NORSK REGNESENTRAL
NORWEGIAN COMPUTING CENTER

N R

FACIES-INNDELING BASERT PÅ LOGG- OG KJERNEDATA FRA BRØNN 31/2-3 VED HJELP AV KLUSTERANALYSE

Av

Jon Helgeland og Erik Mohn

NORSK REGNESENTRAL / NORWEGIAN COMPUTING CENTER

Forskningsvn. 1 B, Blindern, Oslo 3, Norway, telefon (02) 46 69 30

FACIES-INNDELING BASERT PÅ LOGG- OG KJERNEDATA FRA BRØNN 31/2-3 VED HJELP AV KLUSTERANALYSE

Av

Jon Helgeland og Erik Mohn

Prosjekt 388002

Juni 1983

FORORD

I denne rapporten fremlegges resultater av faciesinndeling for brønn 31/2-3 basert på logg- og kjernedata. Rapporten beskriver også de klustringsmetoder som er brukt samt fremgangsmåter for å tolke resultatene fra en klusteranalyse.

Detalj-resultater er gitt i de regnemaskinutskrifter som er vedlegg til denne rapport.

Oslo, 3. juni 1983

Jon Helgeland Erik Mohn

INNHOLDSFORTEGNELSE

1.	Klusteranalyse - et hjelpemiddel ved gruppering	s.	1
2.	Hvor forkjellige er klustrene	"	2
3.	Samsvar mellom logg- og kjernekluster	"	3
4.	Datagrunnlag	"	3
5.	Resultater	"	4
	5.1 Kjerneprøver	"	5
	5.2 Loggprøver	"	15
	5.3 Sammenligning	"	27
Appendiks 1		"	32
Appendiks 2		"	34
Appendiks 3		"	37
Appendiks 4		"	38

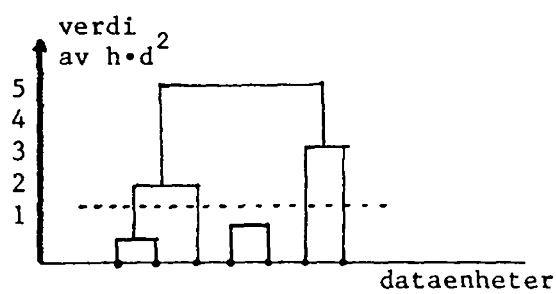
1. KLUSTERANALYSE - ET HJELPEMIDDEL VED GRUPPERING

Klusteranalyse er en samlebetegnelse på en hel rekke forskjellige metoder til å gruppere dataenheter som i en eller annen forstand "ligner" hverandre. I vårt tilfelle er dataenhetene "dyp". Fra et visst antall dyp - la oss si n - foreligger observasjoner av et antall variable - la oss si m stykker. F.eks., fra de n = 185 dypene: 1412.25 m, 1413 m,...,1646.50 m foreligger kjerneprøver hvor man har observert de m = 6 variablene SRT, GRSZ, MICA, CC, CON, BTB. På grunnlag av disse observasjonene skal klusteranalysen gi forslag til grupperinger av dyp. De fremkomne grupper kalles gjerne klustere.

De dyp som tilhører samme kluster, ligner hverandre i en eller annen forstand. Det blir altså nødvendig å definere likhet mellom observasjonssettene tatt fra 2 forskjellige dyp. Dette er gjort ved hjelp av vanlig Euklidsk avstand. Avstandsmålet vil avhenge av de skalaer som variablene registreres i. For å unngå skalaavhengigheten kan vi standardisere variablene. Standardiseringen representerer en bestemt vektlegging av variablene (alle får standardavvik = 1). Ønskes andre vekter, multipliseres de standardiserte variablene med disse. Standardavvikene på de nye variablene blir da lik disse vektene.

Når man har definert et avstandsmål mellom dataenhetene finnes det en rekke klustringskriterier å velge mellom. Vi har valgt en kombinasjon av Wards metode og k-means. Disse kan betraktes som forskjellige algoritmer for å finne en klustring som er optimal i den forstand at den totale variansen innen klustrene er minimal (dvs. kvadratsummen av avstandene fra alle dataenhetene til sine respektive kluster-sentroider).

Wards metode er en såkalt sammenføyningsmetode. Initielt er antall klustre lik n, dvs. antall dataenheter. På hvert trinn slås to klustere sammen til et nytt, nemlig det par for hvilket produktet $h \cdot d^2$ er minst. Her er h det harmoniske gjennomsnitt av antall dataenheter i de to klustrene i paret; d^2 er den kvadrerte Euklidske avstand mellom sentroidene (dvs. middelvektorene). For detaljer: se Appendiks 1. Denne prosedyren fortsettes inntil vi sitter igjen med ett eneste kluster, bestående av alle dataenhetene. Resultatet kan anskueliggjøres ved et dendrogram (se figuren). På den horisontale aksen avsettes dataenhetene (i en passelig rekkefølge). På den vertikale aksen er en skala for sammenføyningskriteriet. Av dendrogrammet kan man avlese hvilke par av klustre som slås sammen på hvert trinn, og for hvilken verdi av $h \cdot d^2$ sammenslåingen skjer. Ønsker man f.eks. c klustre, trekker man en horisontal linje i dendrogrammet som skjærer nøyaktig c



av grenene (på figuren har vi valgt c = 4). Vi kan så avlese hvilke dataenheter som tilhører disse c klustrene. Formen på dendrogrammet kan også brukes til å vurdere hvor mange klustere datamaterialet inneholder.

En svakhet ved sammenføyningsmetodene, slik som Wards, er at to klustere som på et bestemt trinn blir sammenslått, likevel ikke behøver å representere en optimal sammenslåing når man vurderer det ved et senere trinn. Ved hjelp av k-means, som er en såkalt byttemetode, revurderes klustertilhørigheten til alle dataenhetene.

Vi tar utgangspunkt i de c (la oss si) klustrene som vi har bestemt oss for ved Wards metode. Vi beregner de tilsvarende c sentroidene og plasserer hver dataenhet i det kluster som har nærmest sentroide. En del dataenheter vil da kunne bytte kluster. Dernest beregnes sentroidene for de nye c klustrene. Man går så gjennom alle dataenhetene på ny og flytter dataenheter til klustere med nærmeste sentroide. Hver gang man nå foretar en flytting, oppdateres sentroidene. Dette fortsetter man med helt til det ikke skjer flere flyttinger.

2. HVOR FORSKJELLIGE ER KLUSTRENE?

Enhver klustringsmetode vil lede til en partisjon av dataenhetene - selv om det ikke er noen struktur i dataene. Det er derfor viktig å forsøke å karakterisere klustrene ut fra de dataenheter de inneholder, for å se om de virkelig representerer reelle grupperinger i data.

Den enkeleste fremgangsmåten er å regne ut forskjellige deskriptive mål for hver variabel i hvert kluster. Vi har beregnet aritmetrisk gjennomsnitt, varians og standardavvik, samt minimum, maksimum og variasjonsbredde. Kombinert med geologisk/petrofysisk kompetanse vil dette danne grunnlag for en beskrivelse av klustrene og en vurdering av om de har en interessant interpretasjon.

En mer sammensatt fremgangsmåte er følgende (detaljer, se Appendiks 2): la oss bestemme den lineærkombinasjon av variablene som på "best" mulig måte skiller de gitte klustrene fra hverandre. Betydningen av "best" lar vi være at forholdet mellom variansene mellom klustrere og variansen innen klustere blir så stort som mulig. Ved beregningen av denne optimale lineærkombinasjonen får vi også bestemt den nest beste lineærkombinasjon, den tredje beste o.s.v. Disse lineærkombinasjonene blir også ukorrelerte.

Et to-dimensjonalt plott av de to første lineærkombinasjonene vil kunne gi et godt bilde av hvordan dataenhetene i klustrene skiller seg ad. Størrelsene på koeffisientene i lineærkombinasjonen vil fortelle hvilke variable som er de viktigste komponentene i lineærkombinasjonen og kan i visse tilfeller gi en tolkning til disse. I noen tilfeller vil de to første lineærkombinasjonene inneholde mesteparten av variablenes diskriminerende evne. Da vil denne tolkningen kunne være særlig interessant.

3. SAMSVAR MELLOM LOGG- OG KJERNEKLUSTER

En klusteranalyse av brønndata kan gi informasjon om hvor godt loggene kan predikere bestemte interessevariabler som er målt på kjernen, f.eks. porøsitet og permeabilitet. I tilfelle av perfekt prediksjon vil det til hvert loggkluster svare ett og bare ett kjernekluster. En tabell der dypene er kryssklassifisert etter logg- og kjernekluster (slik det gjøres i avsnitt 5.3). viser hvor godt dette idealet er oppfylt, og det er mulig å kvantifisere samsvaret i en slik tabell. Imidlertid sier ikke denne tabellen noe om de innbyrdes avstandene mellom klustrene. Dette betyr at feilaktige prediksjoner blir regnet som like viktige uansett om de klustrene det er snakk om har svært like eller svært ulike verdier på interessevariablen. For å gi en fullgod behandling av prediksjons-problemet må andre statistiske metoder brukes. Man kan tenke seg å bruke spesielle regresjonsanalysemetoder som tar hensyn til korrelasjonen mellom nabodyp.

Samsvaret mellom logg- og kjerneklustrene kan visualiseres ved å plotte de to klustertypene som funksjon av dybden. Ved å f.eks. bruke ulike farvekoder for klustrene kan dette gi et godt bilde av sammenhengen.

4. DATAGRUNNLAGET

Dataene er hentet fra dyp mellom 1370 m og 1670 m. De består av 217 kjerneprøver med observasjoner av variablene

GRSZ, SRT, MICA, CC, CON, BTB

og av 1201 logg-prøver, med observasjoner av

GR, RHOB, PHIN, DT.

For kjerneprøvene foreligger det 185 dyp med verdier på alle 6 variable. For 28 dyp mangler verdien bare på BTB. For disse har vi estimert BTB-verdiene på følgende måte: på grunnlag av klusteranalyse av de 185 dypene har vi laget 10 klustere. For hver av de ufullstendige observasjonssett avgjør vi hvilken av de 10 sentroidene som ligger nærmest - når vi tar hensyn til de 5 første variable. Dernest settes den manglende BTB-verdi lik gjennomsnittet for vedkommende kluster. Til slutt justeres verdien til ett av de to heltallene på hver side av gjennomsnittet, idet det tas hensyn til BTB-verdiene til nabodyp. De estimerte BTB-verdiene er gitt i Appendiks 3.

Variabelen GRSZ har en ganske skjev fordeling, og som input til klusteranalysene er i stedet brukt den naturlige logaritme, ln GRSZ.

For loggprøvene er det fulle verdisett for alle 1201 prøver. Vi har antatt at RHOB og PHIN er korrigert for hydrokarbon-effekt.

En sammenligning av kluster-resultater basert på kjerneprøver og loggprøver forenkles om man kjenner "sant" dyp for prøvene. Det foreliggende prosjekt viser tydelig hvor vanskelig det er å dybdebestemme prøvene. Dataene som analyseres i denne rapporten er dybdeskiftet av Statoil, delvis ut fra preliminære klusteranalyser.

5. RESULTATER

Resultatene fra klusteranalysene på et datasett er dokumentert i detalj i regnemaskinutskrifter. For hvert datasett er det følgende utskrifter.

1) Dendrogram fra Wards metode på hele datasettet. Antall klustere er indikert.

2) Resultatet av k-means når klustrene fra Wards metode danner utgangsposisjonen.

3) Deskriptiv statistikk for klustrene som k-means har dannet.

4) Noen av klustrene er karakterisert ved høye verdier på CC og CON. Dypene fra disse klustrene fjernes fra materialet. Det foretas ny klustring på det reduserte materiale, først med Ward, som så brukes som utgangspartisjon for k-means. Antall klustere er lik antallet vi bruker i punkt 2, minus antall kalk-rike klustere.

5) Deskriptiv statistikk for klustere fra punkt 4.

6) Et plott som gir et bilde av kluster-tilhørigheten som funksjon av dybden. De klustere som inngår er de kalkrike klustere fra punkt 3 og klustrene fra punkt 5.

7) Diskriminantanalyse for klustrene fra punkt 4.

8) Sammenligning av resultatene av to klusteranalyser på samme datasett.

Beregningene er foretatt ved NR's program NCLUST. Programmet for dybdeplottet er skrevet i forbindelse med det foreliggende prosjekt. Beregningene av diskriminantfunksjonene er foretatt ved hjelp av programpakken GENSTAT.

Følgende tabell gir en oversikt over regnemaskinutskriftene.

Tabell 1. Oversikt regnemaskinutskrifter.

Type	Oppdeling	Utskrifter	
Kjerne-prøver	Ikke imputert	1 - 7	
	Imputert	- " -	
Logg-prøver	Hele materialet	1 - 3	
	Red.,veiet, 9 kl	4 - 7	8
	Red.,veiet, 9 kl	- " -	
	Red.,uveiet,7 kl	- " -	8
	Red.,veiet, 7 kl	- " -	

5.1 Kjerneprøver

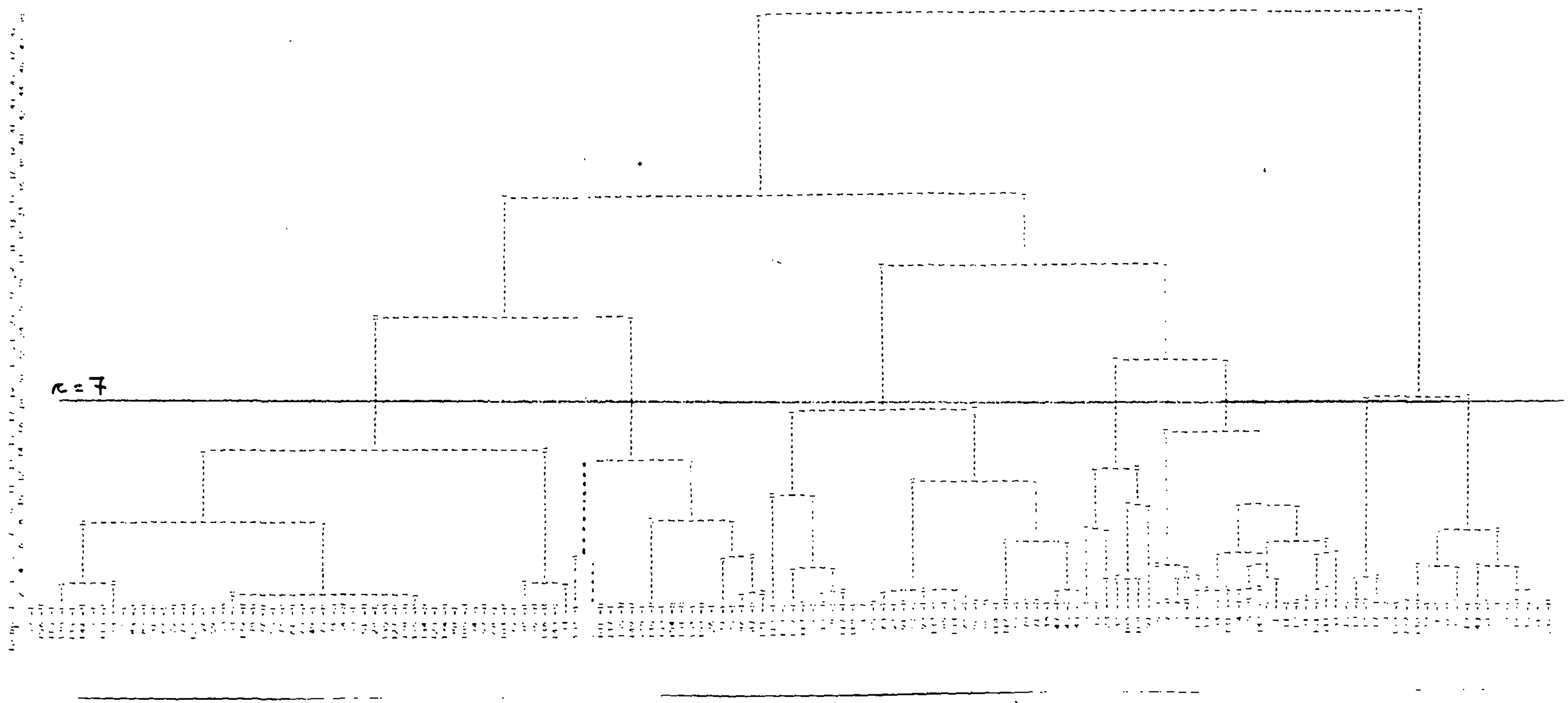
Det er laget 2 facies-inndelinger basert på data fra kjerneprøvene, én uten og én med imputerte verdier på BTB.

Uten imputerte verdier på BTB er det 185 kjerneprøver. Fra dendrogrammet for Wards metode foreslår vi 10 klustere, 3 av disse er kalkholdige. Dette er klusternumrene 3,5 og 6 i utskriften "Cluster Post Analysis". Det reduserte materiale består av 156 kjerneprøver. I Tabell 2 presenterer vi gjennomsnittet av variablene for de 7 klustrene fra det reduserte materiale (dette er klustrene R1, R2,... ..,R7) + de 3 kalkholdige klustrene som er bestemt fra hele materialet (klustrene H3, H5, H6). Klustrene er ordnet efter fallende verdi på GRSZ.

Tabell 2. Gjennomsnittsverdier for kjerneklustere uten imputerte BTB-verdier.

Kluster	R3	H6	R4	R5	H3	R1	R2	R6	H5	R7
Kalkrik?		JA			JA				JA	
Antall prøver	8	6	17	6	12	24	29	21	11	51
lnGRSZ	5.90	5.76	5.76	5.25	5.10	4.99	4.59	4.55	4.54	4.41
SRT	5.50	5.00	6.53	3.50	4.83	6.75	7.62	6.90	7.64	7.84
MICA	1.12	1.33	1.00	3.33	2.00	3.25	3.72	3.52	3.64	3.98
CC	3.00	3.83	1.00	1.67	3.92	1.25	2.28	1.29	4.00	1.00
CON	1.25	4.00	1.47	2.00	3.92	1.71	1.38	3.10	4.00	1.73
BTB	1.50	1.17	1.24	2.67	3.17	2.58	3.14	3.00	2.82	3.02

Figur 1 viser dendrogrammet fra Wards metode anvendt på det reduserte materiale. Dendrogrammet indikerer at det også kan gjøres en finere inndeling enn de 7 klustrene. Dette er imidlertid ikke utforsket videre.



Figur 1.

Dendrogram fra Wards metode anvendt på kjernedata hvor kalkrike soner er fjernet.

Inkluderes imputerte BTB-verdier, blir det 213 kjerneprøver. Tabell 3 viser gjennomsnitt av variablene for 7 klustere bestemt fra det reduserte materiale (hvor kalkrike soner er fjernet) samt for de 3 kalkrike klustrene, bestemt fra alle 213 kjerneprøver. Klustrene er ordnet etter fallende verdi på GRSZ.

Tabell 3. Gjennomsnittsverdier for kjerneklustere hvor imputerte BTB-verdier inkluderes.

Kluster	H6	R6	R5	R3	H3	R4	R1	H8	R2	R7
Kalkrik	JA				JA			JA		
Antall prøver	8	24	16	19	14	17	21	12	53	29
lnGRSZ	5.74	5.71	5.71	5.44	4.99	4.96	4.55	4.53	4.52	4.39
SRT	5.25	6.54	4.25	6.32	4.79	7.18	6.90	7.67	7.62	7.79
MICA	1.37	1.17	1.88	2.42	2.14	3.06	3.52	3.67	3.92	3.93
CC	3.87	1.04	2.25	1.16	3.86	2.76	1.29	4.00	1.11	1.45
CON	4.00	1.33	1.25	1.32	3.86	1.53	3.10	4.00	2.00	1.00
BTB	1.12	1.08	1.75	2.95	3.14	3.24	3.00	2.83	2.89	2.97

Vi skal gi en kort beskrivelse av hovedtrekkene i Tabell 2.

Av de 3 kalkrike klustrene skiller ett seg ut (H8) som finkornet og med høy sortering. De to andre (H6 og H3) adskiller seg særlig høy BTB.

Av de ikke-kalkrike klustrene er det to som til dels har store CC-verdier, mens CON er liten (R5 og R4). Disse to klustrene adskiller seg særlig på MICA og BTB, det ene har lave verdier på disse to variablene, det andre høye.

De øvrige 5 klustrene har CC-verdier lik 1 og 2. Det med groveste korn av disse (R6), adskiller seg fra de 4 andre med meget lave verdier på MICA og BTB.

De 4 gjenværende kan deles i 2 grupper. Den ene meget finkornet, delvis meget godt sortert og meget glimmer-rikt. Den andre litt grovere, ikke så godt sortert og ikke så mye glimmer. De to klustrene i den første gruppen, R2 og R7, adskiller seg på konsolideringen, ved at det ene bare har 2-verdier, det andre bare 1-verdier. I den andre gruppen, R3 og R1, er det også særlig CON som skiller.

Diskriminantanalysen gir opphav til diskriminantfunksjoner (canonical variates) v_i og egenverdier (roots) λ_i som er gitt i tab. 4, 5.

Koeffisientene i v_i-ene er der kalt ladninger (loadings). Diskriminantfunksjonene tolkes etter hvilke variable som har store ladninger. Den relative størrelsen på røttene λ_i gir et mål på de tilhørende diskriminantfunksjonenes evne til å skille mellom klustrene (se Appendiks 2). I begge analysene er de tre første funksjonene klart viktigst.

I analysen uten imputerte BTB-verdier skiller den klart viktigste funksjonen (v_1) mellom finkornet, glimmerrik og grovkornet, lite glimmerholdig. De neste funksjonene er i hovedsak konsolideringsgrad og sortering, og de er omtrent like viktige.

I materialet med imputert BTB er bildet endret. Den viktigste funksjonen skiller nå mellom godt sortert, glimmerrik, finkornet, konsolidert og dårlig sortert, lite glimmerholdig, grovkornet, lite konsolidert. Den neste funksjonen er konsolideringsgrad, og den tredje skiller karbonatholdig (og til en viss grad også lite bioturbert) fra karbonatfattig (og bioturbert).

I tolkningen av en slik analyse må man ha det klart for seg at det er en viss vilkårlighet i utvelgelsen av diskrimiantfunksjoner. Det er tenkelig å lage nye funksjoner som er lineære transformasjoner av de gamle, f.eks. $w_1 = 3v_1 - v_2$, $w_2 = v_1 + v_2$. Vi kan da fremdeles si at settet av variable w_i "forklarer" mesteparten av variasjonen mellom grupper. Dette kan i seg selv innebære en forenkling av de opprinnelige data, dersom antallet av w_i er lite. Det kan også tenkes at et klokt valg av transformasjon gir nye funksjoner w_i som har en klarere tolkning enn v_i-ene. Definisjonen av v_i kan ses på nærmest som én bestemt måte å velge transformasjoner på, med bestemte egenskaper som kan være nyttige. To sett med diskriminantfunksjoner er altså i en viss forstand ekvivalente dersom de kan transformeres lineært over i hverandre.

I vårt tilfelle ser vi at de to settene av funksjoner til en viss grad kan transformeres over i hverandre. Det som skiller er først og fremst at konsolideringsgrad er klart viktig i analyse med imputert BTB mens den er tillagt forholdsvis liten vekt i analyse uten imputert BTB.

Når vi plotter diskriminantfunksjonene mot hverandre, kan vi få et visuelt inntrykk av hvordan klustrene ligger i forhold til hverandre, hvor godt de er adskilt og hvordan spredningen innen klustrne er. Slike plott er presentert i fig. 2-5. Her er klustrene betegnet med A,B,C,.. i stedet for 1,2,3.. . Klustersentroidene er markert med +, og overlapping mellom symboler er markert med : . Av plottene fremgår det at klustrene er rimelig godt adskilte.

Tabell 4. Kjernedata uten imputerte BTB-verdier.

*** LATENT ROOTS ***

ROOT	1	2	3	4	5	6
	8.3669	2.8397	2.3340	0.7728	0.0335	0.0035

PERCENTAGE VARIANCE

ROOT	1	2	3	4	5	6
	58.3452	19.8024	16.2755	5.3886	0.1636	0.0247

*** LATENT VECTORS(LOADINGS) ***

VECT	1	2	3	4	5	6
LNGRSZ	-0.9972	-0.1237	0.0982	-0.6323	-1.7557	0.2250
SRT	0.4377	0.0609	1.6475	0.7138	-0.5707	-0.0216
MICA	1.3398	0.4261	-0.5785	-0.9026	-1.2590	-0.9182
CC	-0.0315	1.7422	-0.2927	0.7159	-0.0431	-0.2408
CON	0.3676	-0.6357	-0.7713	1.0853	-0.3397	-0.1354
BTB	0.5915	-0.0308	-0.2492	-0.4124	-0.0259	1.3645

Tabell 5. Kjernedata med imputerte BTB-verdier.

*** LATENT ROOTS ***

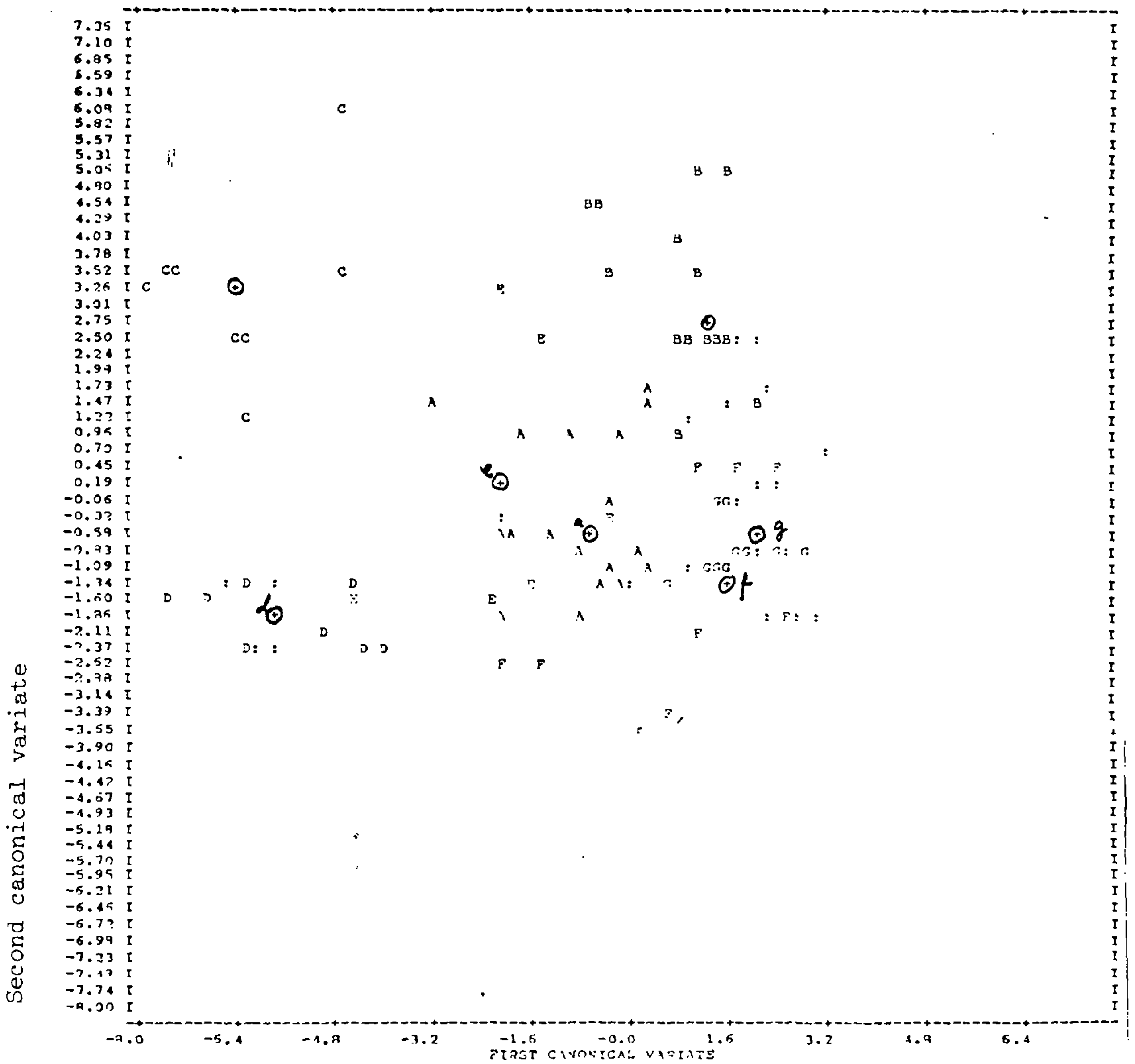
ROOT	1	2	3	4	5	6
	7.2571	3.2698	1.9430	0.5608	0.3663	0.0397

PERCENTAGE VARIANCE

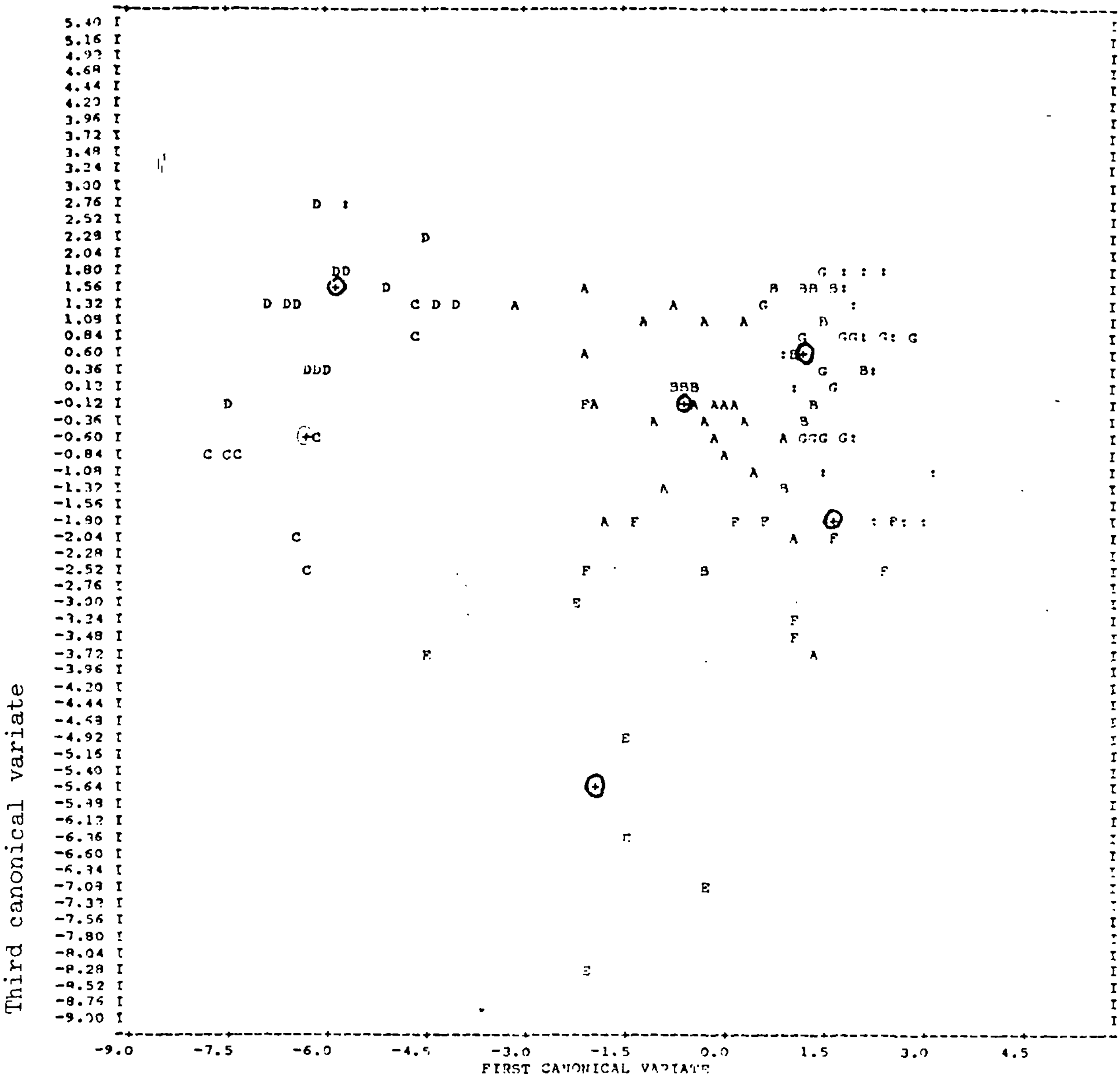
ROOT	1	2	3	4	5	6
	54.0097	24.3351	14.4602	4.1737	2.7259	0.2955

*** LATENT VECTORS(LOADINGS) ***

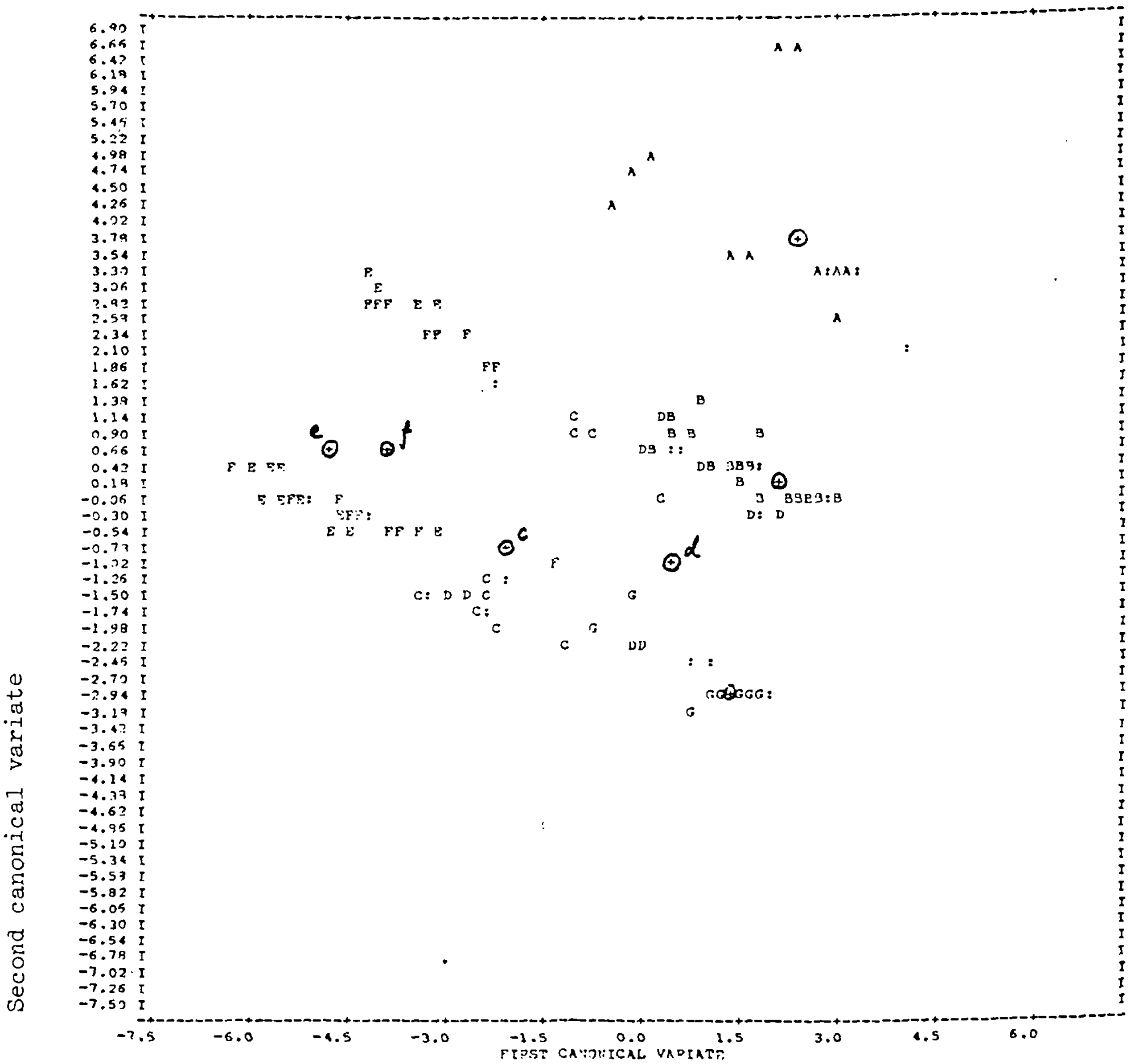
VECT	1	2	3	4	5	6
LNGRSZ	-0.7927	0.0252	0.0919	-1.2515	0.3353	-1.6522
SRT	0.9636	-0.4989	0.8207	0.0238	1.1218	-0.4045
MICA	0.9549	-0.3714	-0.0334	-0.0104	-0.9032	-1.6815
CC	0.1769	-0.0187	-1.2205	0.7833	0.4952	-0.1661
CON	0.7832	2.0141	-0.1290	0.0275	0.2887	-0.1632
BTB	0.2815	-0.4690	-0.8401	-1.4503	0.2756	0.4726



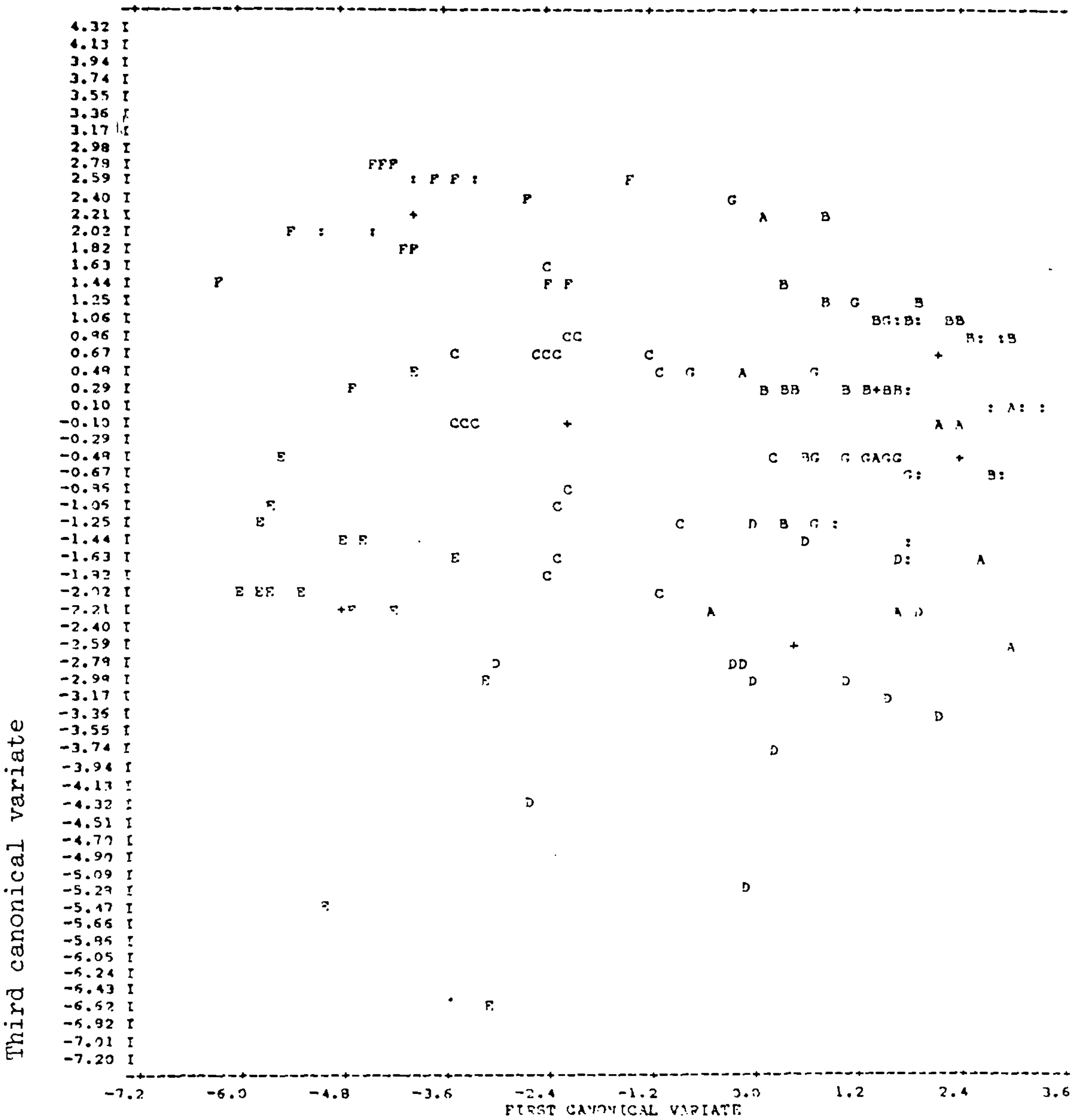
Figur 2. Kjernedata, ikke-imputert BTB.



Figur 3. Kjernedata, ikke-imputert BTB.



Figur 4. Kjernedata, imputert BTB.



Figur 5. Kjernedata, imputert BTB.

Et formelt mål på avstand mellom to klustre er Mahalanobis-avstanden D^2, se tabell 6, 7. (Også her er klustrene betegnet A,B,...).

Tabell 6. Kjernedata uten imputerte BTB-verdier.

***** MAHALANOBIS'S DISTANCES *****
MAHAL

	A	B	C	D	E	F
A	0.0000					
B	3.8172	0.0000				
C	7.0541	7.7066	0.0000			
D	5.7723	8.4318	5.6582	0.0000		
E	5.7824	7.7302	7.8621	8.6122	0.0000	
F	3.7604	4.9984	9.3590	8.3199	6.6733	0.0000
G	2.9225	3.5125	9.4959	8.0303	7.8625	3.5389

Tabell 7. Kjernedata med imputerte BTB-verdier.

***** MAHALANOBIS'S DISTANCES *****
MAHAL

	A	B	C	D	E	F
A	0.0000					
B	3.7356	0.0000				
C	6.6885	4.8295	0.0000			
D	5.6619	4.1491	4.4954	0.0000		
E	8.1203	7.5344	4.6578	6.0485	0.0000	
F	7.6079	6.3370	4.1493	6.7587	4.7201	0.0000
G	6.6592	3.1747	4.7959	3.9755	7.6015	6.7198

D^2 er et avstandsmål som bygger på at kovariansmatrisene for de to klustrene som inngår, er like. Under denne forutsetning representerer en D^2-verdi på ca. 5 eller større at en hypotetisk, fremtidig observasjon klassifiseres med rimelig sikkerhet til den ene av de to klustrene. I vår situasjon tyder fig. 2-5 på at for flere av klusterparene er denne forutsetningen neppe oppfylt. En vurdering av adskillelsen mellom klustrene bør derfor skje både ut fra D^2-verdiene og plottene av diskriminantfunksjonene.

5.2 Logg-prøver

For logg-prøvene er det først kjørt en klusteranalyse på hele materialet. Dendrogrammet gir gode indikasjoner på grupperinger. Vi har brukt 12 klustere, men andre valg ville antagelig også være interessante å undersøke nærmere.

Tabell 8 viser gjennomsnittsverier for de 4 variablene for hvert av de 12 klusterne, ordnet etter økende verdier på GR.

Tre av klustrene viser lave verdier på DT, nemlig 3, 10 og 11. For 10 og 11 er også RHOB stor. Dette tyder på at disse klustrene representerer karbonatsonene. Dette bekreftes, stort sett, når man sammenholder med kjerneklustringen og med "graphical log-presentation".

I den videre analyse har vi fjernet dypene fra disse 3 klustrene. Det reduserte materiale består da av 1069 dyp. På dette materialet foretas 4 sett av klusteranalyser, definert ved følgende oppsett

		Veiing	
		Uveiet	GR får vekt = 2
Antall	9	x	x
klustere	7	x	x

Noen resultater fra tilfellet med 9 klustere er gitt i Tabell 9, hvor vi presenterer gjennomsnittsverdier, ordnet etter voksende GR. Vi ser at mange av klusterparene har forholdsvis like sentroider. Noen par er mer ulike, det gjelder særlig (6,1) og (1,6). En grundigere sammenligning av den uveiete og veiete analysen er gitt i utskrift 8 , "Cluster analysis comparison".

Tabell 8. Gjennomsnittsverdier for loggklustere basert på hele materialet.

Kl.no	2	8	3	5	10	1	9	6	11	7	12	4
n_a	36	114	28	132	43	23	115	74	61	121	145	209
GR	28.56	37.45	38.36	42.08	43.71	48.96	51.51	54.53	57.62	65.09	68.29	70.62
RHOB	2.28	2.01	2.15	1.91	2.44	2.09	2.12	1.93	2.37	2.00	2.24	2.20
PHIN	.30	.14	.12	.15	.15	.44	.36	.23	.31	.32	.39	.37
DT	110.18	125.25	98.34	137.01	72.28	140.75	107.09	143.67	87.16	150.29	108.71	131.27

Tabell 9. Gjennomsnittsverdier for 9 uveiete og veiete logg-klustere basert på de ikke-kalkrike soner. Klustrene er nummerert i henhold til "Cluster Post Analysis".

UVEIET	Kluster	3	8	5	2	6	1	7	9	4
	Antall	37	117	124	115	71	84	86	242	193
1	GR	28.84	37.88	41.81	51.51	53.20	61.05	64.81	68.29	70.93
1	RHOB	2.28	2.01	1.91	2.12	1.92	2.08	1.97	2.24	2.20
1	PHIN	.30	.15	.15	.36	.22	.36	.31	.39	.37
1	DT	110.59	125.41	135.98	107.09	145.28	140.52	153.93	108.56	130.58
VEIET	Kluster	3	7	5	2	1	6	4	9	8
	Antall	35	108	152	95	23	63	124	262	207
2	GR	28.21	36.36	43.07	48.80	50.46	55.77	65.44	67.57	71.08
1	RHOB	2.28	2.00	1.92	2.12	2.08	1.93	2.01	2.23	2.20
1	PHIN	.30	.14	.16	.37	.44	.23	.31	.39	.37
1	DT	110.21	125.19	136.94	108.19	141.44	145.61	148.85	108.10	130.88

Her skal vi bare vise en matrise som viser graden av overensstemmelse rent antallsmessig.

Tabell 10. Sammenfall-matrise for loggklustring basert på ikke-kalkrike soner. 9 klustere. Klustrene nummerert som i Tabell 9.

		VEIET									
		3	7	5	2	1	6	4	9	8	SUM
UVEIET	3	35			2						37
	8		96	20			1				117
	5		12	111			1				124
	2				92			1	22		115
	6			20			49	2			71
	1			1		22	5	42		14	84
	7					1	7	78			86
	9				1				236	5	242
	4								4	188	193
SUM		35	108	152	95	23	63	124	262	207	1063

Tabellen viser at innføring av vekter endrer klustertilhørigheten for mange dyp. Det er også tydelig at noen klustere er mer stabile enn andre; dette gjelder særlig for klustrene med ekstreme GR-verdier.

Vi skal gi en kort beskrivelse av klustrene fra den veiete analysen. Klustrne 3, 2 og 9 har forholdsvis lave DT-verdier. To av dem har ganske høye verdier på RHOB. De adskiller seg særlig på GR.

Klustrene 7 og 5 skiller seg fra resten, ved at PHIN-verdiene er svært lave. Innbyrdes er de ulike på GR og DT.

De øvrige 4 klustrene kan inndeles i 2 hovedgrupper, de med høye GR-verdier (klustrene 4 og 8) og de med moderate GR-verdier (klustrene 1 og 6). Klustrene 4 og 8 adskiller seg innbyrdes med ca. 10% på alle 4 variable. Klustrene 1 og 6 er mer like, bortsett fra i variabelen PHIN. Sammenfalls-matrisen i Tabell 10 tyder jo også på at disse klustrene er mer ustabile.

Resultater for tilfellet med 7 klustere er gitt i Tabell 11.

Tabell 11. Gjennomsnittsverdier for 7 uveiete og veiete loggklustere basert på de ikke kalkrike soner. Klustrene er nummerert i henhold til "Cluster Post Analysis".

UVEIET	Kluster	3	6	5	2	1	7	4
	Antall	37	135	150	119	167	246	215
	GR	28.84	38.30	44.50	51.70	61.92	68.39	70.12
	RHOB	2.28	2.00	1.91	2.12	1.99	2.24	2.19
	PHIN	.30	.14	.17	.36	.31	.39	.37
	DT	110.59	126.08	138.74	107.51	149.40	108.83	131.77
VEIET	Kluster	2	4	1	5	3	7	6
	Antall	37	227	110	85	136	265	209
	GR	28.43	39.43	48.72	52.13	64.51	67.48	71.07
	RHOB	2.28	1.96	2.12	1.93	2.00	2.23	2.20
	PHIN	.29	.15	.38	.21	.32	.39	.37
	DT	110.58	130.83	113.28	143.90	149.06	108.10	130.94

Klustrene i parene (5,1) og (2,5) har særlig forskjellige sentroider. For de andre parene er de forholdsvis like. At det er noe spesielt med de to nevnte par fremgår også av sammenfallsmatrisen i Tabell 12.

Tabell 12. Sammenfall-matrise for loggklustring basert på ikke-kalkrike soner. 7 klustre. Klustrene nummerert som i tabell 11.

		VEIET							
		2	4	1	5	3	7	6	SUM
UVEIET	3	35		2					37
	6	2	129		3	1			135
	5		97		53				150
	2			92		1	26		119
	1		1	12	29	125			167
	7						238	8	246
	4		.	4		9	1	201	215
SUM		37	227	110	85	136	265	209	1069

For grundigere sammenligning refereres til utskrift 8 .

Diskriminantfunksjoner og egenverdier fra diskriminantanalysen av de fire logg-klustringene er vist i tabell 13-16 og i fig 6-9. For alle analysene gjelder det at de to første diskrimiantfunksjonene (v_1 og v_2) kan forklare mesteparten av variasjonen mellom klustrene. Videre er det små endringer når man går fra 7 til 9 klustre.

I analysene uten veiing måler første diskriminantfunksjon avtakende GR, RHOB, PHIN og voksende DT. Den andre funksjonen måler voksende GR og DT.

I analysene med veiing måler den første funksjonen avtakende GR og PHIN. Den andre komponenten måler voksende PHIN og avtakende GR, DT.

Den viktigste forskjellen mellom veiet og uveiet ser ut til å være at RHOB får liten vekt i det veiete tilfellet.

Mahalanobis-avstandene er presentert i tabell 17-20. Helt parallelt med analysen av kjernedataene kan vi konkludere at klustrene er klart separerte i alle de fire klustringene.

Tabell 13. 7 uveide loggklustre.

*** LATENT ROOTS ***

ROOT	1	2	3	4
	12.0051	5.4633	1.1951	0.5353

PERCENTAGE VARIANCE

ROOT	1	2	3	4
	62.5633	28.4715	6.1758	2.7894

*** LATENT VECTORS(LOADINGS) ***

VECT	1	2	3	4
GR	-1.4265	1.4710	-0.6807	-1.5645
RHOB	-0.8530	-0.5143	2.7093	-0.2570
PHIN	-1.4015	0.2961	-1.0410	1.9701
DT	0.8091	1.8985	1.5994	0.6588

Tabell 14. 9 uveide loggklustre.

*** LATENT ROOTS ***

ROOT	1	2	3	4
	17.1380	7.0605	1.3194	0.6010

PERCENTAGE VARIANCE

ROOT	1	2	3	4
	65.6153	27.0321	5.0514	2.3012

*** LATENT VECTORS(LOADINGS) ***

VECT	1	2	3	4
GR	-1.2534	1.8680	-0.8957	1.5119
RHOB	-1.2471	-0.4224	2.7385	0.4922
PHIN	-1.7875	0.1553	-0.8702	-2.0703
DT	0.9479	2.0374	1.7134	-0.6032

Tabell 15. 7 veide loggklustre.

*** LATENT ROOTS ***

ROOT			
1	2	3	4
15.1732	4.9339	0.7657	0.6601

PERCENTAGE VARIANCE

ROOT			
1	2	3	4
70.4652	22.9134	3.5560	3.0655

*** LATENT VECTORS(LOADINGS) ***

	VECT 1	2	3	4
GR	-1.2040	-1.0193	-0.2609	0.6709
RHOB	-0.3430	0.4491	2.4546	0.2759
PHIN	-1.4773	1.1121	-0.9652	-1.9615
DT	0.6181	-1.4001	1.4044	-0.9900

Tabell 16. 9 veide loggklustre.

*** LATENT ROOTS ***

ROOT			
1	2	3	4
16.5851	7.6207	1.1019	0.8143

PERCENTAGE VARIANCE

ROOT			
1	2	3	4
63.4911	29.1734	4.2184	3.1172

*** LATENT VECTORS(LOADINGS) ***

	VECT 1	2	3	4
GR	-1.4155	-1.0630	-0.8221	-0.1111
RHOB	-0.1400	0.8115	0.9121	-2.3405
PHIN	-1.4539	1.0948	1.5371	1.8068
DT	0.4410	-1.7279	1.6440	-0.8669

Tabell 17. 7 uveide loggklustre.

***** MAHALANOBIS'S DISTANCES *****

MAHAL

A	0.0000					
B	6.9132	0.0000				
C	9.9910	6.6602	0.0000			
D	4.4140	5.6054	8.8682	0.0000		
E	5.0702	7.1968	8.6450	7.8098	0.0000	
F	6.9729	6.7718	6.5519	8.3133	2.8836	0.0000
G	7.2244	3.8768	8.2615	3.6465	9.2267	8.9694
	A	B	C	D	E	F

Tabell 18. 9 uveide loggklustre.

***** MAHALANOBIS'S DISTANCES *****

MAHAL

A	0.0000							
B	6.1155	0.0000						
C	9.5185	6.8727	0.0000					
D	3.4017	6.3427	9.7240	0.0000				
E	7.6037	8.8234	9.9752	10.0217	0.0000			
F	5.2499	8.5176	10.9556	7.9930	3.3122	0.0000		
G	3.4494	8.9714	12.1388	6.0369	7.0356	3.8044	0.0000	
H	7.8392	7.6848	7.5781	9.6425	2.7746	5.2183	8.3582	0.0000
I	6.0658	4.2763	8.8146	3.9755	11.3012	10.0942	9.1304	10.3134
	A	B	C	D	E	F	G	H

Tabell 19. 7 uveide loggklustre.

***** MAHALANOBIS'S DISTANCES *****

MAHAL

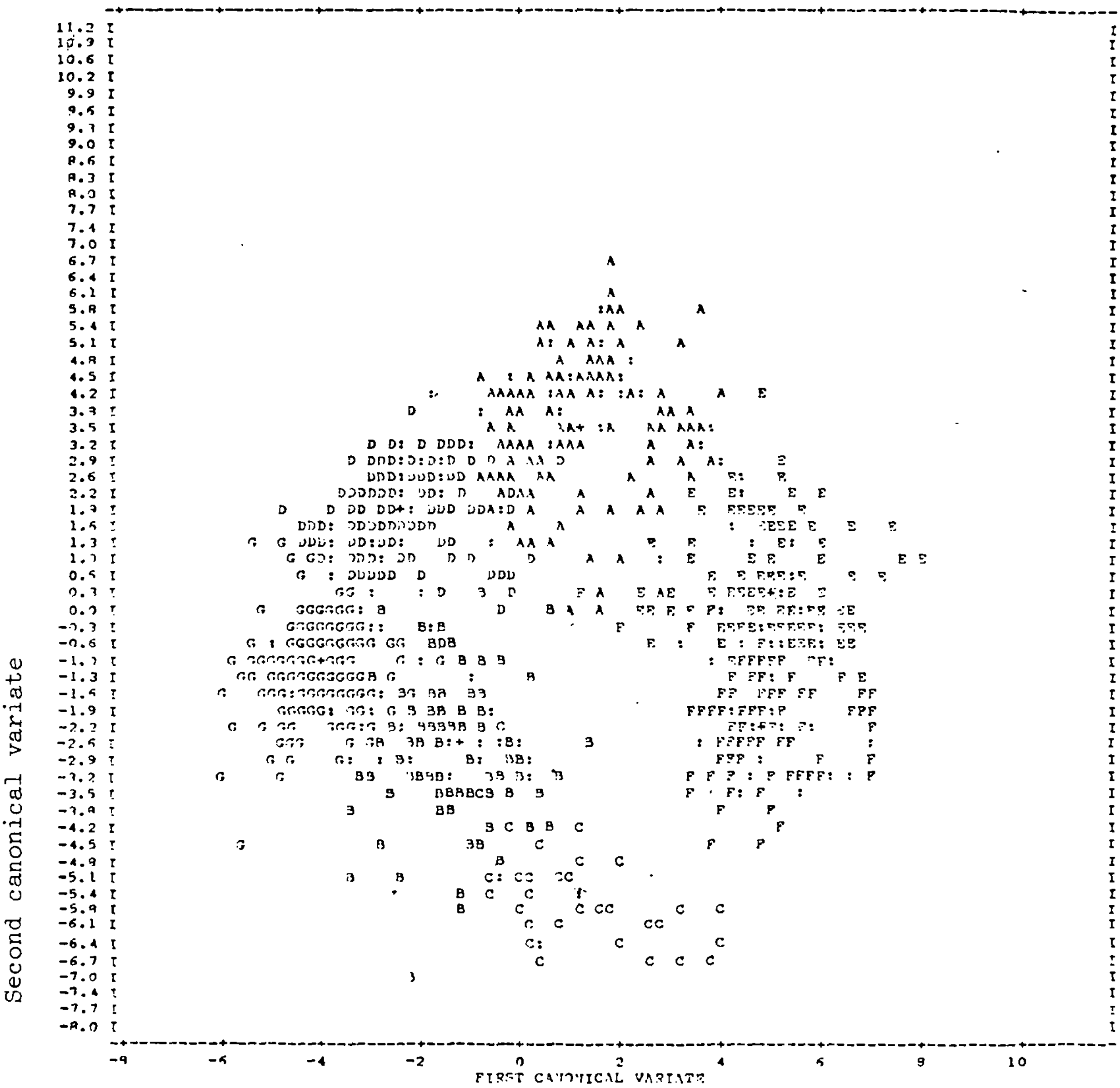
A	0.0000					
B	6.6720	0.0000				
C	6.5076	10.7514	0.0000			
D	7.2139	7.1772	7.2548	0.0000		
E	6.6870	9.2036	3.9038	3.6503	0.0000	
F	6.2636	10.8665	3.9075	9.4852	6.9790	0.0000
G	4.8641	10.2173	6.1379	9.9787	8.2536	3.2819
	A	B	C	D	E	F

Tabell 20. 9 uveide loggklustre.

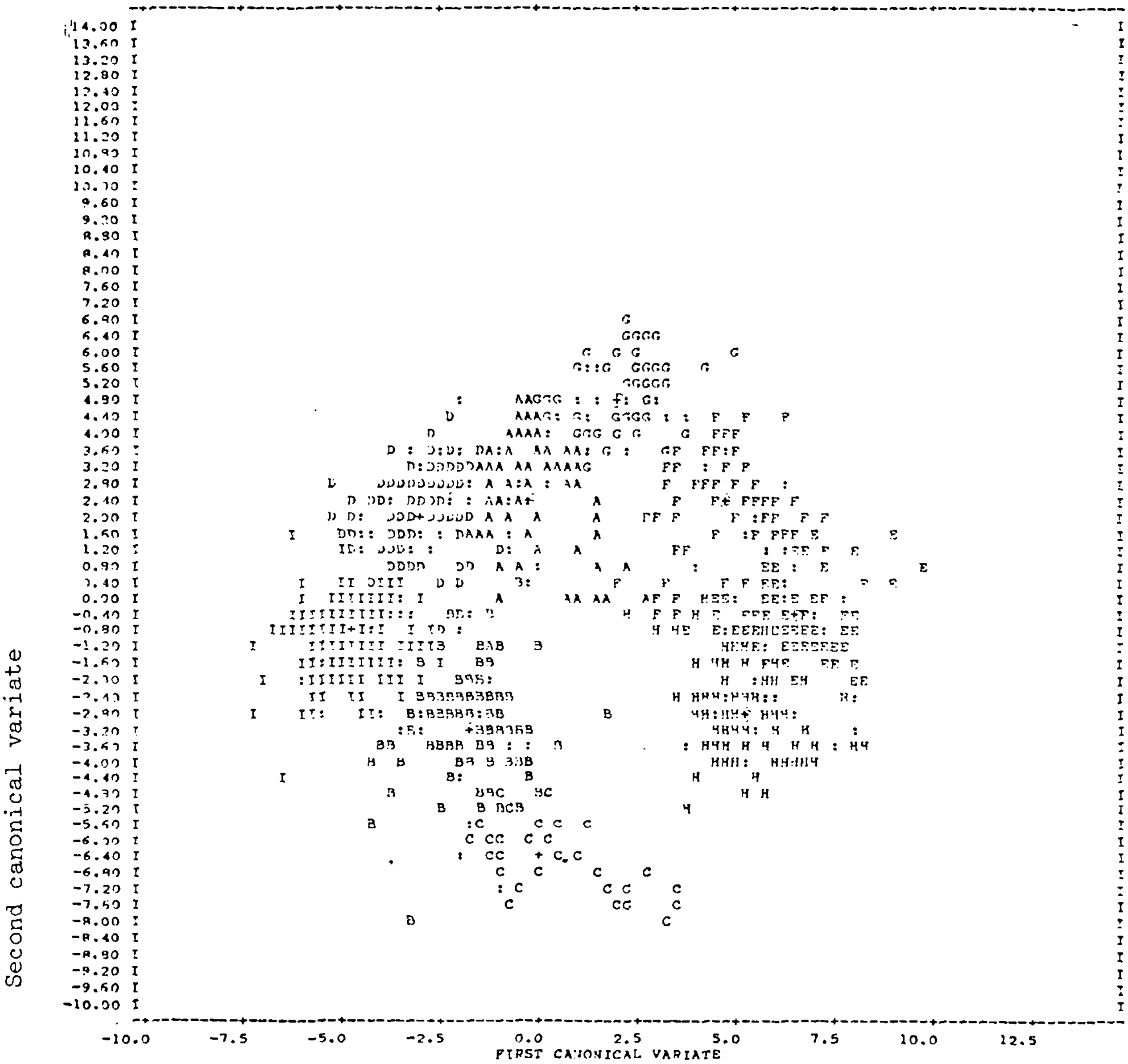
***** MAHALANOBIS'S DISTANCES *****

MAHAL

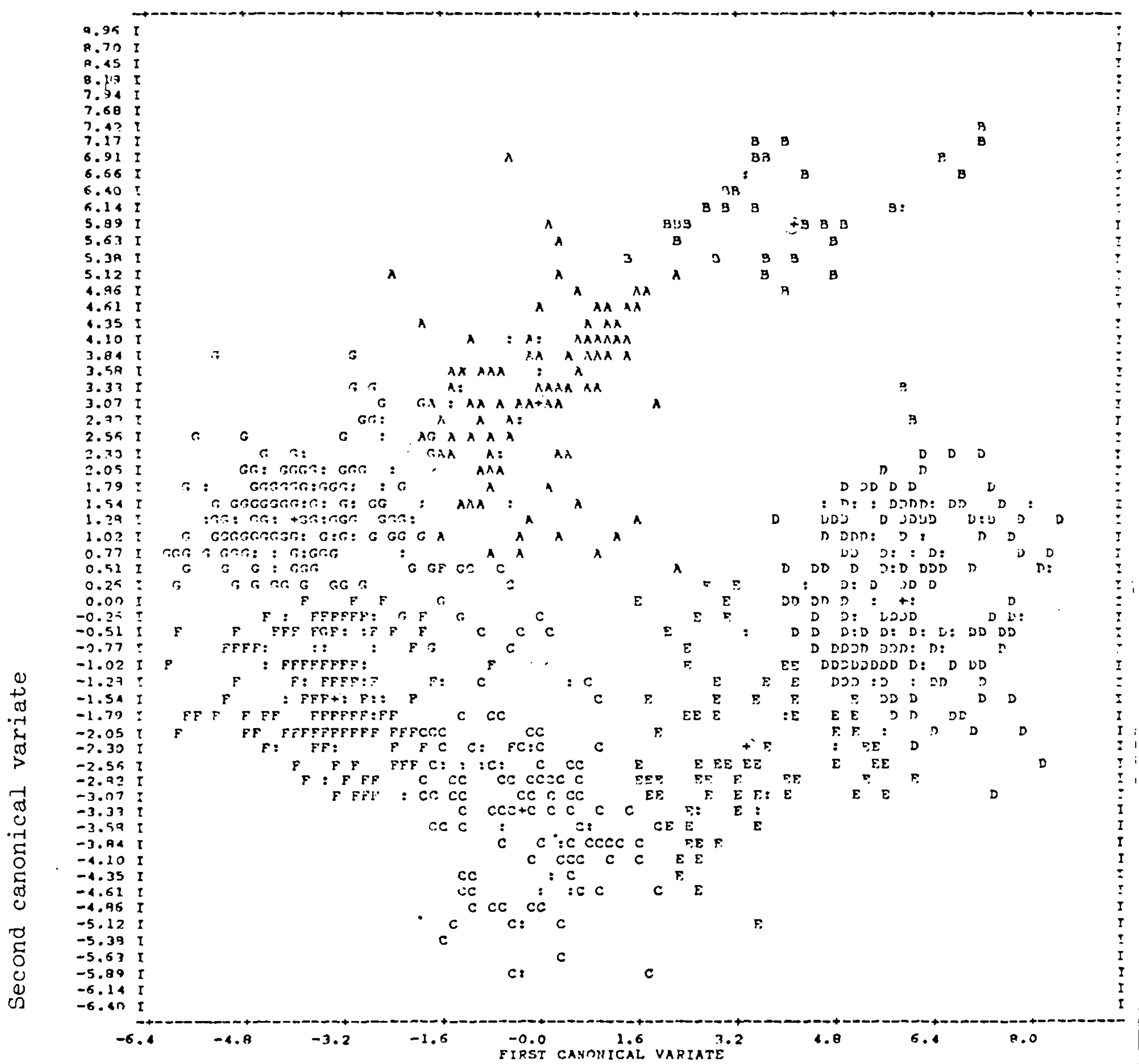
A	0.0000							
B	5.0924	0.0000						
C	9.4325	7.5757	0.0000					
D	6.1677	8.3296	13.4164	0.0000				
E	7.7979	7.7054	9.6435	7.1310	0.0000			
F	6.5928	7.8752	12.0185	3.3150	4.0159	0.0000		
G	8.7697	7.5107	7.4311	9.4736	2.9582	6.6971	0.0000	
H	6.3616	7.2482	12.7690	4.0950	9.2639	6.4475	10.7061	0.0000
I	6.7819	5.1453	11.5556	6.9290	9.9752	8.2371	10.6654	3.7539
	A	B	C	D	E	F	G	H



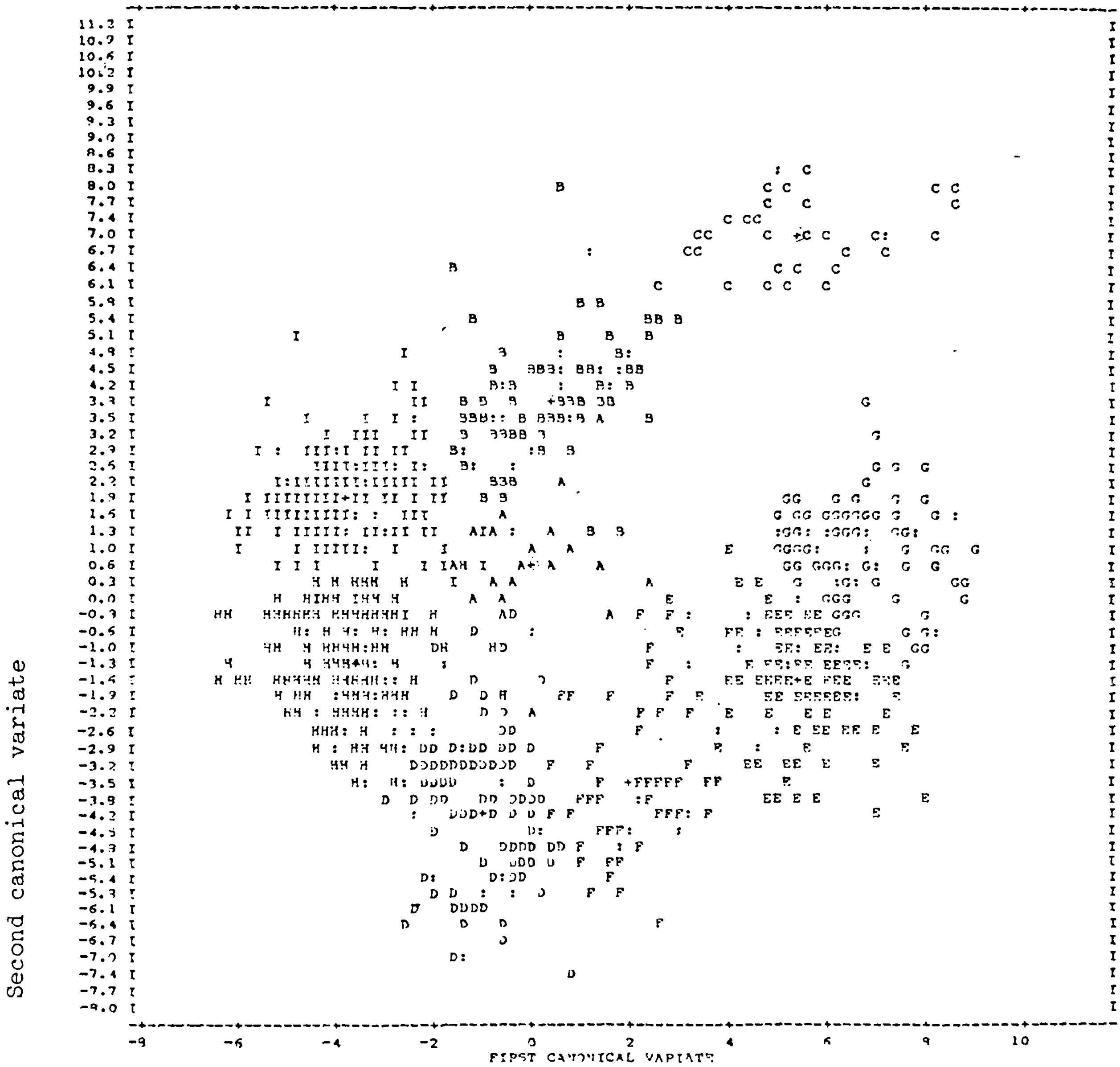
Figur 6. Uveide loggdata, 7 klustre.



Figur 7. Uveide loggdata, 9 klustre.



Figur 8. Veide loggdata, 7 klustre.



Figur 9. Veide loggdata, 9 klustre.

5.3 Sammenligning av klustringsresultater fra kjerne- og loggdata.

For hver kombinasjon av de to kjerneklustringene og de fire loggklustringene har vi satt opp en tabell som viser samsvaret mellom logg- og kjernekluster. I tabellene er alle dyp som har verdier både på logg- og kjernevariable tatt med, også de kalkrike sonene. Som en hjelp i tolkningen av disse tabellene (tabell 21-28) har vi beregnet noen mål på samvariasjon. Disse målene er beskrevet i Appendiks 4. Det som kanskje er mest interessant er $\lambda_{R|C}$, som måler hvor mye oftere (relativt sett) man kan predikere kjernekluster riktig ut fra kjennskap til loggkluster, enn når man ikke kjenner loggkluster. Verdiene 0 og 1 på $\lambda_{R|C}$ svarer til hhv. ingen sammenheng og total sammenheng.

Vi får verdier på $\lambda_{R|C}$ mellom 0.19 og 0.32, så det er ikke noen sterk sammenheng mellom kjerne- og loggkluster. Det ser ut som om samsvaret er størst uten imputerte BTB-verdier. Dette kan skyldes at det på de nye dypene som kommer med er dårligere samsvar mellom logger og kjerner, eller at de "kunstige" BTB-verdiene har falt langt fra de sanne verdiene.

I diskriminantanalysen trengte vi 3 diskriminantfunksjoner til å "forklare" kjernedataene, men bare 2 til å "forklare" loggdataene. Det kan således se ut som om kjernedataene "har en ekstra dimensjon" som disse fire loggene ikke inneholder informasjon om. Dette kan kanskje underbygge det dårlige samsvaret og antyde hva forklaringen kan være.

Tabell 21. Uveide loggdata, 10 klustre, ikke-imputerte kjernedata.

KRYSSKLASSIFIKASJON AV LOG- OG KJERNEKLUSTRE

KJERNEKLUSTER	LOGKLUSTER R1	R4	R2	R3	R5	R6	R7	H3	H10	H11	
R6	3	2	1		1		4				11
R1	10	3	4		1	1	2		1		22
R2	4	12	1				9			3	29
R3			1			5		1	1		8
R4	1				7	5					13
R7	1	23	2				21				47
R5		1				1	2		1	1	6
H3							3	2	3	3	11
H5	2								1	5	8
H6		1				2		1	1		5
	21	42	9		9	14	41	4	8	12	160

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS TAU(R!C) : .220

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS TAU(C!R) : .219

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS LAMBDA(C!R) : .263

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS LAMBDA(R!C) : .292

Tabell 22. Uveide loggdata, 10 klustre, imputerte kjernedata.

KRYSSKLASSIFIKASJON AV LOG- OG KJERNEKLUSTRE

LOGKLUSTER

KJERNEKLUSTER	R1	R4	R2	R3	R5	R6	R7	H3	H10	H11	
R1	3	2	1		1		4				11
R2	6	24	2		1		15		1	1	50
R3	4	1	2		4	5	1				17
R6	2				6	9		1			18
R7	1	12					14			1	28
R4	4	3	4				3	1	1	1	17
R5	1				1	9	2	1	1	1	16
H3	1						3	2	4	3	13
H6		1			1	3		1	1		7
H8	2								1	6	9
	24	43	9		14	26	42	6	9	13	186

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS TAU(R!C) : .171

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS TAU(C!R) : .200

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS LAMBDA(C!R) : .273

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS LAMBDA(R!C) : .191

Tabell 23. Uveide loggdata, 12 klustre, ikke-imputerte kjernedata.

KRYSSKLASSIFIKASJON AV LOG- OG KJERNEKLUSTRE

LOGKLUSTER

KJERNEKLUSTER	R1	R3	R4	R6	R5	R7	R8	R2	R9	H3	H10	H11	
R6	2		1	1		2		1	4				11
R1	6		1		1	6	1	4	2		1		22
R2	3		10			3		1	9			3	29
R3							5	1		1	1		8
R4				4	4		5						13
R7	1		24					2	20				47
R5	1						1		2		1	1	6
H3									3	2	3	3	11
H5						2					1	5	8
H6	1						2			1	1		5
	14		36	5	5	13	14	9	40	4	8	12	160

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS TAU(R!C) : .253

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS TAU(C!R) : .208

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS LAMBDA(C!R) : .217

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS LAMBDA(R!C) : .319

Tabell 24. Uveide loggdata, 12 klustre, imputerte kjernedata.

KRYSSKLASSIFIKASJON AV LOG- OG KJERNEKLUSTRE

LOGKLUSTER

KJERNEKLUSTER		R1	R3	R4	R6	R5	R7	R8	R2	R9	H3	H10	H11	
	R1	2		1	1		2		1	4				11
	R2	4		24		1	3		2	14		1	1	50
	R3	2			1	4	2	5	2	1				17
	R6				4	6	1	6			1			18
	R7	1		12						14			1	28
	R4	4					3		4	3	1	1	1	17
	R5	1				2		8		2	1	1	1	16
	H3	1								3	2	4	3	13
	H6	1				1		3			1	1		7
	H8						2					1	6	9
		16		37	6	14	13	22	9	41	6	9	13	186

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS TAU(R!C) : .203

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS TAU(C!R) : .197

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS LAMBDA(C!R) : .255

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS LAMBDA(R!C) : .213

Tabell 25. Veide loggdata, 10 klustre, ikke-imputerte kjernedata.

KRYSSKLASSIFIKASJON AV LOG- OG KJERNEKLUSTRE

LOGKLUSTER

KJERNEKLUSTER		R3	R1	R2	R6	R5	R4	R7	H3	H10	H11	
	R6	4	1		1	1		4				11
	R1	11	3		2	2		3		1		22
	R2	5			12			9			3	29
	R3		1				5		1	1		8
	R4					6	7					13
	R7	1	1		24			21				47
	R5				1	1		2		1	1	6
	H3							3	2	3	3	11
	H5	2								1	5	8
	H6	1					2		1	1		5
		24	6		40	10	14	42	4	8	12	160

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS TAU(R!C) : .234

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS TAU(C!R) : .241

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS LAMBDA(C!R) : .280

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS LAMBDA(R!C) : .310

Tabell 26. Veide loggdata, 10 klustre, imputerte kjernedata.

KRYSSKLASSIFIKASJON AV LOG- OG KJERNEKLUSTRE

KJERNEKLUSTER	LOGKLUSTER R3	R1	R2	R6	R5	R4	R7	H3	H10	H11	
R1	4	1		1	1		4				11
R2	6	1		26	1		14		1	1	50
R3	3	1		1	4	6	2				17
R6	1				5	11		1			18
R7	1			12			14			1	28
R4	6	3		1			4	1	1	1	17
R5		1	1			9	2	1	1	1	16
H3					1		3	2	4	3	13
H6	1					4		1	1		7
H8	2								1	6	9
	24	7	1	41	12	30	43	6	9	13	186

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS TAU(R!C) : .194

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS TAU(C!R) : .230

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS LAMBDA(C!R) : .329

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS LAMBDA(R!C) : .206

Tabell 27. Veide loggdata, 12 klustre, ikke-imputerte kjernedata.

KRYSSKLASSIFIKASJON AV LOG- OG KJERNEKLUSTRE

KJERNEKLUSTER	LOGKLUSTER R1	R2	R3	R8	R4	R5	R6	R7	R9	H3	H10	H11	
R6		1		1	4	1			4				11
R1		3		2	10	2	1		3		1		22
R2				11	7				8			3	29
R3		1						5		1	1		8
R4						5	3	5					13
R7		1		24	1				21				47
R5				1			1		2		1	1	6
H3									3	2	3	3	11
H5					2						1	5	8
H6					1			2		1	1		5
		6		39	25	8	5	12	41	4	8	12	160

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS TAU(R!C) : .236

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS TAU(C!R) : .221

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS LAMBDA(C!R) : .252

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS LAMBDA(R!C) : .301

Tabell 28. Veide loggdata, 12 klustre, imputerte kjernedata.

KRYSSKLASSIFIKASJON AV LOG- OG KJERNEKLUSTRE

KJERNEKLUSTER \ LOGKLUSTER	R1	R2	R3	R8	R4	R5	R6	R7	R9	H3	H10	H11	
R1		1		1	4	1			4				11
R2		1		26	5	1	1		14		1	1	50
R3		1		1	3	5	2	3	2				17
R6					1	5	3	8		1			18
R7				12	1				14			1	28
R4		3			8				3	1	1	1	17
R5	1					1		9	2	1	1	1	16
H3							1		3	2	4	3	13
H6					1	1		3		1	1		7
H8					2						1	6	9
	1	6		40	25	14	7	23	42	6	9	13	186

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS TAU(R!C) : .207

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS TAU(C!R) : .219

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS LAMBDA(C!R) : .313

SAMVARIASJONSMAAL:GOODMAN/KRUSKALS LAMBDA(R!C) : .228

APPENDIKS 1. OM KLUSTRINGSMETODENE WARD OG K-MEANS

La x_{ik} være observasjonen på k-te variabel for i-te dataenhet ($i=1,\ldots,n$, $k=1,\ldots,m$), eventuelt efter standardisering, veiing og transformering (f.eks. er

$$x_{ik} = w_k \frac{x'_{ik} - \bar{x}'_k}{s'_k},$$

hvor $w_1,\ldots,w_m$ er vektene, $\{x'_{ik}\}$ råobservasjonene, $\bar{x}'_k$ og s'_k gjennomsnitt og standardavvik for k-te variabel).

Ved Wards metode starter vi med n klustere: hver dataenhet utgjør ett kluster. Metoden går ut på å slå sammen på hvert trinn de to klustrene som gir minst økning i variansen innen klustrene. La det efter et visst antall trinn være c klustere. La

$$x_{ik}^{(a)} = \text{i-te observasjon fra kluster a av variabel k,}$$

$i=1,\ldots,n_a$ (= antall dataenheter i kluster a), $a=1,\ldots,c$.

La videre $\bar{x}_k^{(a)} = \sum_{i=1}^{n_a} x_{ik}^{(a)}/n_a$ og definer vektorene

$$\underset{\sim}{x}_i^{(a)\prime} = (x_{i1}^{(a)},\ldots, x_{im}^{(a)}), \quad \underset{\sim}{\bar{x}}^{(a)\prime} = (\bar{x}_1^{(a)},\ldots,\bar{x}_m^{(a)})$$

(her er $\underset{\sim}{\bar{x}}^{(a)}$ <u>sentroiden</u> for kluster a). Variansen innen kluster a er

$$(1) \qquad E_a = \sum_{i=1}^{n_a} (\underset{\sim}{x}_i^{(a)} - \underset{\sim}{\bar{x}}^{(a)})'(\underset{\sim}{x}_i^{(a)} - \underset{\sim}{\bar{x}}^{(a)}),$$

og variansen innen alle klustrene $E = \sum_{a=1}^{c} E_a$. Vi ser at

$$(2) \qquad E_a = \sum_{i=1}^{n_a} \underset{\sim}{x}_i^{(a)\prime}\underset{\sim}{x}_i^{(a)} - n_a\, \underset{\sim}{\bar{x}}^{(a)\prime}\underset{\sim}{\bar{x}}^{(a)}.$$

Slår vi sammen klustrene a og b til et nytt kluster t vil økningen i E bli

$$\begin{aligned}\Delta E_{ab} &= E_t - (E_a+E_b)\\ &= n_a\, \underset{\sim}{\bar{x}}^{(a)\prime}\underset{\sim}{\bar{x}}^{(a)} + n_b\, \underset{\sim}{\bar{x}}^{(b)\prime}\underset{\sim}{\bar{x}}^{(b)} - n_t\, \underset{\sim}{\bar{x}}^{(t)\prime}\underset{\sim}{\bar{x}}^{(t)}\end{aligned}$$

på grunn av (2). Setter vi her inn at

$$n_t \underset{\sim}{\bar{x}}^{(t)} = n_a \underset{\sim}{\bar{x}}^{(a)} + b_b \underset{\sim}{\bar{x}}^{(b)}$$

og at $n_t = n_a + n_b$, leder en enkel algebraisk manipulasjon til at vi kan skrive

$$\Delta E_{ab} = \frac{n_a n_b}{n_a + n_b} (\underset{\sim}{\bar{x}}^{(a)} - \underset{\sim}{\bar{x}}^{(b)})' (\underset{\sim}{\bar{x}}^{(a)} - \underset{\sim}{\bar{x}}^{(b)})$$

$$= h(n_a, n_b) d^2 (\underset{\sim}{\bar{x}}^{(a)}, \underset{\sim}{\bar{x}}^{(b)}) ,$$

altså produktet av det harmoniske gjennomsnitt av n_a og n_b og den kvadratiske Euklidske avstand mellom sentroidene i klustrene a og b.

Man slår altså sammen det klusterpar for hvilket dette produktet er minst.

Anta vi fra Wards metode bestemmer oss for c klustrene. La $\underset{\sim}{\bar{x}}_o^{(1)}, \ldots, \underset{\sim}{\bar{x}}_o^{(k)}$ være de tilhørende sentroider. Første trinn går ut på å plassere hver dataenhet i det kluster hvor avstanden til sentroiden er minst. Hvis altså

$$\min_{1 \leq a \leq c} d^2 (\underset{\sim}{x}_i, \underset{\sim}{x}_o^{(a)})$$

oppnås for $a = a_o$, plasseres i-te dataenhet i kluster a_o. Når alle dataenhetene er gjennomgått på denne måten, beregnes nye sentroider $\underset{\sim}{\bar{x}}_1^{(1)}, \ldots, \underset{\sim}{\bar{x}}_1^{(k)}$.

I neste runde går man påny gjennom alle dataenhetene; hver dataenhet plasseres i det kluster hvis sentroide ligger nærmest. Men hver gang dette leder til skifte av klustertilhørighet, oppdateres de to tilsvarende sentroidene. Prosessen fortsetter inntil vi har hatt en runde uten skifte av klustertilhørighet.

(Litteratur: se f.eks. Anderberg, M.R. "Cluster Analysis for Applications", Academic Press 1973 (sec. 6.2.6 og 7.2)).

APPENDIKS 2. KARAKTERISERING AV KLUSTERE VED DISKRIMINANTFUNKSJONER

Anta klusteranalysen gir forslag til c klustere. Vi bruker samme notasjon som i Appendiks 1 og lar

$$\underset{\sim}{x}_1^{(a)}, \ldots, \underset{\sim}{x}_{n_a}^{(a)}$$

være observasjonsvektorene for de n_a dataenhetene i kluster a, med sentroide $\underset{\sim}{\bar{x}}^{(a)}$, (a=1,...,c).

Forskjellen mellom klustrene kan vurderes ved hjelp av Wilks Λ-observator

$$\Lambda^{-1} = \frac{|W + B|}{|W|}$$

hvor

$$W = \sum_a \sum_{i=1}^{n_a} (\underset{\sim}{x}_i^{(a)} - \underset{\sim}{\bar{x}}^{(a)})(\underset{\sim}{x}_i^{(a)} - \underset{\sim}{\bar{x}}^{(a)})'$$

$$B = \sum_a n_a (\underset{\sim}{\bar{x}}^{(a)} - \underset{\sim}{\bar{x}})(\underset{\sim}{\bar{x}}^{(a)} - \underset{\sim}{\bar{x}})' ,$$

hvor $\underset{\sim}{\bar{x}} = \frac{1}{n} \sum n_a \underset{\sim}{\bar{x}}^{(a)}$.

La nå $y_i^{(a)} = \underset{\sim}{v}' \underset{\sim}{x}_i^{(a)}$, i=1,...,n_a, a=1,...,c, være en lineærtransformasjon av $\{\underset{\sim}{x}_i^{(a)}\}$. Bestem $\underset{\sim}{v}$ slik at F-observatoren for å teste forskjell på klustrene basert på $\{\tilde{y}_i^{(a)}\}$ blir størst mulig. F-observatoren er

$$F = \text{konstant x } \frac{ss_b}{ss_w}$$

hvor

$$ss_b = \sum_a n_a (\bar{y}^{(a)} - \bar{y})^2 = \underset{\sim}{v}' B \underset{\sim}{v}$$

$$ss_w = \sum_a \sum_i (y_i^{(a)} - \bar{y}^{(a)})^2 = \underset{\sim}{v}' W \underset{\sim}{v} .$$

Vi søker derfor den $\underset{\sim}{v}$ som maksimerer

$$\lambda(\underset{\sim}{v}) = \frac{\underset{\sim}{v}'B\underset{\sim}{v}}{\underset{\sim}{v}'W\underset{\sim}{v}} \cdot$$

Settes de partielt deriverte lik null, får vi å løse

$$(W^{-1}B - \lambda I)\underset{\sim}{v} = 0.$$

La $\lambda_1 > \lambda_2 > \ldots . > \lambda_r > 0$, med $r = \min(c-1,m)$, være egenverdiene til $W^{-1}B$ og la $\underset{\sim}{v}_1,\ldots,\underset{\sim}{v}_r$ være de tilsvarende egenvektorene. Da er

$$y_{ij}^{(a)} = \underset{\sim}{v}_j' \underset{\sim}{x}_i^{(a)} \quad , \; i=1,\ldots,n_a, \quad a=1,\ldots,c$$

de y-er som gir den j-te største F-observator ($j=1,\ldots,r$). Disse r lineære transformasjonene kalles diskriminantfunksjonene.

De er ukorrelerte, både innen og mellom klustere. Betrakt f.eks.

$$c_{12} = \sum_{a,i} (y_{i1}^{(a)} - \bar{y}_1^{(a)})(y_{i2}^{(a)} - \bar{y}_2^{(a)}) = \underset{\sim}{v}_1'W\underset{\sim}{v}_2$$

Nå er $(B - \lambda_j W)\underset{\sim}{v}_j = 0$ for $j = 1,2$, dsv.

$$B\underset{\sim}{v}_1 = \lambda_1 W\underset{\sim}{v}_1 \; , \; B\underset{\sim}{v}_2 = \lambda_2 W\underset{\sim}{v}_2$$

hvorav

$$\underset{\sim}{v}_2'B\underset{\sim}{v}_1 = \lambda_1 \, \underset{\sim}{v}_2'W\underset{\sim}{v}_1 \; , \quad \underset{\sim}{v}_1'B\underset{\sim}{v}_2 = \lambda_2 \, \underset{\sim}{v}_1'W\underset{\sim}{v}_2$$

Da B er symmetrisk, er venstresidene like. Og da W også er symmetrisk, er $(\lambda_1 - \lambda_2)\underset{\sim}{v}_1'W\underset{\sim}{v}_2 = 0$. Hvis $\lambda_1 < \lambda_2$ er derfor $c_{12} = 0$.

Det to-dimensjonale plottet av $(y_{i1}^{(a)}, y_{i2}^{(a)})$, $i=1, \ldots, n_a$, $a=1,\ldots,c$, og også av klustergjennomsnittene $(\bar{y}_1^{(a)}, \bar{y}_2^{(a)})$, $a=1,\ldots,c$, vil kunne gi en interessant karakteristikk av klustrene.

Hvorvidt de to første diskriminantfunksjonene inneholder mesteparten av variablenes diskriminerende evne vil kunne vurderes ved hjelp av en trinnvis variant av Wilkes test.

Ifølge Bartlett er

$$V = -(n-1 - \frac{m+c}{2})\ln\Lambda = (n-1 - \frac{m+c}{2})\sum_{j=1}^{r}\ln(1+\lambda_j)$$

tilnærmet $\chi^2_{m(c-1)}$ under H_o (ingen forskjell).

Og

$$V_j = (n-1 - \frac{m+c}{2})\ln(1+\lambda_j)$$

er tilnærmet uavhengige og χ^2_{m+r-2j}. Følgende trinnvise prosedyre er da foreslått: hvis V-testen forkaster, har det fulle sett diskriminerende evne. Betrakt så $V-V_1$, som under H_o er tilnærmet $\chi^2_{(m-1)(c-2)}$. Hvis denne differensen forkaster, har det fulle sett, minus den beste diskriminantfunksjon, også diskriminerende evne. Betrakt så $V-V_1-V_2$, som under H_o er tilnærmet $\chi^2_{(m-2)(c-3)}$. Hvis denne differensen ikke er større enn den kritiske verdi, konkluderer vi med at det er de to første diskriminantfunksjonene som har diskriminerende evne.

(Litteratur: Tatsuoka, M.M. "Multivariate Analysis", John Wiley 1971 (ch.6)).

APPENDIKS 3. IMPUTERTE BTB-VERDIER

Dyp	GRSZ	SRT	MICA	CC	CON	Estimert BTB
1424.50	200	7	2	1	1	2
1434.50	120	3	2	1	1	3
1434.75	275	6	2	1	1	3
1435.50	300	6	2	1	1	3
1436.50	280	7	2	1	1	3
1437.50	250	7	2	1	1	3
1438.50	240	7	2	1	1	3
1439.50	270	6	2	1	1	3
1440.50	250	6	2	4	4	1
1441.50	330	4	2	1	1	2
1443.25	350	6	1	4	4	1
1447.75	30	5	3	4	4	3
1448.75	320	4	2	2	1	2
1449.75	350	5	2	2	1	2
1450.75	420	4	2	4	1	2
1452.75	500	5	1	2	1	2
1461.75	380	7	1	1	1	1
1462.75	320	7	1	1	1	1
1463.75	400	6	2	1	1	1
1465.75	310	6	1	2	1	1
1471.25	400	6	2	1	1	1
1472.25	440	5	2	2	1	2
1480.50	150	6	4	4	1	3
1490.75	160	6	2	2	1	3
1540.25	70	8	4	1	2	3
1548.75	80	8	4	4	4	3
1622.00	230	7	1	1	1	1
1623.00	190	7	1	1	1	1

APPENDIKS 4. SAMVARIASJONSMÅL I REKTANGULÆRE KONTINGENSTABELLER

Anta at ett av kjennetegnene $A_1, \dots, A_r$ og ett av $B_1, \dots, B_s$ kan inntreffe. Her lar vi A_i være kjernekluster i og B_j loggkluster j. Sannsynligheten for at A_i, B_j skal inntreffe betegnes p_{ij} og vi har data, ordnet i en tabell, som viser antall ganger X_{ij} denne begivenheten har inntruffet. Samvariasjonsmålene vi vil bruke tar utgangspunkt i prediksjonsevne: hvor godt kan kjennskap om at B_j har inntruffet brukes til å predikere A_i? To slike mål er Goodman og Kruskals $\tau_{R|C}$ og $\lambda_{R|C}$.

$\tau_{R|C}$ måler forklart varians på lignende måte som i varians- eller regresjonsanalyse, men med utgangspunkt i Ginis variasjonsmål for kategoriske variable.

Vi har X_{i+} betegne antall ganger A_i opptrer, dvs. $X_{i+} = \sum_j X_{ij}$.

$N = X_{++} = \sum_{ij} X_{ij}$ er antall dyp i alt. Da er Ginis variasjon for variabelen A definert som

$$T = \frac{N}{2} - \frac{1}{2N} \sum_i X_{i+}^2 .$$

Variasjonen innen grupper for B_j er på tilsvarende vis

$$W = \sum \{ \frac{X_{+j}}{2} - \frac{1}{2X_{+j}} \sum_i X_{i+}^2 \} .$$

Dersom B_j predikerer A_i godt, er W liten i forhold til T, og vi definerer

$$\tau_{R|C} = \frac{T-W}{W} .$$

$\lambda_{R|C}$ måler reduksjonen i feilsannsynlighet i prediksjon av A. Dersom vi ikke kjenner B, vil den beste gjetning være at A_m inntreffer, bestemt av at $p_{m+} = \max_i (p_{i+})$. Sannsynligheten for å ta feil blir da $1 - p_{m+}$. Dersom B_j har inntruffet, blir feilsannsynligheten, helt analogt, $1 - p_{mj}/p_{+j}$ der $p_{mj} = \max_i (p_{ij})$. Den totale feilsannsynligheten for prediksjon på grunnlag av B blir derfor

$$\sum_j p_{+j}(1 - p_{mj}/p_{+j}) = 1 - \sum_j p_{mj} .$$

Vi definerer $\lambda_{R|C}$ som den relative reduksjon i feilsannsynlighet:

$$\lambda_{R|C} = \frac{(1-p_{m+}) - (1-\sum p_{mj})}{1 - p_{m+}} = \frac{\sum p_{mj} - p_{m+}}{1 - p_{m+}} .$$

For å estimere $\lambda_{R|C}$ fra materialet, erstatter vi p_{ij} med estimatet X_{ij}/N.

Størrelsene $\tau_{C|R}$ og $\lambda_{C|R}$ måler A's evne til prediksjon av B, og fremkommer ved bare å bytte om "A_i" og "B_j" i definisjonene ovenfor.