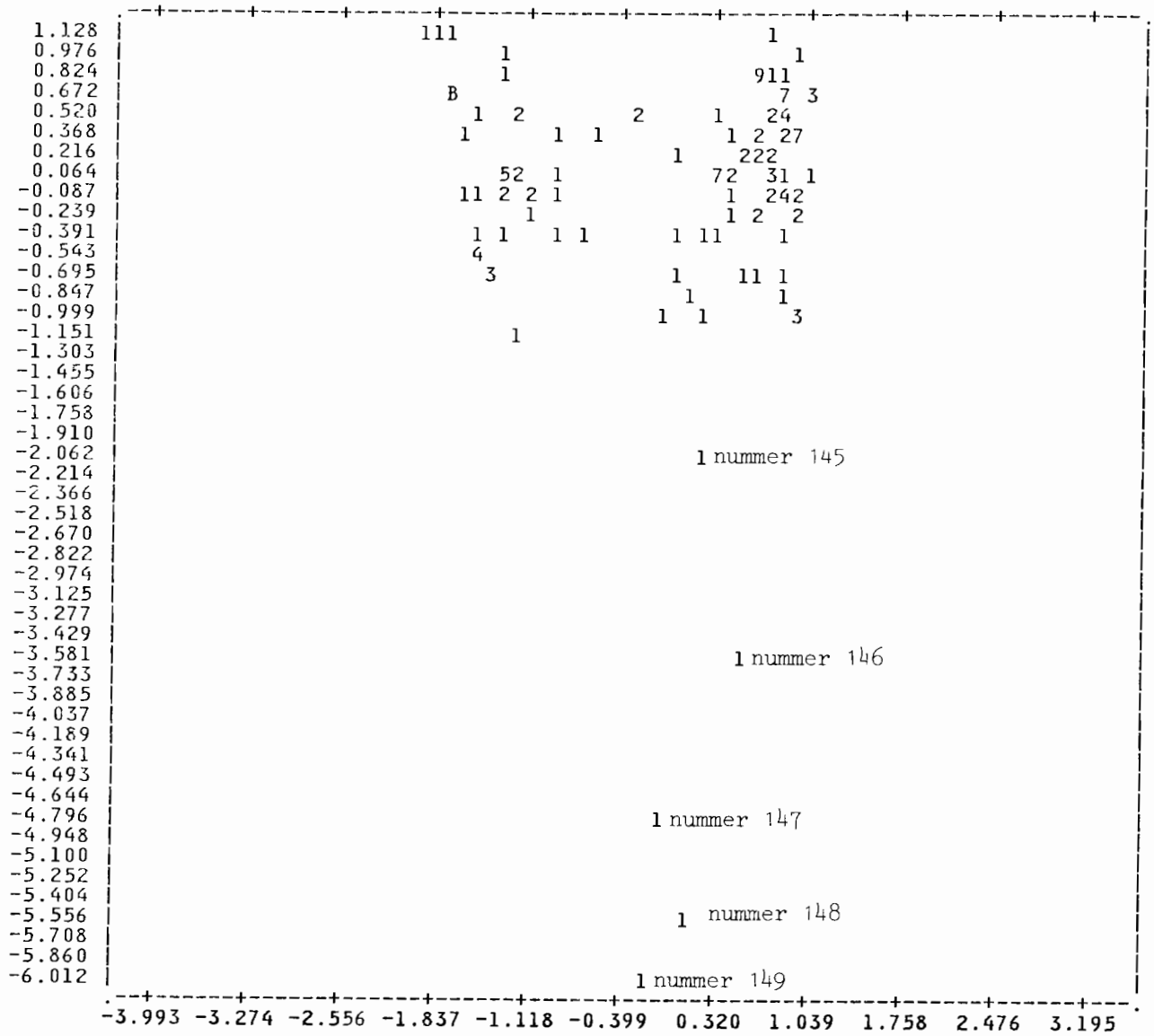
De plot van objectscores, 149 respondenten (blz. 19)

De plot van objectscores bestaat wederom uit één groot cluster en een paar uitbijters. Voordat we het cluster interpreteren, bekijken we eerst de extreme individuen. Hun oorspronkelijke kenmerken zijn:

	WO	KA	GK	DB	KK	BG	VW	BW	KO	EI
145	3	5	4^2	1	1	1	2	1	2	1
146	2	5	3^3	2	1	1	3	1	2	4
147	1^3	5	3^3	1	1	2	1	1	2	4
148	1^3	3	3^3	1	1	2	1	1	2	4
149	1^3	4	4^2	1	1	2	1	1	2	4

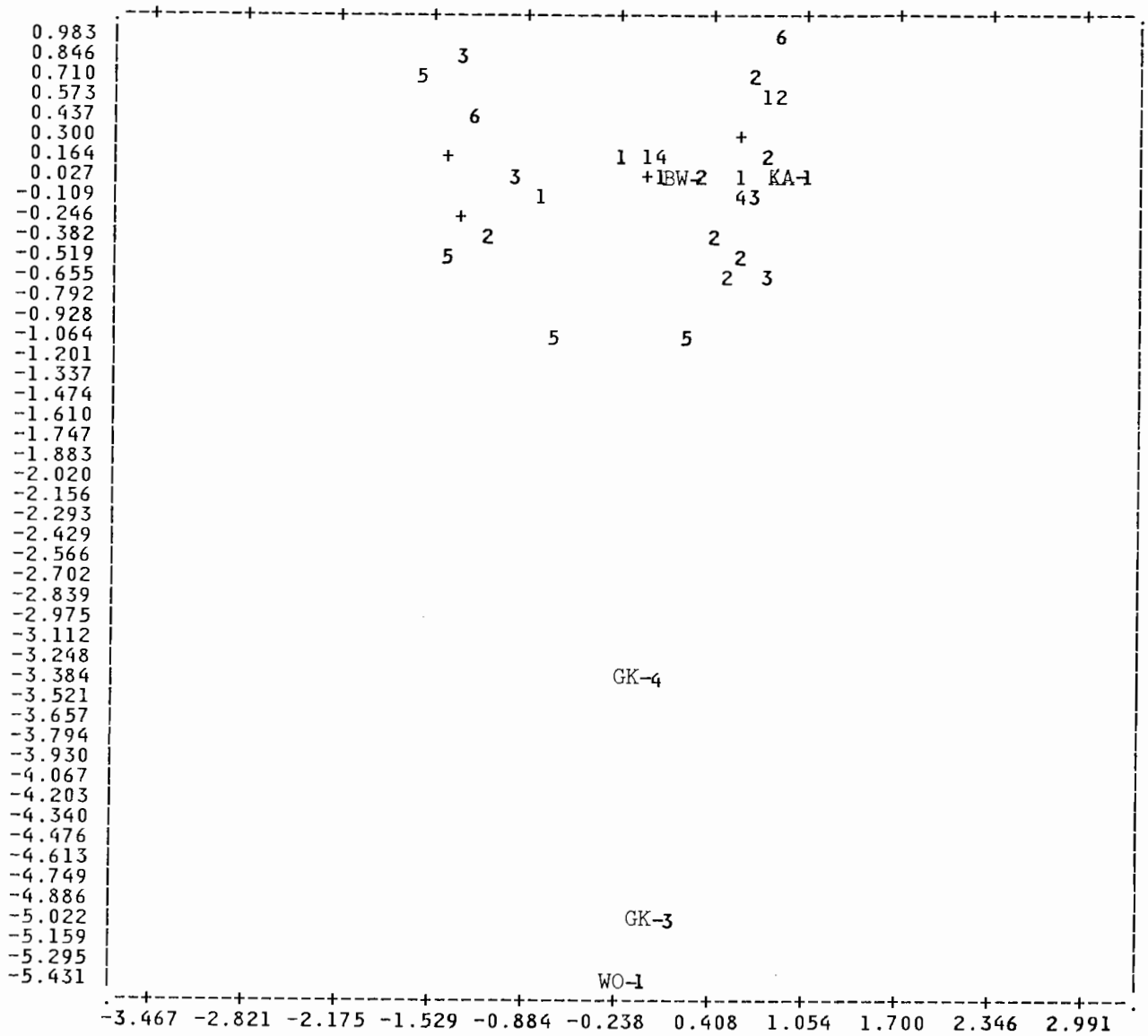

THE PLOT OF THE DISCRIMINATION MEASURES (* = ORIGIN)
0.831
0.812
0.793
0.773
0.754
0.735
0.715
0.696
0.677
0.658
0.638
0.619
0.600
0.580
0.561
0.542
0.522
0.503
0.484
0.465
0.445
0.426
0.407
0.387
0.368
0.349
0.329
0.310
0.291
0.272
0.252
0.233
0.214
0.194
0.175
0.156
0.136
0.117
0.098
0.079
0.059
0.040
0.021
0.001
-0.018
-0.037
-0.057
-0.076
0.0
0.091
0.183
0.274
0.365
0.457
0.548
0.639
0.731
0.822
0.913
1 WO
2 KA
3 GK
4 DB
5 KK
6 BG
7 VW
8 BW
9 KO
A EI

De laatste drie personen zijn de enigen die in een enkele kamer wonen (WO-1). Verder worden deze personen gekenmerkt door het aantal gemeenschappelijke kamers (GK-3 of GK-4). Waarschijnlijk worden de respondenten met de nummers 145 en 146 weggetrokken van het cluster doordat zij kenmerken delen met de kamerbewoners (GK-3 en GK-4) en niet doordat zij zo'n uniek profiel bezitten.



De plot van categoriekwantificaties, 149 respondenten, gelabeled met oorspronkelijke categorienummers

Onder in de plot vinden we drie categorieën, WO-1, GK-3 en GK-4, die overeenkomen met de posities van de uitbijters in de plot van objectscores. Vergeleken met de vorige analyse zijn de categorieën KA-1 en BW-2 in het cluster opgegaan.

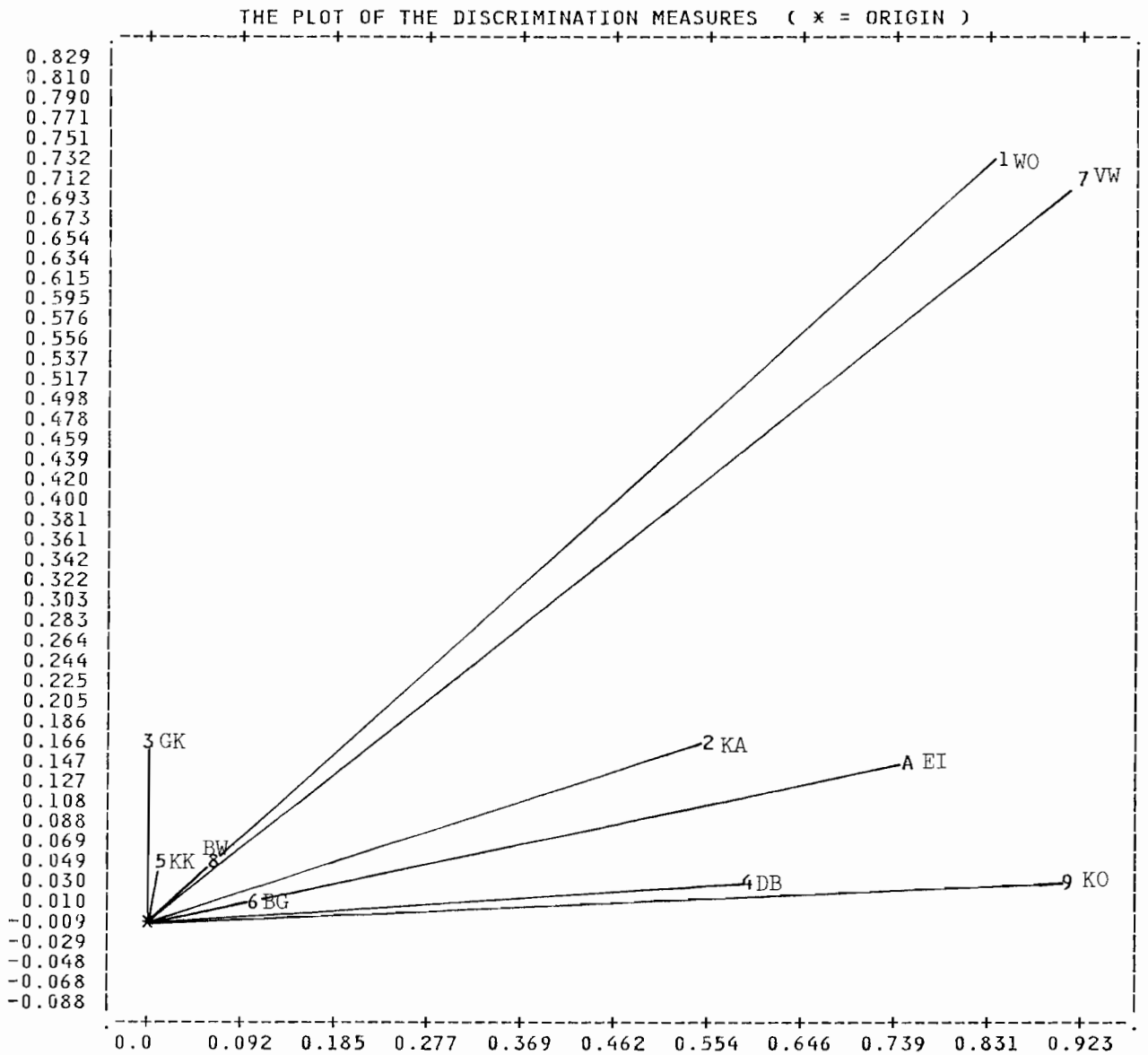


De volgende stap

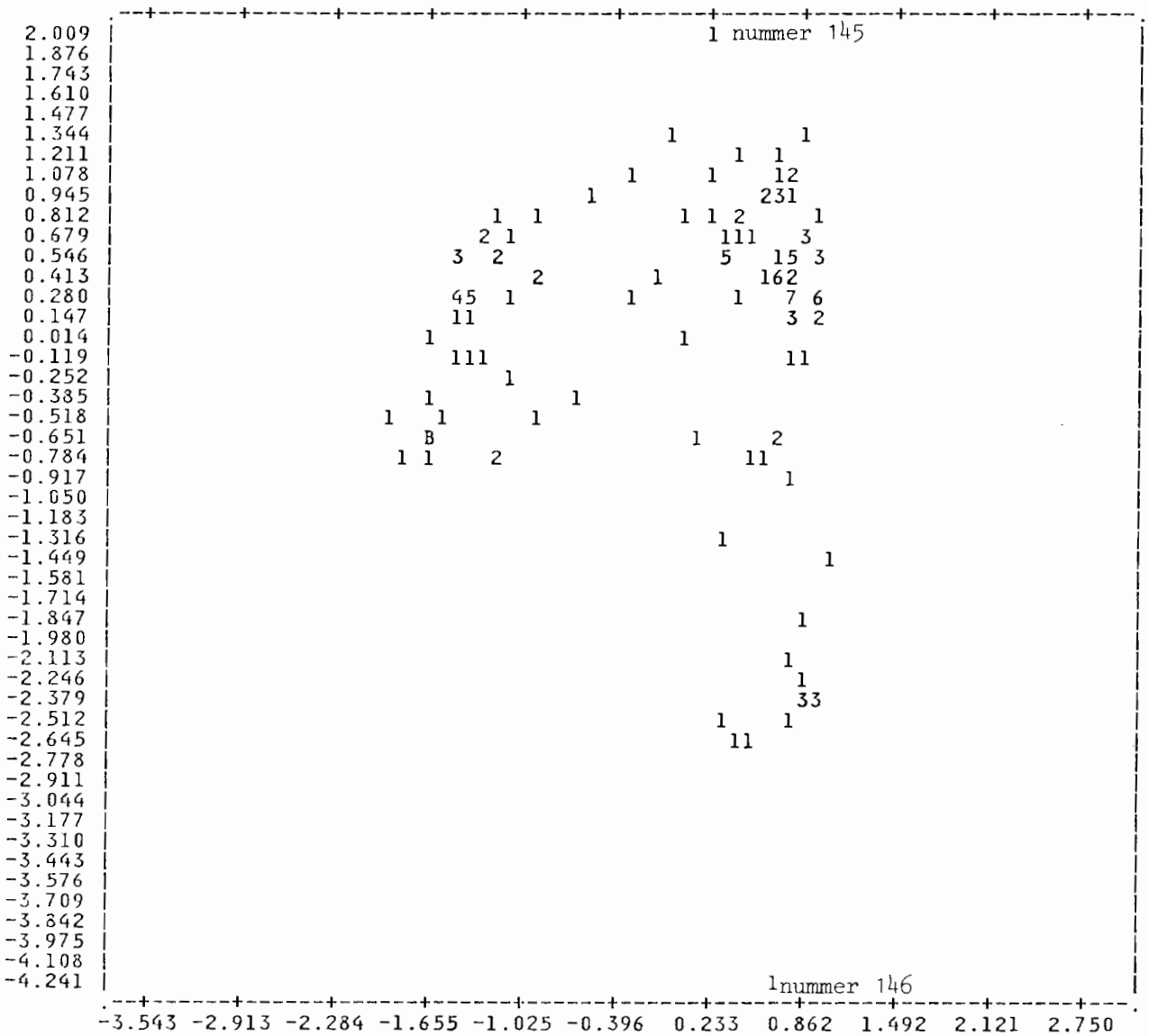
We verwijderen opnieuw een aantal uitbijters voordat we de volgende analyse uitvoe - ren. Het lijkt aannemelijk de kamerbewoners als een aparte categorie te beschouwen en hen daarom uit de data te verwijderen.

De plot van discriminatiewaarden, 146 respondenten

Opnieuw hebben de variabelen "woning", "verwarming", "kosten", "eigenaar", "dou - che/bad" en "kamers" een hoge discriminatiewaarde op de eerste dimensie. De tweede



dimensie wordt bepaald door "woning" en "verwarming". De invloed van variabele 3 "gemeenschappelijke kamers" op de tweede dimensie is grotendeels verdwenen, vergeleken met de analyse met 149 respondenten. De variabele "verwarming" heeft slechts drie categorieën. De variabele "gemeenschappelijke kamers" heeft zes categorieën. In dit voorbeeld zien we dat een variabele met minder categorieën niet persé onbelangrijker is dan een variabele met meer categorieën.
We bekijken nu de plots van objectscores en categoriekwantificaties om na te gaan welke categorieën met elkaar samenhangen.



De plots van objectscores (blz. 22) en van categoriekwantificaties (blz. 24)

Aan de linkerkant van de plot van categoriekwantificaties vinden we een groep samen - hangende categoriekwantificaties. Dit zijn grote eengezinswoningen of flats (WO-5 en WO-4), centrale verwarming (VW-1), douche of bad (DB-1), hoge kosten (KO-3 en KO-4) en veel kamers (KA-5 en KA-6). Deze huizen zijn het eigendom van de be - woners (EI-2) of van een woningbouwvereniging (EI-3). In vergelijking zijn dit de dure huizen. De rijkere mensen liggen dus links in de plot met objectscores.

In de rechterbovenhoek van de plot van categoriekwantificaties vinden we categorieën als portiekwoning (WO-3), kachel met uigang via een schoorsteen (VW-2), twee tot vier kamers (KA-2, KA-3 en KA-4), geen douche of bad (DB-2), lage woonlasten (KO-1 en KO-2) en een woning die het eigendom is van een particulier (EI-1) of van de overheid (EI-4). Dit zijn de goedkopere woningen en dus vinden we de relatief armere respondenten in het rechtergedeelte van de plot van objectscores.

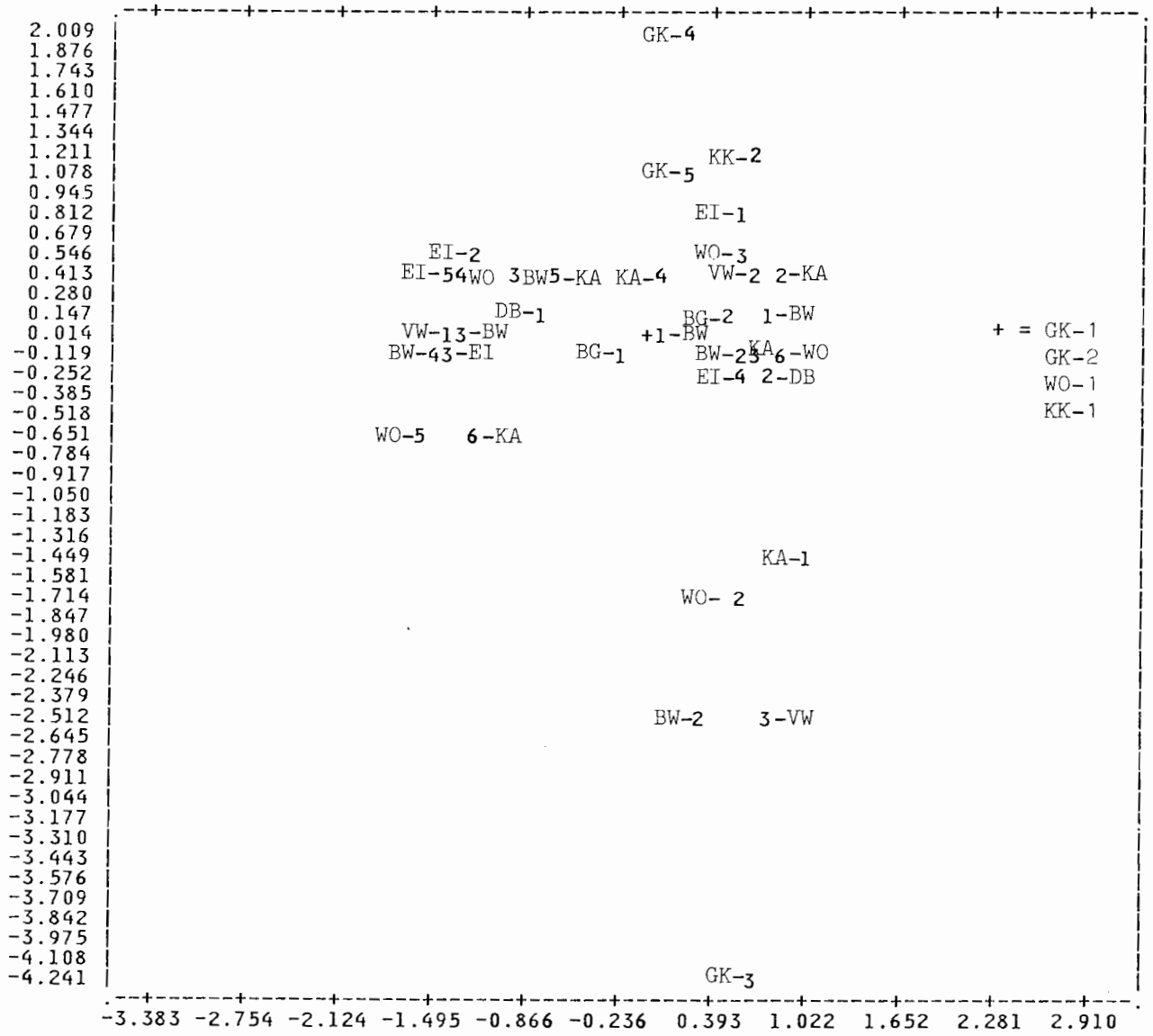
Aan de onderkant van de plot van categoriekwantificaties vinden we bijzondere ken - merken, namelijk etages (WO-2) waarvan de lager in de plot gelegenen geen fatsoen - lijke verwarming hebben (VW-3, geen schoorsteen, geen centrale verwarming). Waar - schijnlijk wonen er alleenstaanden of jonge stellen in dit type woningen. Deze groep respondenten hoort eerder bij de "arme" groep dan bij de "rijke". Zij hebben dezelfde positieve scores op de eerste dimensie.

Bekijken we de categoriekwantificaties van variabele 1 (woning) en variabele 3 (ge - meenschappelijke kamers) op bladzijde 24, dan zien we dat de categoriekwantificaties van variabele 3 meer spreiding vertonen dan die van variabele 1. Over het algemeen komt meer spreiding overeen met hogere discriminatiewaarden, maar in dit geval zijn de respondenten zeer ongelijk verdeeld over de categorieën van de variabele "gemeen - schappelijke kamers". In feite vallen 140 personen in categorie GK-1. Deze categorie vinden we dan ook vlak bij de oorsprong in de plot.

Gelabelde plots van objectscores, 146 respondenten

De oorspronkelijke categorienummers van de variabelen met de hoogste discrimi - natiewaarden zijn gebruikt om de plot van objectscores van labels te voorzien. Dit zijn de variabelen "woning" (blz. 25), "verwarming" (blz. 26) en "kosten" (blz.27).

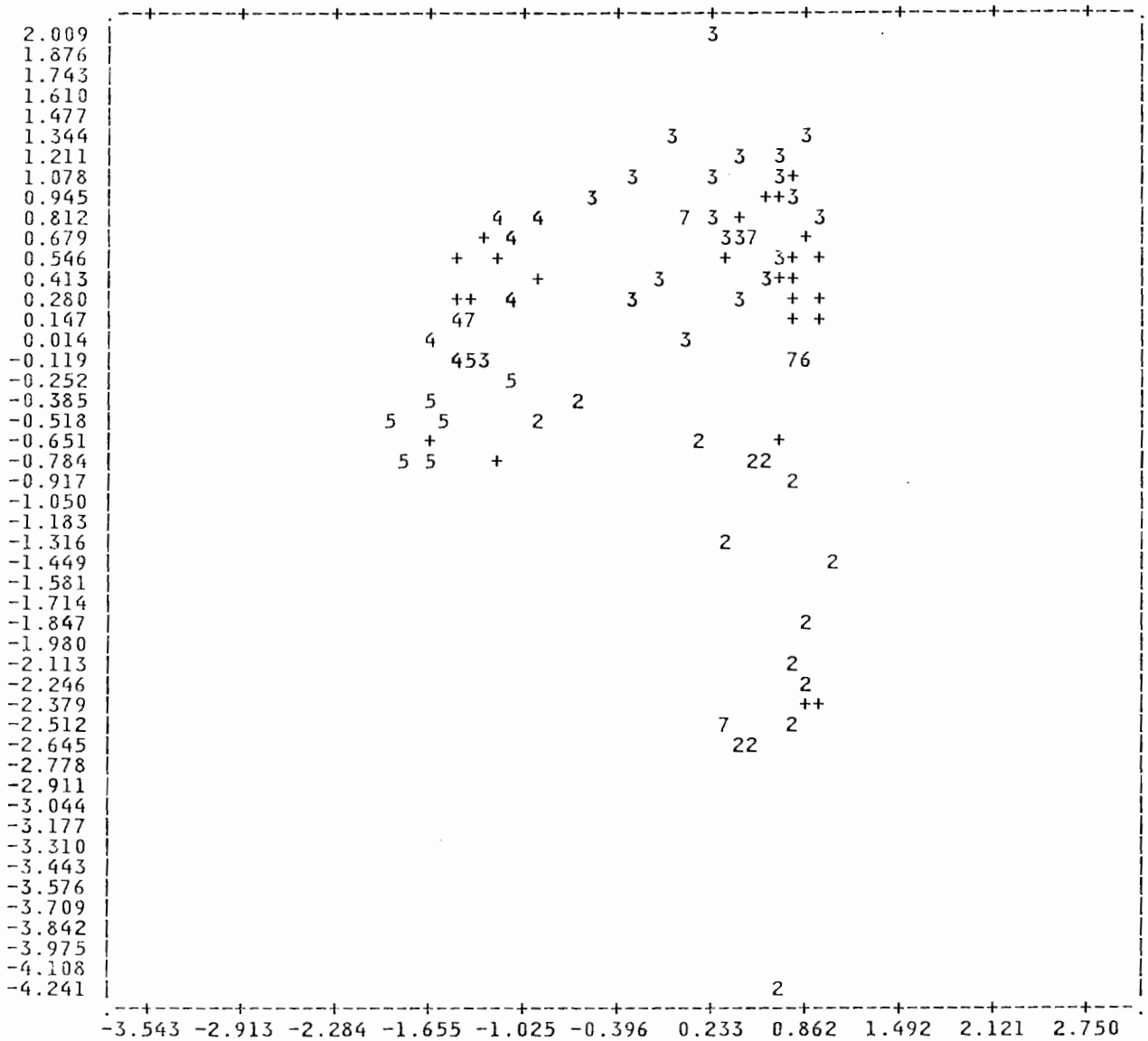
Op bladzijde 25 vinden we zeer homogene groepen voor variabele WO (woning). Het overzicht van samenvallende punten onder de plot geeft aan dat al die punten uit één categoriesoort bestaan en dit wijst ook op homogeniteit. Alleen de respondenten zonder score op "woning" (WO-7) zijn over de plot verdeeld. We kunnen de puntenwolk inter - preteren als een ééndimensionele schaal, lopend van 2 via 3 en 4 naar 5. Dat betekent van kleine woningen naar grote huizen.



Variabele VW (verwarming) verdeelt de respondenten eveneens in homogene groepen (blz. 26). Deze komen goed overeen met de groepen die in de plot van objectscores te vinden zijn. Aan de linkerkant vinden we de centraal verwarmde huizen (VW-1), aan de rechterkant de huizen met kachels met een uitgang via een schoorsteen of de gevel (VW-2) en onderaan de andere vormen van verwarming (VW-3).

De laatste variabele die we gebruikt hebben om de plot van objectscores te labelen, is variabele KO (kosten). Deze plot staat op blz. 27. De variabele verdeelt de respondenten op de eerste dimensie in twee groepen. Links staan de respondenten die meer dan *f* 200,- voor hun huisvesting betalen, rechts de respondenten met lagere woonlasten.

OBJECT SCORES, LABELED BY VARIABLE : WO

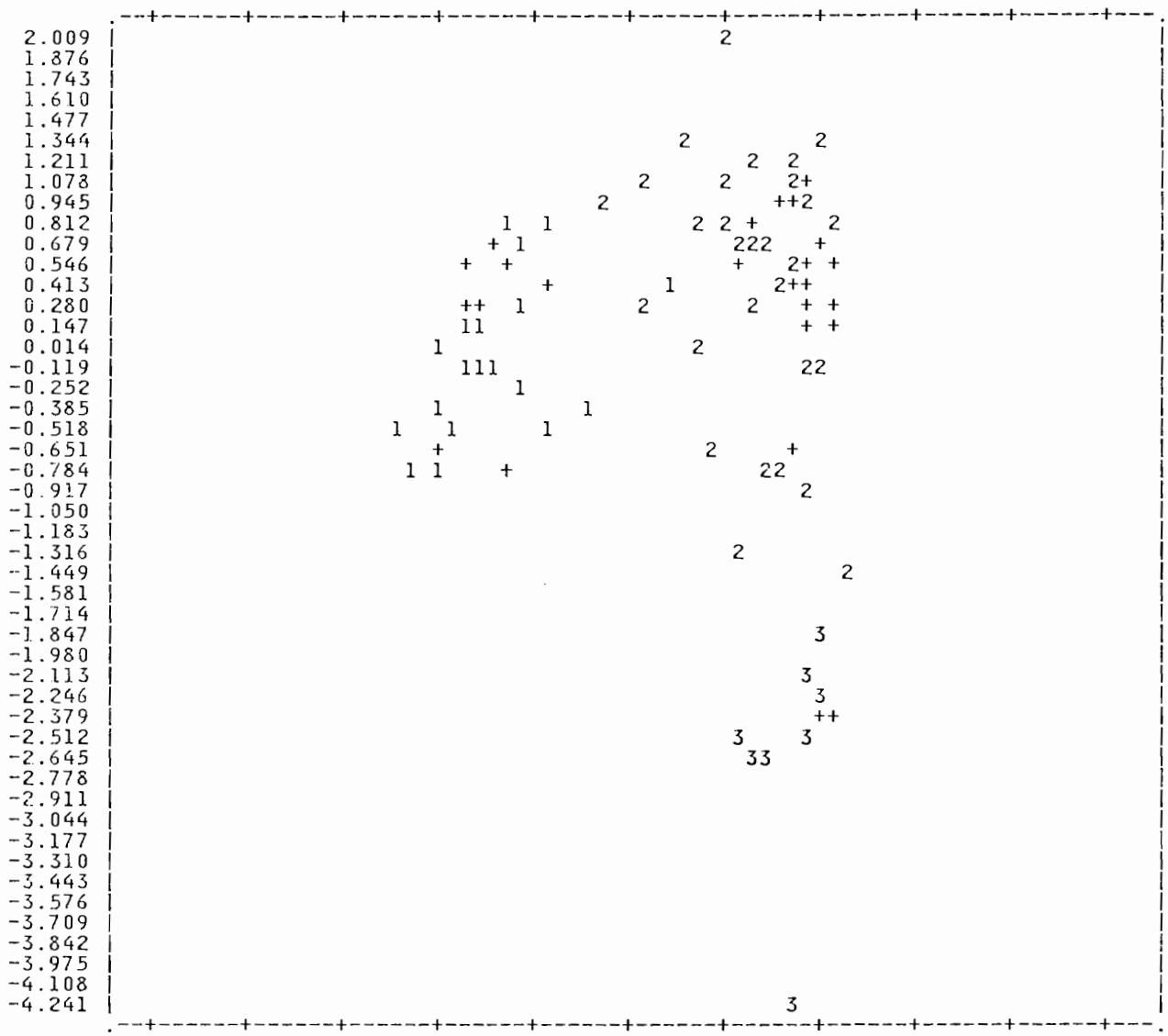


22 DEC 86 HOUSING CHARACTERISTICS
12:54:27 146 RESPONDENTEN

SUMMARY OF ALL CELLS (X,Y), MARKED : + IN THE PLOT, CONTAINING MORE THAN 1 POINT IDENTIFICATION

X	Y	NUMBER OF POINTS	POINT-IDENTIFICATION
0.773	1.078	2	33
0.593	0.945	2	33
0.683	0.945	3	333
0.413	0.812	2	33
-1.295	0.679	2	44
0.862	0.679	3	333
-1.475	0.546	3	444
-1.205	0.546	2	44
0.323	0.546	5	33333
0.773	0.546	5	33333
0.952	0.546	3	333
-0.936	0.413	2	34
0.683	0.413	6	333333
0.773	0.413	2	33
-1.475	0.280	4	4444
-1.385	0.280	5	44444
0.773	0.280	7	3333333
0.952	0.280	6	333333
0.773	0.147	3	333
0.952	0.147	2	33
-1.655	-0.651	11	55555555555
0.683	-0.651	2	22
-1.205	-0.784	2	55
0.862	-2.379	3	222
0.952	-2.379	3	222

OBJECT SCORES, LABELED BY VARIABLE : VW

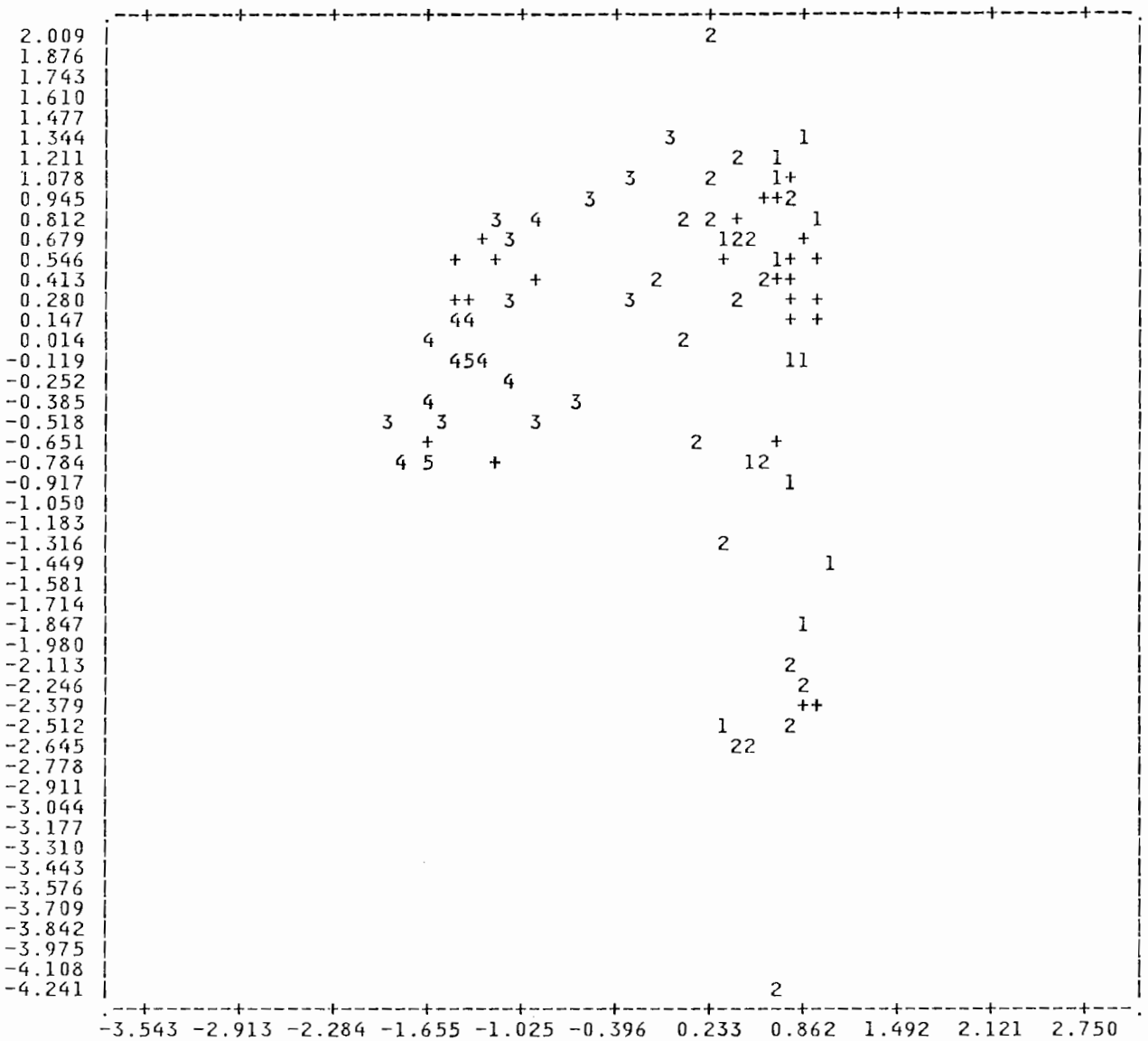


22 DEC 86 HOUSING CHARACTERISTICS
12:54:27 146 RESPONDENTEN

SUMMARY OF ALL CELLS (X,Y), MARKED : + IN THE PLOT, CONTAINING MORE THAN 1 POINT IDENTIFICATION

X	Y	NUMBER OF POINTS	POINT-IDENTIFICATION
0.773	1.078	2	22
0.593	0.945	2	22
0.683	0.945	3	222
0.413	0.812	2	22
-1.295	0.679	2	11
0.862	0.679	3	222
-1.475	0.546	3	111
-1.205	0.546	2	11
0.323	0.546	5	22222
0.773	0.546	5	22222
0.952	0.546	3	222
-0.936	0.413	2	11
0.683	0.413	6	222222
0.773	0.413	2	22
-1.475	0.280	4	1111
-1.385	0.280	5	11111
0.773	0.280	7	2222222
0.952	0.280	6	222222
0.773	0.147	3	222
0.952	0.147	2	22
-1.655	-0.651	11	11111111111
0.683	-0.651	2	22
-1.205	-0.784	2	11
0.862	-2.379	3	333
0.952	-2.379	3	333

OBJECT SCORES, LABELED BY VARIABLE : KO



22 DEC 86 HOUSING CHARACTERISTICS
12:54:27 146 RESPONDENTEN

SUMMARY OF ALL CELLS (X,Y), MARKED : + IN THE PLOT, CONTAINING MORE THAN 1 POINT IDENTIFICATION

X	Y	NUMBER OF POINTS	POINT-IDENTIFICATION
0.773	1.078	2	11
0.593	0.945	2	22
0.683	0.945	3	111
0.413	0.812	2	22
-1.295	0.679	2	44
0.862	0.679	3	222
-1.475	0.546	3	444
-1.205	0.546	2	33
0.323	0.546	5	22222
0.773	0.546	5	22111
0.952	0.546	3	111
-0.936	0.413	2	43
0.683	0.413	6	111111
0.773	0.413	2	22
-1.475	0.280	4	4444
-1.385	0.280	5	44444
0.773	0.280	7	1111111
0.952	0.280	6	111111
0.773	0.147	3	222
0.952	0.147	2	22
-1.655	-0.651	11	44444444444
0.683	-0.651	2	11
-1.205	-0.784	2	44
0.862	-2.379	3	111
0.952	-2.379	3	222

Conclusie

De respondenten die naar de kenmerken van hun woning zijn gevraagd, vormen drie groepen. Deze groepen worden bepaald door het type woning, het soort verwarming en de woonlasten. We zouden dit in feite kunnen opvatten als een ééndimensionele huis - vestingsschaal die van etages via portiekwoningen en galerijflats naar eengezins - woningen loopt. Deze classificatie komt overeen met goedkope versus dure woningen, niet versus wel centraal verwarmde huizen, weinig kamers versus veel kamers, geen douche of bad versus wel douche of bad, particulier of overheidsbezit versus woning - bouwverenigingen of onroerend-goed maatschappijen als eigenaars. De variabelen "keuken", "berging" en "bewoners" lijken weinig belangrijk te zijn voor deze schaal.
Proberen we de kamerbewoners in deze schaal te passen, dan moeten we hen aan het uiteinde van de schaal plaatsen, voorbij de mensen die etages bewonen. Dit valt af te lei - den uit de plot van objectscores in de eerste analyse. De variabele "gemeenschappelijke kamers" is zeer belangrijk voor kamerbewoners, aangezien er in de eerste analyse slechts zeven personen zijn die kamers delen en drie van de vier kamer bewoners meer dan één kamer delen.
De huidige analyse laat zien dat de structuur van de data overtuigender naar voren komt wanneer de vijf extreme respondenten verwijderd worden die in weinig frequente cate - gorieën scoren (categorie 1 van de variabelen "woning" en "kamers" en categorie 2 van de variabele "bewoner"). Een andere strategie zou zijn geweest deze categorieën als "missing" te beschouwen. De resultaten van zo'n analyse, met alle 151 respondenten, blijken in zeer grote mate overeen te komen met de resultaten die hierboven zijn ge - presenteerd. Daarom zal die analyse hier niet aan worden toegevoegd.

AANGERADEN LITERATUUR

De theorie waarop HOMALS is gebaseerd, wordt beschreven in:

Gifi, A. (1981). *Nonlinear multivariate analysis*. Leiden: Department of Data Theory.

Een uitgebreidere beschrijving van het programma en vele suggesties voor wie zich wil verdiepen, biedt:

Van de Geer, J.P. (1985). *HOMALS, Research Report UG-85-02*. Leiden: Department of Data Theory.

Toepassingen voor HOMALS zijn o.a. te vinden in:

De Leeuw, J. & Kreft, I. (1985). *Over definitie en kwantificatie van schoolloopbanen.* Leiden: Department of Data Theory (RR-85-20).

Van den Berg, G.M. (1986). Optimalisering van procesgerichte diagnostiek. In: W.J. van der Linden & J.M. Wijnstra (red.), *Ontwikkelingen in de methodologie van het onderwijsonderzoek*. Lisse: Swets & Zeitlinger B.V., pp. 141-151.

Specifiek over de behandeling van missing data gaat:

Meulman, J. (1982). *Homogeneity analysis of incomplete data*. Leiden: DSWO Press.

De beschrijving van het verwante programma PRIMALS is te vinden in:

Van de Geer, J.P. & Meulman, J. (1985). *PRIMALS, Research Report UG-85-01*. Leiden: Department of Data Theory.