Chapter 12

12.1 a) Codes:

1 → South	Yes	Democrat	Yes	Yes
0 → non-South	No	Republican	No	No

e.g. Reagan - Carter:

	1	0
1	1	0
0	2	2

, $\frac{a+d}{p} = 3/5 = .60$

Pair	Coefficient (a+d)/p
R-C	.6
R-F	.4
R-N	.6
R-J	0
R-K	.6
C-F	0
C-N	.2
C-J	.4
C-K	.6
F-N	.8
F-J	.6
F-K	.4
N-J	.4
N-K	.6
J-K	.4

12.1 b)

Pair	Coefficient 1	2	3	Rank Order 1	2	3
R-C	.6	.75	.429	4.5	4.5	4.5
R-F	.4	.571	.25	10	10	10
R-N	.6	.75	.429	4.5	4.5	4.5
R-J	0	0	0	14.5	14.5	14.5
R-K	.6	.75	.429	4.5	4.5	4.5
C-F	0	0	0	14.5	14.5	14.5
C-N	.2	.333	.111	13	13	13
C-J	.4	.571	.25	10	10	10
C-K	.6	.75	.429	4.5	4.5	4.5
F-N	.8	.889	.667	1	1	1
F-J	.6	.75	.429	4.5	4.5	4.5
F-K	.4	.571	.25	10	10	10
N-J	.4	.571	.25	10	10	10
N-K	.6	.75	.429	4.5	4.5	4.5
J-K	.4	.571	.25	10	10	10

12.2

Pair	Coefficient 5	6	7	Rank Order 5	6	7
R-C	.333	.5	.2	9	9	9
R-F	0	0	0	14	14	14
R-N	.333	.5	.2	9	9	9
R-J	0	0	0	14	14	14
R-K	.333	.5	.2	9	9	9
C-F	0	0	0	14	14	14
C-N	.2	.333	.111	12	12	12
C-J	.4	.571	.25	6	6	6
C-K	.5	.667	.333	3	3	3
F-N	.667	.8	.5	1	1	1
F-J	.5	.667	.333	3	3	3
F-K	.25	.4	.143	11	11	11
N-J	.4	.571	.25	6	6	6
N-K	.5	.667	.333	3	3	3
J-K	.4	.571	.25	6	6	6

12.3

x\y	1	0	Total
1	a	b	a+b
0	c	d	c+d
Total	a+c	b+d	p=a+b+c+d

$$r = \frac{\sum_{i=1}^{p}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{p}(x_i-\bar{x})^2 \sum_{i=1}^{p}(y_i-\bar{y})^2}}$$

$\bar{x} = (a+b)/p; \quad \bar{y} = (a+c)/p$

$$\Sigma(x_i-\bar{x})^2 = (a+b)(1-(a+b)/p)^2 + (c+d)(0-(a+b)/p)^2 = \frac{(c+d)(a+b)}{p}$$

$$\Sigma(y_i-\bar{y})^2 = (a+c)(1-(a+c)/p)^2 + (b+d)(0-(a+c)/p)^2 = \frac{(a+c)(b+d)}{p}$$

$$\Sigma(x_i-\bar{x})(y_i-\bar{y}) = \Sigma(x_iy_i-y_i\bar{x}-x_i\bar{y}+\bar{x}\bar{y})$$

$$= a - \frac{(a+c)(a+b)}{p} - \frac{(a+b)(a+c)}{p} + p\,\frac{(a+b)(a+c)}{p^2}$$

$$= \frac{a(a+b+c+d)-(a+c)(a+b)}{p} = \frac{ad-bc}{p}$$

Therefore

$$r = \frac{(ad-bc)/p}{\left[\frac{(c+d)(a+b)(a+c)(b+d)}{p^2}\right]^{\frac{1}{2}}} = \frac{ad-bc}{[(a+b)(c+d)(a+c)(b+d)]^{\frac{1}{2}}}$$

12.4 Let $c_1 = \frac{a+d}{p}$, $c_2 = \frac{2(a+d)}{2(a+d)+(b+c)}$ and $c_3 = \frac{a+d}{(a+d)+2(b+c)}$

then $c_3 = \frac{1}{1+2(c_1^{-1}-1)}$ so c_3 increases as c_1 increases

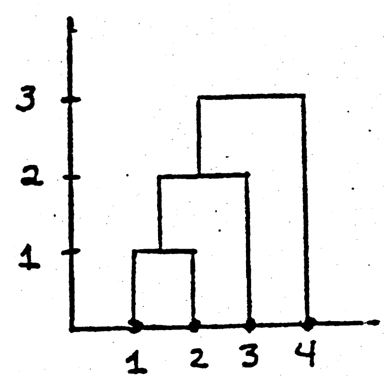
Also, $c_2 = \frac{2}{c_1^{-1}+1}$ so c_2 increases as c_1 increases

Finally, $c_2 = \frac{4}{c_3^{-1}+3}$ so c_2 increases as c_3 increases

12.5 a) Single linkage

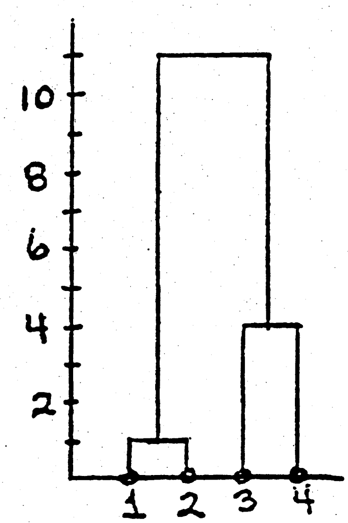
$$\begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & 0 & & & \\ 2 & 1 & 0 & & \\ 3 & 11 & 2 & 0 & \\ 4 & 5 & 3 & 4 & 0 \end{array} \rightarrow \begin{array}{c|ccc} & (12) & 3 & 4 \\ \hline (12) & 0 & & \\ 3 & 2 & 0 & \\ 4 & 3 & 4 & 0 \end{array} \rightarrow \begin{array}{c|cc} & (123) & 4 \\ \hline (123) & 0 & \\ 4 & 3 & 0 \end{array}$$

Dendogram



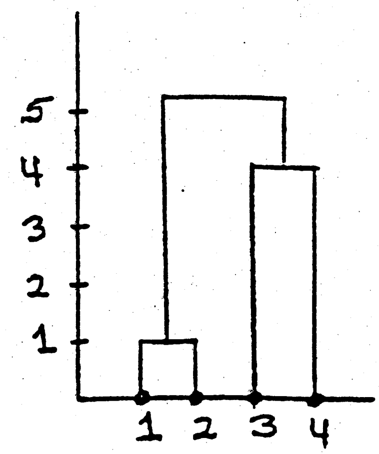
12.5 **b)** Complete Linkage

Dendogram

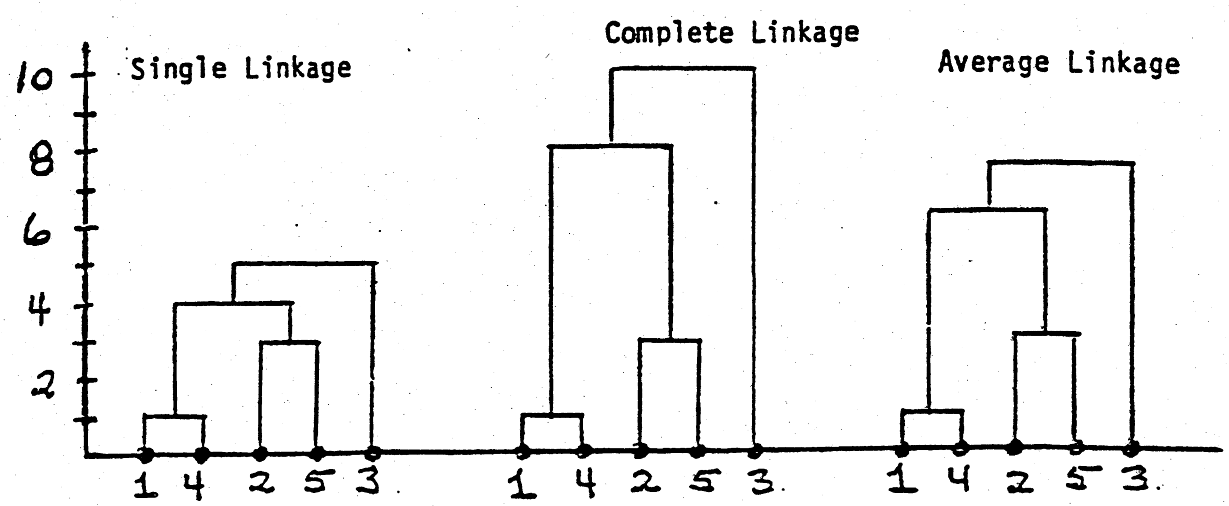


c) Average Linkage

Dendogram



12.6 Dendograms



All three methods produce the same hierarchical arrangements. Item 3 is somewhat different from the other items.

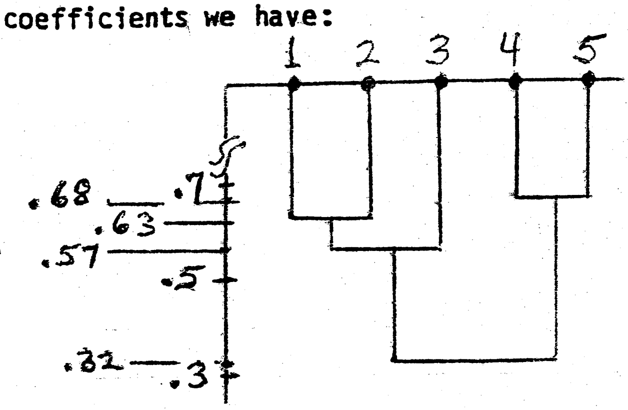
12.7 Treating correlations as <u>similarity</u> coefficients we have:

<u>Single linkage</u>

$S_{45} = .68$

$S_{(45)1} = \max(S_{41}, S_{51}) = .16$

$S_{(45)2} = .32, \quad S_{(45)3} = .18,$ and so forth.

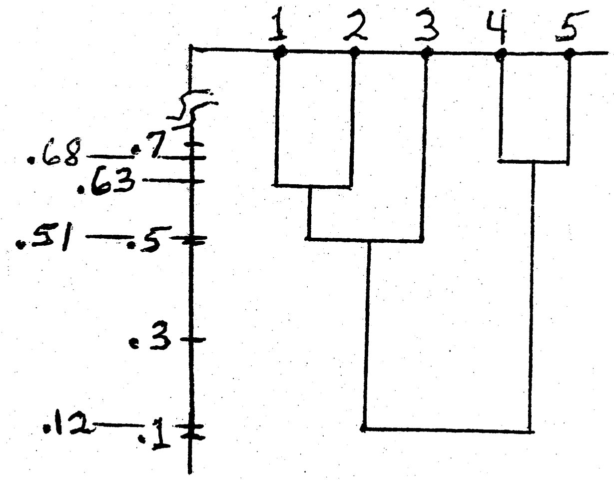


<u>Complete linkage</u>

$S_{45} = .68$

$S_{(45)1} = \min(S_{41}, S_{51}) = .12$

$S_{(45)2} = .21, \quad S_{(45)3} = .15$, and so forth.

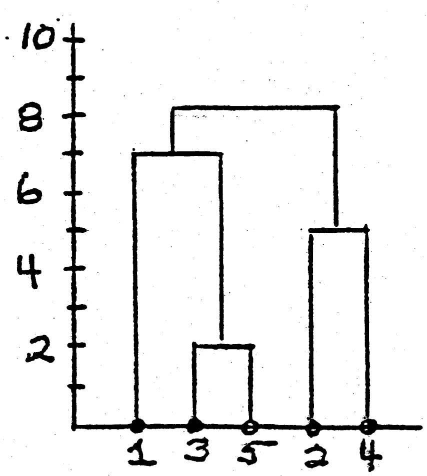


Both methods arrive at nearly the same clustering.

12.8

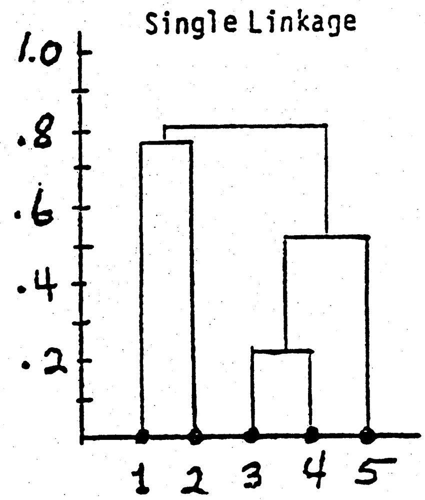
$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{bmatrix} \end{array} \rightarrow \begin{array}{c} \\ 1 \\ 2 \\ (35) \\ 4 \end{array}\begin{array}{c} \begin{array}{cccc} 1 & 2 & (35) & 4 \end{array} \\ \begin{bmatrix} 0 & & & \\ 9 & 0 & & \\ 7 & 8.5 & 0 & \\ 6 & 5 & 8.5 & 0 \end{bmatrix} \end{array} \rightarrow$$

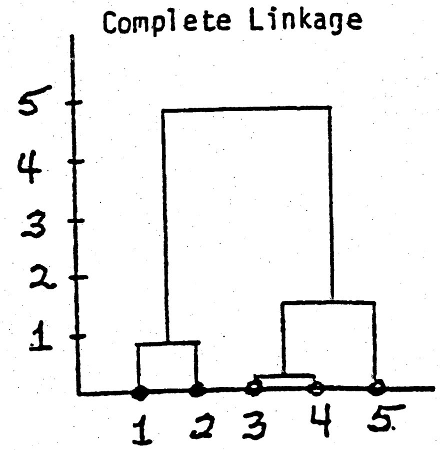
$$\begin{array}{c} \\ 1 \\ (35) \\ (24) \end{array}\begin{array}{c} \begin{array}{ccc} 1 & (35) & (24) \end{array} \\ \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 7.5 & 8.5 & 0 \end{bmatrix} \end{array} \rightarrow \begin{array}{c} \\ (135) \\ (24) \end{array}\begin{array}{c} \begin{array}{cc} (135) & (24) \end{array} \\ \begin{bmatrix} 0 & \\ 8.167 & 0 \end{bmatrix} \end{array}$$

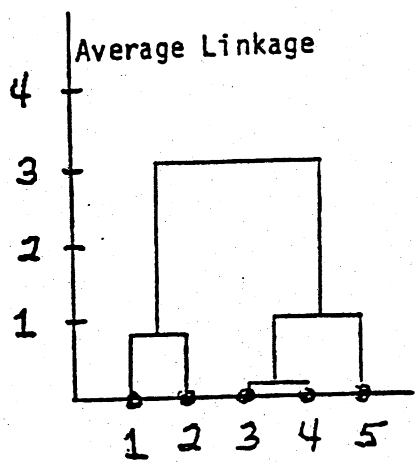


Average linkage produces results similar to single linkage.

12.9 Dendograms







Although the vertical scales are different, all three linkage methods produce the same groupings. (Note different vertical scales.)

12.10 (a) $ESS_1 = (2-2)^2 = 0$, $ESS_2 = (1-1)^2 = 0$, $ESS_3 = (5-5)^2 = 0$, and $ESS_4 = (8-8)^2 = 0$.

(b) At step 2

Clusters			Increase in ESS
{12}	{3}	{4}	.5
{13}	{2}	{4}	4.5
{14}	{2}	{3}	18.0
{1}	{23}	{4}	8.0
{1}	{24}	{3}	24.5
{1}	{2}	{34}	4.5

(c) At step 3

Clusters		Increase in ESS
{12}	{34}	5.0
{123}	{4}	8.7

Finally all four together have
$ESS = (2-4)^2 + (1-4)^2 + (5-4)^2 + (8-4)^2 = 30$

12.11 K = 2 initial clusters (AB) and (CD)

	$\bar{x}_1$	$\bar{x}_2$
(AB)	3	1
(CD)	1	1

Final clusters (AD) and (BC)

	$\bar{x}_1$	$\bar{x}_2$
(AD)	4	2.5
(BC)	0	−.5

Cluster	Squared distance to group centroids			
	A	B	C	D
(AD)	3.25	29.25	27.25	3.25
(BC)	45.25	3.25	3.25	11.25

12.12 K = 2 initial clusters (AC) and (BD)

	$\bar{x}_1$	$\bar{x}_2$
(AC)	3	.5
(BD)	−2	−.5

Final clusters (A) and (BCD)

	$\bar{x}_1$	$\bar{x}_2$
(A)	5	3
(BCD)	−1	−1

Cluster	Squared distance to group centroids			
	A	B	C	D
(A)	0	40	41	89
(BCD)	52	4	5	5

As expected, this result is the same as the result in Example 12.11. A graph of the items supports the (A) and (BCD) groupings.

12.13 K = 2 initial clusters (AB) and (CD)

	$\bar{x}_1$	$\bar{x}_2$
(AB)	2	2
(CD)	−1	−2

Final clusters (A) and (BCD)

	$\bar{x}_1$	$\bar{x}_2$
(A)	5	3
(BCD)	−1	−1

Cluster	Squared distance to group centroids			
	A	B	C	D
A	0✓	40	41	89
(BCD)	52	4✓	5✓	5✓

The final clusters (A) and (BCD) are the same as they are in Example 12.11. In this case we start with the same initial groups and the first, and only, reassignment is the same. It makes no difference if you start at the top or bottom of the list of items.

12.14. (a) The Euclidean distances between pairs of cereal brands

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
C1	0.0											
C2	116.0	0.0										
C3	15.5	121.7	0.0									
C4	6.4	117.9	10.0	0.0								
C5	103.2	61.6	100.6	102.1	0.0							
C6	72.8	44.1	78.4	74.4	54.3	0.0						
C7	86.4	71.9	82.5	84.9	22.3	52.4	0.0					
C8	15.3	121.5	1.4	10.1	100.6	78.3	82.4	0.0				
C9	46.2	72.6	54.7	48.9	75.8	32.1	65.2	54.5	0.0			
C10	54.9	123.0	68.9	59.5	134.7	87.8	122.5	68.8	65.7	0.0		
C11	81.3	154.7	94.7	85.8	169.6	121.3	157.0	94.6	94.5	47.1	0.0	
C12	42.3	114.2	31.3	38.5	81.1	75.3	60.2	31.0	59.8	92.9	121.9	0.0
C13	163.2	163.4	177.9	168.1	208.0	155.4	205.1	177.9	148.9	112.4	110.7	198.0
C14	46.7	90.8	60.4	51.5	103.8	55.4	92.9	60.3	28.5	44.3	67.5	75.9
C15	60.3	170.5	50.0	56.6	141.5	127.8	121.5	50.0	103.8	101.7	115.6	62.0
C16	46.9	90.8	60.5	51.6	103.8	55.5	92.9	60.3	28.5	44.3	67.6	75.8
C17	23.1	101.0	21.6	21.6	81.4	58.5	63.6	21.4	37.5	70.1	100.7	26.0
C18	265.7	221.1	280.0	270.6	278.9	233.9	283.3	280.0	235.6	227.7	218.6	294.5
C19	68.2	181.9	60.5	65.2	155.9	138.7	136.2	60.5	113.2	102.7	111.7	76.6
C20	116.6	71.0	113.2	115.3	19.7	69.9	32.1	113.1	89.3	150.5	183.5	90.6
C21	103.0	217.7	96.6	100.6	191.7	174.7	171.6	96.6	148.1	129.7	130.5	111.7
C22	98.6	160.1	112.6	103.4	181.3	130.5	170.2	112.6	106.9	54.1	22.5	139.2
C23	58.0	102.8	49.1	54.9	62.4	68.1	41.3	48.9	61.2	105.4	136.9	20.7
C24	68.1	181.8	60.4	65.2	155.8	138.7	136.1	60.4	113.1	102.7	111.6	76.5
C25	49.4	121.0	36.2	44.8	82.5	82.1	62.8	36.2	68.9	101.7	130.2	14.7
C26	182.8	290.3	186.0	183.8	285.6	250.4	267.2	185.9	220.2	173.8	145.7	210.7
C27	134.7	99.9	148.2	139.1	150.9	101.1	152.2	148.2	104.2	99.6	113.7	160.9
C28	16.1	128.3	14.2	14.2	111.1	85.7	92.3	13.7	59.2	63.5	86.3	39.4
C29	107.5	159.0	120.3	111.6	180.7	132.1	170.7	120.3	116.0	54.1	64.6	144.1
C30	33.5	120.1	21.2	29.2	90.7	78.8	71.2	21.0	61.7	83.1	113.7	17.2
C31	78.9	80.5	90.9	82.8	108.5	59.2	103.1	90.8	56.9	52.6	90.6	101.7
C32	32.1	122.6	43.5	36.0	120.8	83.1	105.0	43.3	51.3	50.9	60.0	65.9
C33	143.1	68.0	141.3	142.4	42.0	84.5	61.1	141.2	109.8	170.6	203.8	120.8
C34	173.0	157.7	187.8	177.9	207.5	155.6	206.8	187.8	151.8	127.0	123.8	205.9
C35	116.2	70.4	112.7	114.9	16.9	69.2	30.4	112.6	89.9	148.8	183.8	90.0
C36	114.1	230.0	111.1	112.9	210.2	186.9	190.8	111.1	158.8	129.8	122.7	131.2
C37	53.1	78.2	51.4	52.4	51.6	41.3	34.2	51.1	38.1	91.1	124.5	36.6
C38	54.2	100.4	45.8	51.0	61.8	63.5	43.5	45.8	59.0	99.2	133.6	25.8
C39	48.3	93.5	42.5	45.9	61.0	55.1	43.3	42.5	49.6	90.7	125.9	27.3
C40	40.6	140.9	51.6	44.3	139.8	100.7	123.8	51.4	70.3	44.1	46.2	79.4
C41	197.8	309.6	194.3	196.6	288.1	268.0	268.1	194.3	237.8	215.5	194.4	209.9
C42	191.1	301.3	190.3	190.8	286.6	260.4	267.3	190.2	229.3	200.8	174.0	209.7
C43	185.2	290.7	189.2	186.6	288.1	251.4	270.2	189.2	221.4	173.6	143.7	214.8

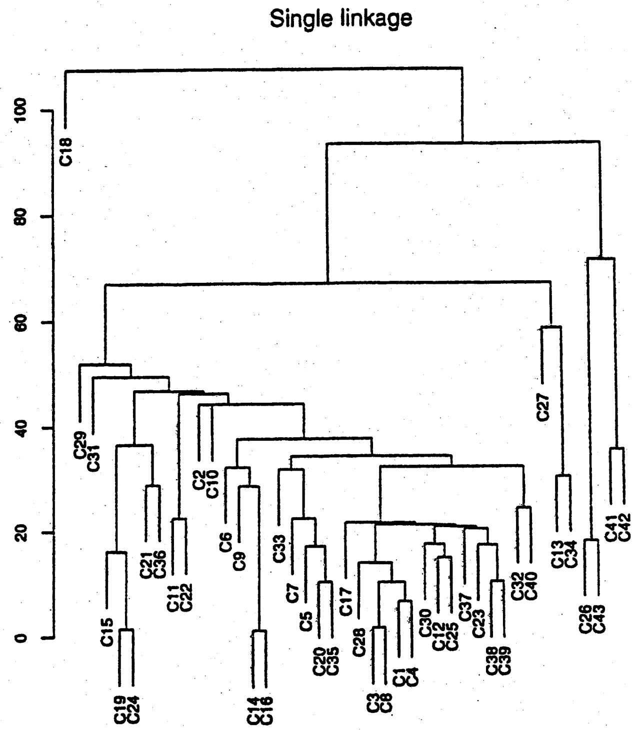
	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23	C24
C13	0.0											
C14	127.4	0.0										
C15	213.2	105.0	0.0									
C16	127.4	1.0	105.0	0.0								
C17	173.1	51.3	69.7	51.3	0.0							
C18	134.4	220.7	321.2	220.8	270.1	0.0						
C19	212.5	110.8	16.2	110.9	81.2	322.6	0.0					

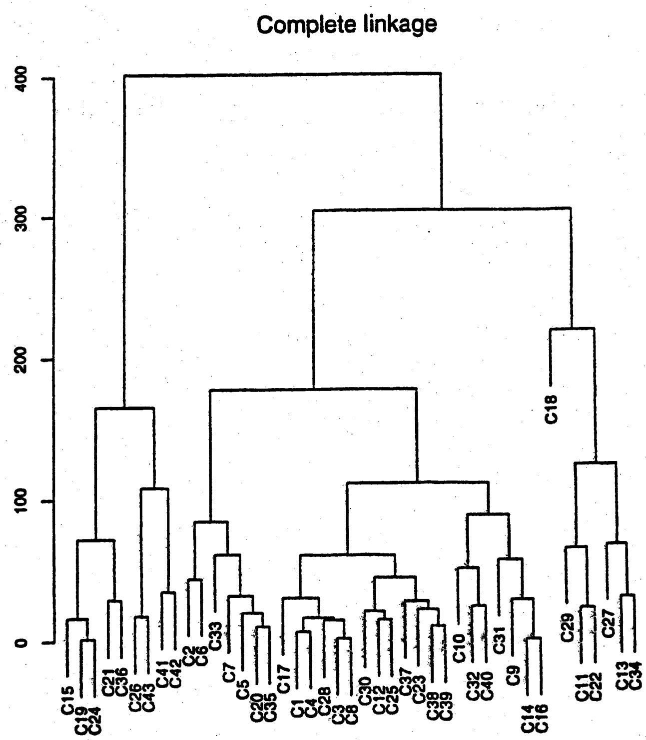
C20	223.2	117.3	151.2	117.3	94.3	288.6	166.1	0.0				
C21	234.6	142.8	50.3	142.8	117.2	347.4	36.5	201.2	0.0			
C22	91.5	79.1	135.2	79.2	116.8	204.1	131.1	195.9	148.8	0.0		
C23	204.9	83.3	81.1	83.2	36.8	295.9	96.2	70.9	130.9	153.2	0.0	
C24	212.5	110.7	16.0	110.8	81.1	322.6	1.4	166.0	36.5	131.1	96.1	0.0
C25	207.5	86.0	60.0	86.1	35.2	303.9	75.3	91.8	110.1	147.9	23.2	75.3
C26	233.8	200.3	159.3	200.3	204.2	342.0	143.8	297.3	121.0	152.7	231.2	143.8
C27	67.1	92.1	193.3	92.2	136.5	141.1	197.4	164.6	227.0	105.1	162.0	197.4
C28	174.0	59.3	46.7	59.3	30.1	278.3	55.0	123.1	89.7	104.7	58.5	54.9
C29	83.1	93.3	144.4	93.3	122.6	214.5	141.7	197.4	160.4	51.8	156.3	141.7
C30	191.2	73.8	53.3	73.8	24.6	293.2	66.8	102.5	102.5	130.6	34.3	66.8
C31	104.8	49.4	135.7	49.3	78.9	207.0	141.7	124.7	173.2	91.2	104.5	141.7
C32	150.5	37.5	75.3	37.5	47.4	248.1	78.9	132.4	108.8	79.4	80.7	78.7
C33	230.0	136.6	181.8	136.5	121.5	283.5	196.3	31.7	231.9	214.1	101.6	196.3
C34	30.1	132.2	226.4	132.3	180.7	107.3	226.8	221.3	250.8	107.0	210.8	226.8
C35	221.6	117.8	150.9	117.7	93.7	289.9	165.8	10.1	201.0	195.7	70.2	165.7
C36	226.8	148.7	71.8	148.7	131.9	341.0	56.0	221.0	28.8	139.2	151.3	56.0
C37	182.4	63.6	95.5	63.6	31.1	270.0	108.7	64.4	144.7	138.6	27.7	108.6
C38	198.4	80.8	81.3	80.9	34.1	292.4	95.7	74.1	131.3	148.9	17.1	95.7
C39	188.6	71.5	83.1	71.6	27.4	282.6	96.8	74.6	132.8	140.6	21.8	96.7
C40	146.6	52.5	71.8	52.6	62.1	252.4	70.9	152.7	96.8	66.6	96.6	70.8
C41	301.1	227.1	153.1	227.1	213.8	401.5	140.2	295.1	108.9	210.5	228.7	140.1
C42	277.2	214.8	154.9	214.9	209.3	375.5	140.8	294.9	112.9	188.1	229.2	140.7
C43	229.1	200.6	165.0	200.7	207.1	335.7	149.7	300.2	128.8	149.4	235.2	149.6

	C25	C26	C27	C28	C29	C30	C31	C32	C33	C34	C35	C36
C25	0.0											
C26	213.9	0.0										
C27	170.1	257.2	0.0									
C28	46.5	175.0	148.2	0.0								
C29	152.5	172.5	103.0	113.8	0.0							
C30	20.8	200.3	158.2	30.2	132.8	0.0						
C31	111.4	225.7	66.9	91.2	79.1	97.2	0.0					
C32	75.0	170.7	126.2	36.4	101.6	62.2	81.5	0.0				
C33	122.5	324.8	167.2	151.1	214.1	131.9	137.3	157.0	0.0			
C34	215.5	253.2	58.3	184.8	107.8	201.1	112.6	158.5	225.1	0.0		
C35	91.3	297.5	163.7	122.7	194.6	101.0	121.9	133.6	33.3	220.7	0.0	
C36	131.0	93.2	227.1	102.7	152.9	120.7	178.1	114.7	250.8	244.4	220.8	0.0
C37	43.5	234.6	136.1	60.4	141.6	44.5	81.7	72.4	91.2	186.6	63.7	161.4
C38	24.7	230.4	156.4	57.3	148.9	30.7	97.7	81.1	103.2	205.3	72.0	150.5
C39	30.1	227.7	146.5	53.6	140.6	30.7	87.9	74.5	102.6	195.3	72.6	150.5
C40	86.9	150.1	132.6	41.9	88.9	71.1	88.4	24.1	177.4	158.4	153.0	98.1
C41	209.3	98.9	305.4	186.0	236.3	204.2	264.3	190.2	325.4	315.9	297.0	96.8
C42	210.6	71.2	286.8	180.8	216.6	203.0	251.2	179.4	324.1	292.0	296.8	94.0
C43	218.2	17.7	254.4	178.3	170.3	204.2	225.5	172.3	327.1	248.4	300.5	100.9

	C37	C38	C39	C40	C41	C42	C43
C37	0.0						
C38	27.0	0.0					
C39	20.2	10.1	0.0				
C40	90.2	94.6	88.5	0.0			
C41	241.1	232.1	233.1	177.4	0.0		
C42	237.9	231.7	231.2	164.5	35.2	0.0	
C43	237.2	233.9	230.8	151.2	108.2	78.7	0

(b) Complete linkage produces results similar to single linkage.





12.15. In K-means method, we use the means of the clusters identified by average linkage as the initial cluster centers.

Final cluster centers for K = 4

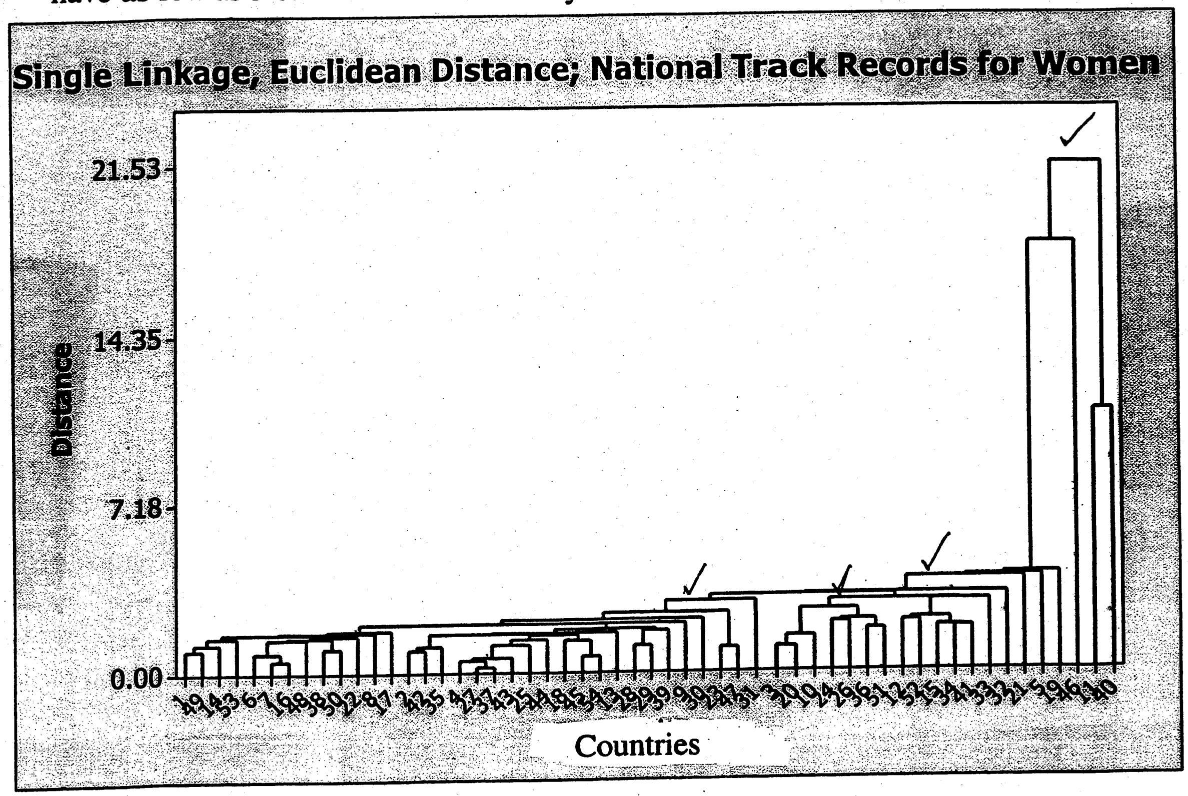
	1	2	3	4	5	6	7	8
1	110.0	2.1	0.9	215.0	0.7	15.3	7.9	50.0
2	114.4	3.1	1.7	171.1	2.8	15.0	6.6	123.9
3	86.7	2.3	0.5	26.7	1.4	10.0	5.8	55.8
4	112.5	3.2	0.8	225.0	5.8	12.5	10.8	245.0

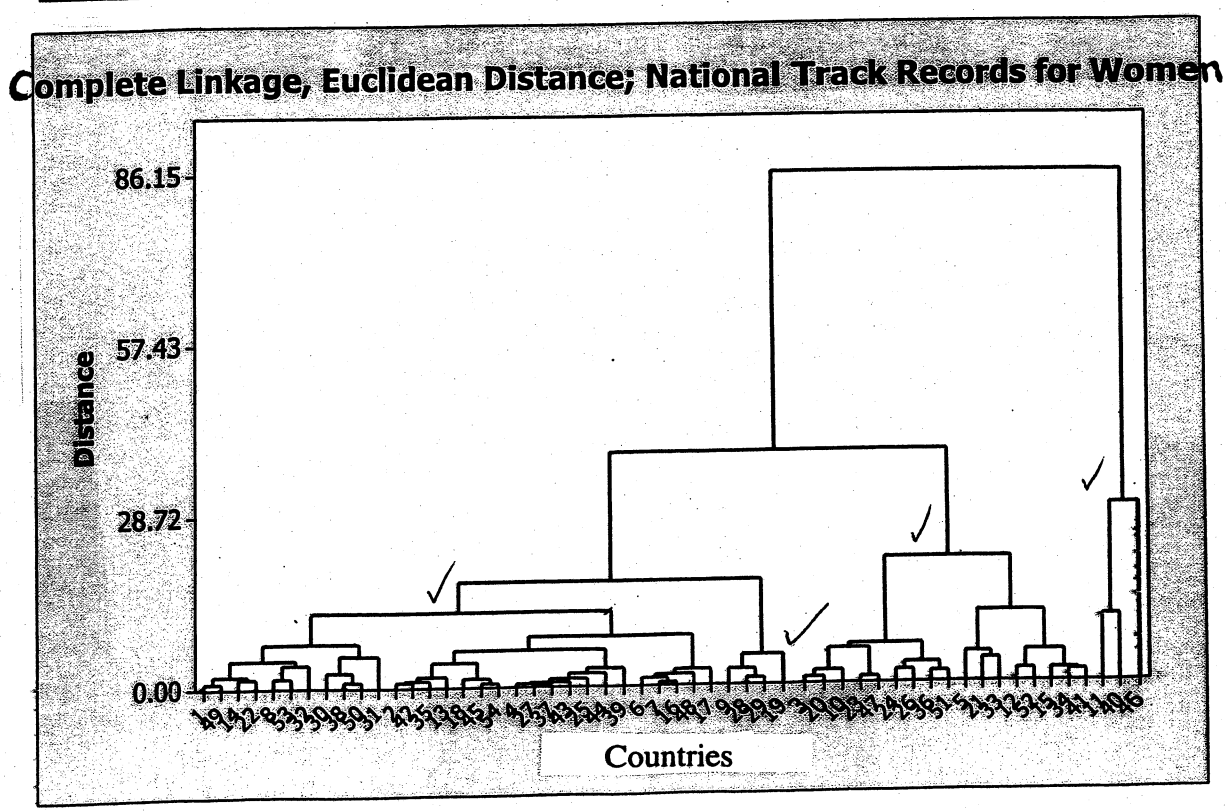
Distances between centers

	1	2	3	4
1	0.0			
2	86.1	0.0		
3	190.0	162.2	0.0	
4	195.4	132.7	275.4	0.0

	K-means						4 clusters			
	K = 2		K = 3		K = 4		Single		Complete	
1	C1	1	C1	1	C1	1	C1	1	C1	1
2	C2	1	C2	1	C2	1	C2	1	C2	1
3	C3	1	C3	1	C3	1	C3	1	C3	1
4	C4	1	C4	1	C4	1	C4	1	C4	1
5	C5	1	C5	1	C5	1	C5	1	C5	1
6	C6	1	C6	1	C6	1	C6	1	C6	1
7	C7	1	C7	1	C7	1	C7	1	C7	1
8	C8	1	C8	1	C8	1	C8	1	C8	1
9	C9	1	C9	1	C9	1	C9	1	C9	1
10	C10	1	C10	1	C12	1	C10	1	C10	1
11	C12	1	C12	1	C15	1	C11	1	C12	1
12	C14	1	C14	1	C17	1	C12	1	C14	1
13	C15	1	C15	1	C19	1	C13	1	C16	1
14	C16	1	C16	1	C20	1	C14	1	C17	1
15	C17	1	C17	1	C23	1	C15	1	C20	1
16	C19	1	C19	1	C24	1	C16	1	C23	1
17	C20	1	C20	1	C25	1	C17	1	C25	1
18	C21	1	C23	1	C28	1	C19	1	C28	1
19	C23	1	C24	1	C30	1	C20	1	C30	1
20	C24	1	C25	1	C33	1	C21	1	C31	1
21	C25	1	C28	1	C35	1	C22	1	C32	1
22	C26	1	C30	1	C37	1	C23	1	C33	1
23	C28	1	C31	1	C38	1	C24	1	C35	1
24	C30	1	C32	1	C39	1	C25	1	C37	1
25	C32	1	C33	1	C10	2	C27	1	C38	1
26	C33	1	C35	1	C11	2	C28	1	C39	1
27	C35	1	C37	1	C14	2	C29	1	C40	1
28	C36	1	C38	1	C16	2	C30	1	C11	11
29	C37	1	C39	1	C22	2	C31	1	C13	11
30	C38	1	C40	1	C29	2	C32	1	C22	11
31	C39	1	C21	2	C31	2	C33	1	C27	11
32	C40	1	C26	2	C32	2	C34	1	C29	11
33	C41	1	C36	2	C40	2	C35	1	C34	11
34	C42	1	C41	2	C21	3	C36	1	C15	15
35	C43	1	C42	2	C26	3	C37	1	C19	15
36	C11	2	C43	2	C36	3	C38	1	C21	15
37	C13	2	C11	3	C41	3	C39	1	C24	15
38	C18	2	C13	3	C42	3	C40	1	C26	15
39	C22	2	C18	3	C43	3	C18	18	C36	15
40	C27	2	C22	3	C13	4	C26	26	C41	15
41	C29	2	C27	3	C18	4	C43	26	C42	15
42	C31	2	C29	3	C27	4	C41	41	C43	15
43	C34	2	C34	3	C34	4	C42	41	C18	18

12.16 (a), (b) Dendrograms for single linkage and complete linkage follow. The dendrograms are similar; as examples, in both procedures, countries 11, 40 and 46 form a group at a relatively high level of distance, and countries 4, 27, 37, 43, 25 and 44 form a group at a relatively small distance. The clusters are more apparent in the complete linkage dendrogram and, depending on the distance level, might have as few as 3 or 4 clusters or as many as 6 or 7 clusters.





(c) The results for K = 4 and K = 6 clusters are displayed below. The results seem reasonable and are consistent with the results for the linkage procedures. Depending on use, K = 4 may be an adequate number of clusters.

Data Display

Country	ClustMemK=6	ClustMemK=4
1	6	2
2	2	4
3	1	2
4	4	4
5	3	1
6	6	2
7	6	2
8	1	2
9	4	4
10	1	2
11	5	3
12	3	1
13	2	4
14	6	2
15	3	1
16	6	2
17	6	2
18	2	4
19	4	4
20	1	2
21	3	1
22	6	2
23	1	2
24	1	1
25	4	4
26	1	2
27	4	4
28	4	4
29	4	4
30	6	2
31	6	2
32	6	2
33	3	1
34	3	1
35	2	4
36	1	1
37	4	4
38	6	2
39	4	4
40	5	3
41	3	1
42	2	4
43	4	4
44	2	4
45	2	4
46	5	3
47	1	2
48	6	4
49	6	2
50	6	4
51	1	1
52	3	1
53	6	2
54	2	4

Number of clusters: 4

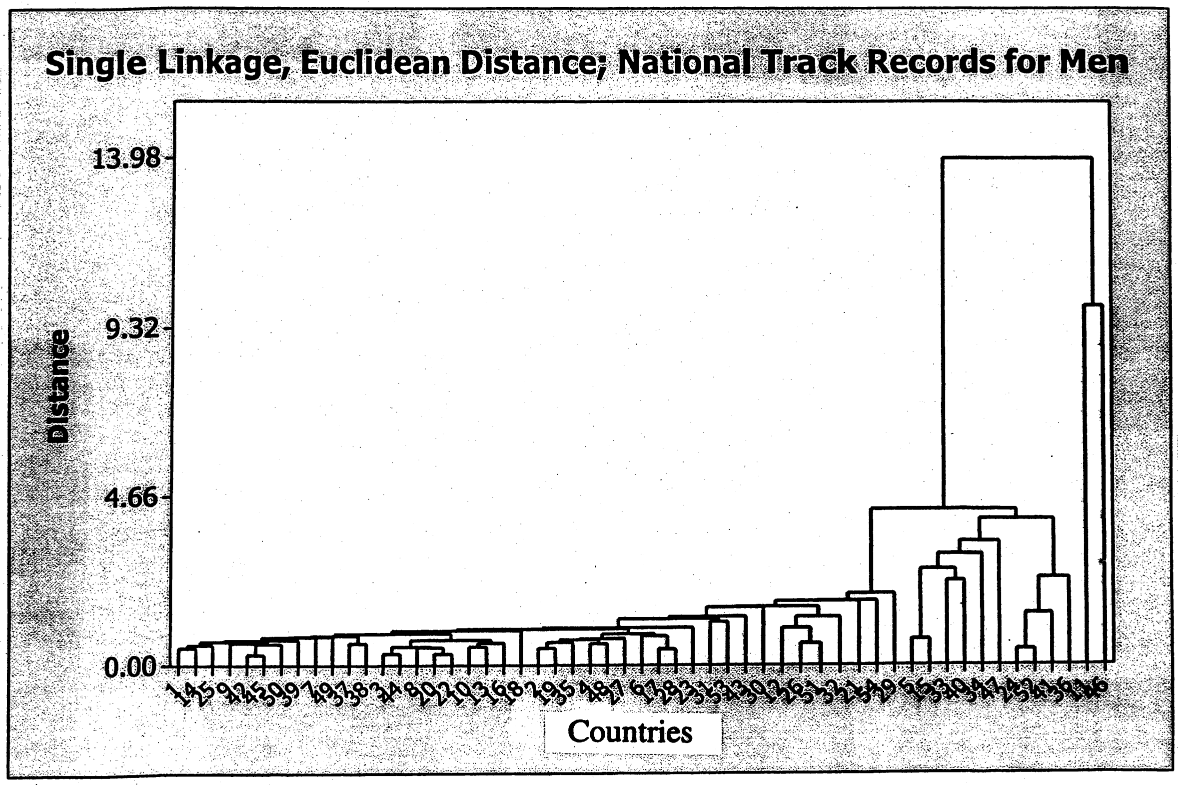
	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	11	298.660	4.494	9.049
Cluster2	20	318.294	3.613	6.800
✓Cluster3	3	490.251	11.895	16.915
Cluster4	20	182.870	2.681	7.024

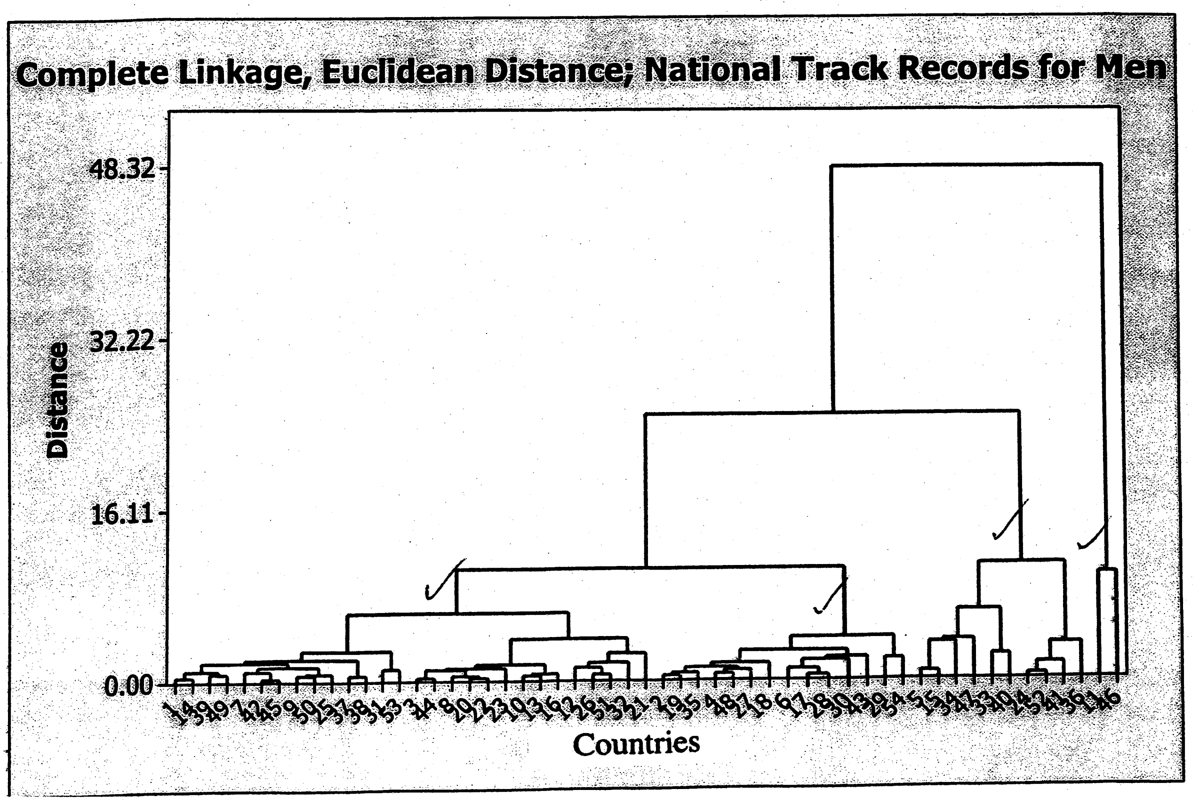
Number of clusters: 6

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	10	90.154	2.884	4.008
Cluster2	8	22.813	1.613	2.428
Cluster3	8	116.518	3.346	6.651
Cluster4	10	78.508	2.513	5.977
✓Cluster5	3	490.251	11.895	16.915
Cluster6	15	128.783	2.669	5.521

✓ Identical

12.17 (a), (b) Dendrograms for single linkage and complete linkage follow. The dendrograms are similar; as examples, in both procedures, countries 11 and 46 form a group at a relatively high level of distance, and countries 2, 19, 35, 4, 48 and 27 form a group at a relatively small distance. The clusters are more apparent in the complete linkage dendrogram and, depending on the distance level, might have as few as 3 or 4 clusters or as many as 6 or 7 clusters.





(c) The results for K = 4 and K = 6 clusters are displayed below. The results seem reasonable and are consistent with the results for the linkage procedures. Depending on use, K = 4 may be an adequate number of clusters. The results for the men are similar to the results for the women.

Data Display

Country	ClustMemK=4	ClustMemK=6
1	2	2
2	4	4
3	2	1
4	4	4
5	1	3
6	4	6
7	2	2
8	2	1
9	4	2
10	2	2
11	3	5
12	2	1
13	2	1
14	2	2
15	1	3
16	2	2
17	4	6
18	4	4
19	4	4
20	2	1
21	2	1
22	2	1
23	2	1
24	1	3
25	4	2
26	2	1
27	4	4
28	4	6
29	4	6
30	4	4
31	2	2
32	2	1
33	1	3
34	1	3
35	4	4
36	1	3
37	4	4
38	4	2
39	2	2
40	1	3
41	1	3
42	4	2
43	4	4
44	2	1
45	4	2
46	3	5
47	1	3
48	4	4
49	2	2
50	4	2
51	2	1
52	1	3
53	2	2
54	4	6

Number of clusters: 4

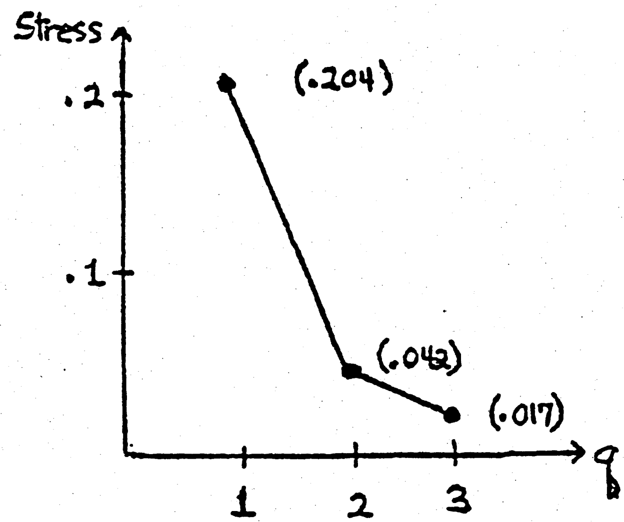
	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	10	169.042	3.910	5.950
Cluster2	21	73.281	1.684	3.041
✓Cluster3	2	49.174	4.959	4.959
Cluster4	21	56.295	1.481	3.249

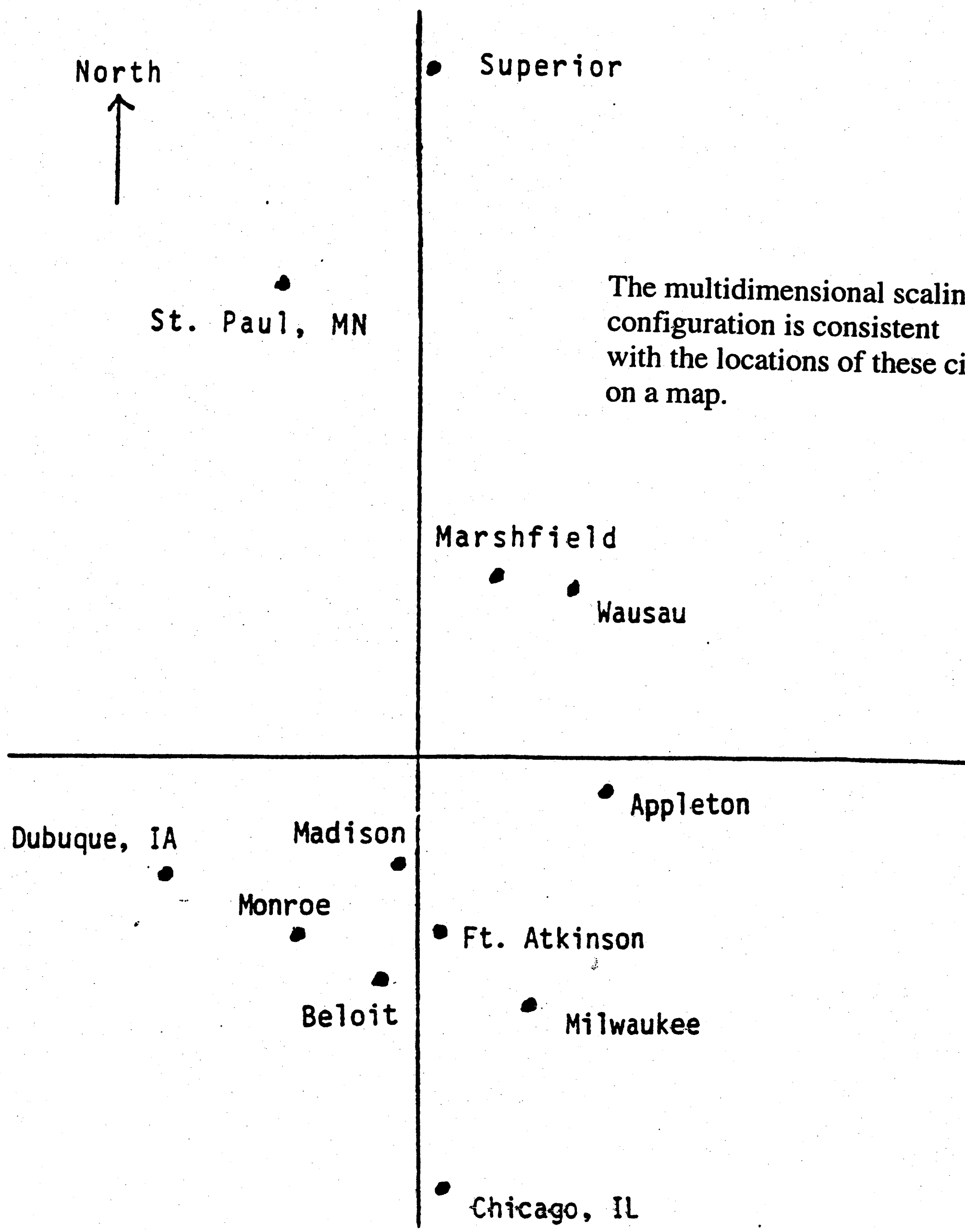
Number of clusters: 6

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	12	26.806	1.418	2.413
Cluster2	15	18.764	1.048	1.844
Cluster3	10	169.042	3.910	5.950
Cluster4	10	10.137	0.935	1.559
✓Cluster5	2	49.174	4.959	4.959
Cluster6	5	6.451	1.092	1.606

✓ Identical

12.18.





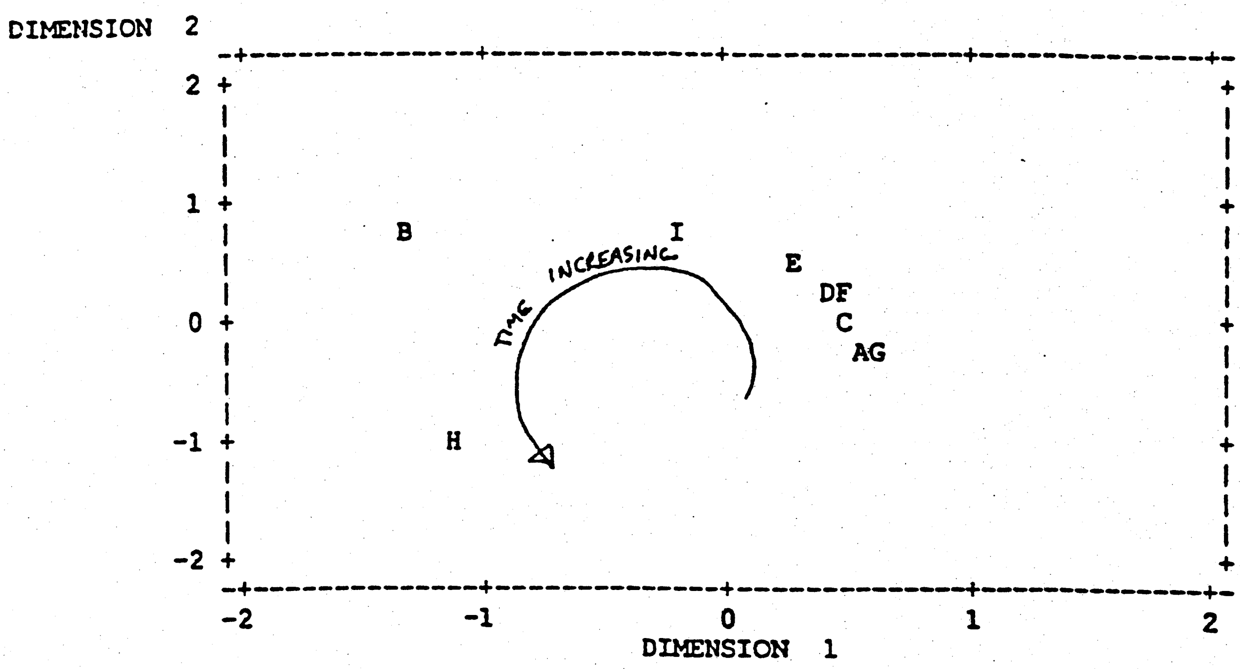
The multidimensional scaling configuration is consistent with the locations of these cities on a map.

12.19.

The stress of final configuration for q=5 is 0.000. The sites in 5 dimensions and the plot of the sites in two dimensions are

COORDINATES IN 5 DIMENSIONS

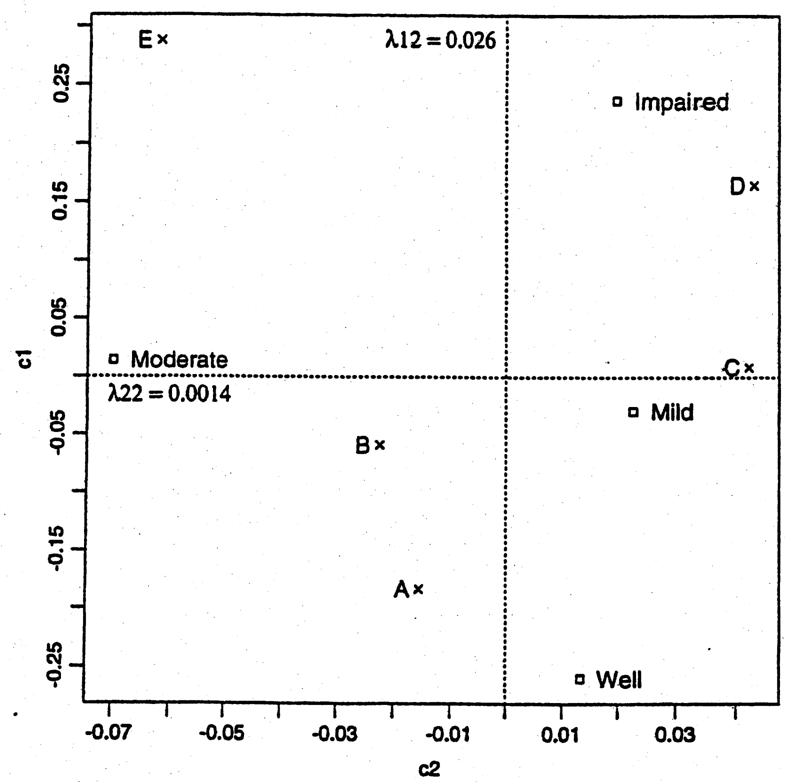
VARIABLE	PLOT	DIMENSION 1	2	3	4	5
P1980918	A	.51	-.28	.24	-.68	.12
P1931131	B	-1.32	.69	.62	-.05	-.02
P1550960	C	.47	-.07	.19	.30	.06
P1530987	D	.39	.09	.05	.34	.10
P1361024	E	.23	.30	-.32	.05	.12
P1351005	F	.47	.14	-.22	-.14	-.28
P1340945	G	.58	-.35	.46	.18	-.10
P1311137	H	-1.12	-1.12	-.31	.05	-.01
P1301062	I	-.22	.61	-.70	-.06	.01



The results show a definite time pattern (where time of site is frequently determined by C-14 and tree ring (lumber in great houses) dating).

12.20. A correspondence analysis of the mental health-socioeconomic data

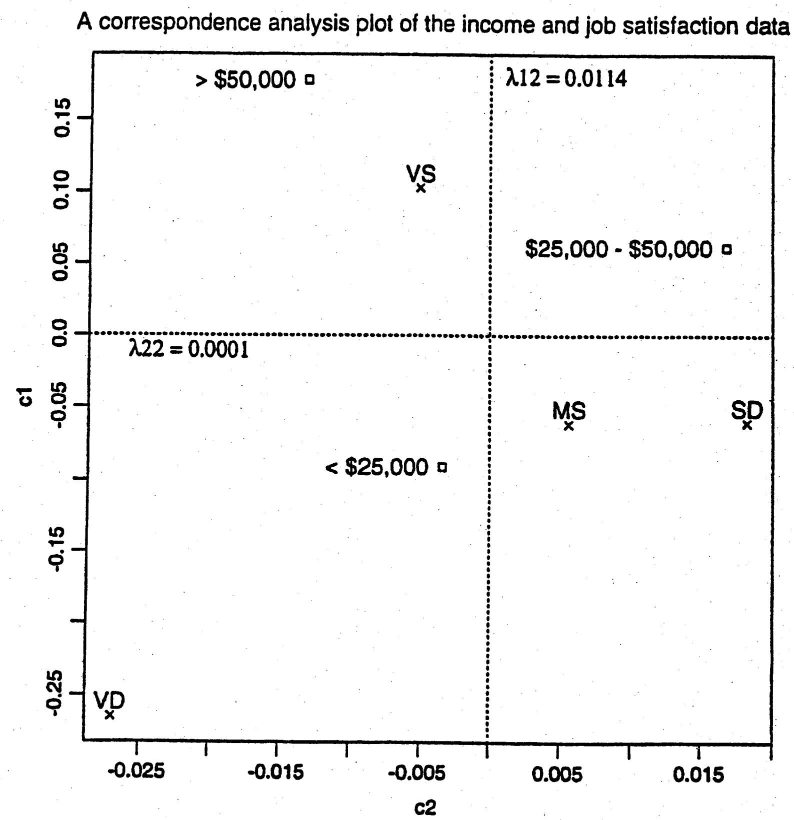
A correspondence analysis plot of the mental health-socioeconomic data




```
U                                      V
-0.6922  0.1539  0.5588 0.4300         -0.6266 -0.2313  0.0843 -0.3341
-0.1100  0.3665 -0.7007 0.6022         -0.1521 -0.2516 -0.5109 -0.6407
 0.0411 -0.8809 -0.0659 0.4670          0.0265  0.5490  0.5869 -0.5756
 0.7121  0.2570  0.4388 0.4841          0.4097  0.4668 -0.5519 -0.2297
                                        0.6448 -0.6032  0.2879 -0.3062

lambda
0.1613 0.0371 0.0082 0.0000

Cumulative inertia
0.0260 0.0274 0.0275

Cumulative proportion
0.9475 0.9976 1.0000
```

The lowest economic class is located between moderate and impaired. The next lowest class is closest to impaired.

12.21. A correspondence analysis of the income and job satisfaction data



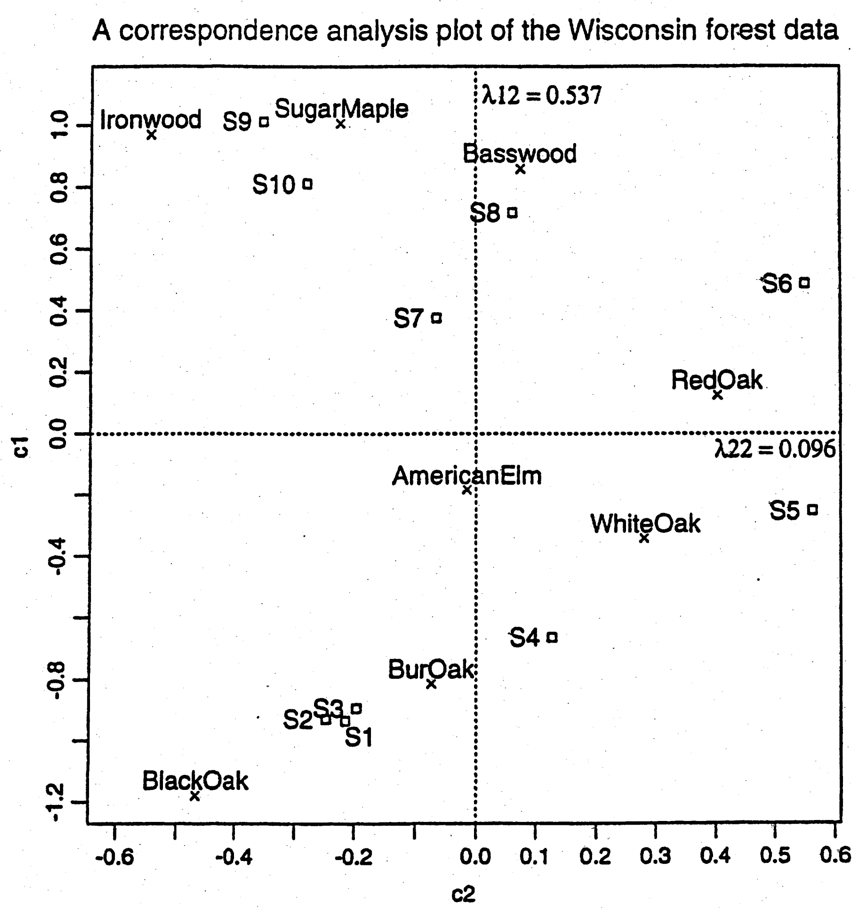

```
U
-0.6272 -0.2392 0.7412
 0.2956  0.8073 0.5107
 0.7206 -0.5394 0.4356
```

```
V
-0.6503 -0.6661 -0.3561
-0.1944  0.5933 -0.7758
-0.3400  0.3159  0.2253
 0.6510 -0.3233 -0.4696
```

```
lambda
0.1069 0.0106 0.0000

Cumulative inertia
0.0114 0.0116

Cumulative proportion
0.9902 1.0000
```

Very satisfied is closest to the highest income group, and very dissatisfied is below the lowest income group. Satisfaction appears to increase with income.

12.22. A correspondence analysis of the Wisconsin forest data



U

-0.3877	-0.2108	-0.0616	0.4029	-0.0582	0.3269	0.4247	-0.1590
-0.3856	-0.2428	-0.0106	0.4345	-0.1950	-0.1968	-0.2635	-0.3835
-0.3495	-0.1821	0.4079	-0.5718	0.2343	-0.1167	0.3294	-0.1272
-0.3006	0.1355	0.0540	-0.2646	0.0006	-0.0826	-0.6644	-0.3192
-0.1108	0.5817	-0.4856	-0.1598	-0.2333	0.1607	0.0772	-0.0518
0.2022	0.5400	0.4626	0.2687	-0.0978	-0.3943	0.2668	-0.3606
0.1852	-0.0756	-0.5090	-0.0291	0.6026	-0.1955	0.1520	-0.5154
0.3140	0.0644	0.3394	0.1567	0.3366	0.6573	-0.2507	-0.2267
0.4200	-0.3484	-0.0394	0.1165	-0.0625	-0.3772	-0.1456	0.1381
0.3549	-0.2897	-0.0345	-0.3393	-0.5994	0.2002	0.1262	-0.4907

V

-0.3904	-0.0831	-0.4781	0.4562	-0.0377	0.3369	0.4071	-0.3511
-0.5327	-0.4985	0.4080	0.0925	-0.0738	-0.3420	-0.2464	-0.3310
-0.1999	0.3889	0.4089	-0.3622	0.4391	0.3217	0.1808	-0.4260
0.0698	0.5382	-0.1726	0.3181	-0.0544	-0.1596	-0.6122	-0.4138
-0.0820	-0.0151	-0.4271	-0.7086	-0.4160	-0.1685	0.0307	-0.3258
0.4005	0.0831	0.1478	0.1866	-0.0042	-0.5895	0.5587	-0.3412
0.3634	-0.4850	-0.3232	-0.0937	0.6298	0.0164	-0.2172	-0.2745
0.4689	-0.2476	0.3150	0.0726	-0.4771	0.5142	-0.0763	-0.3412

lambda

0.7326 0.3101 0.2685 0.2134 0.1052 0.0674 0.0623 0.0000

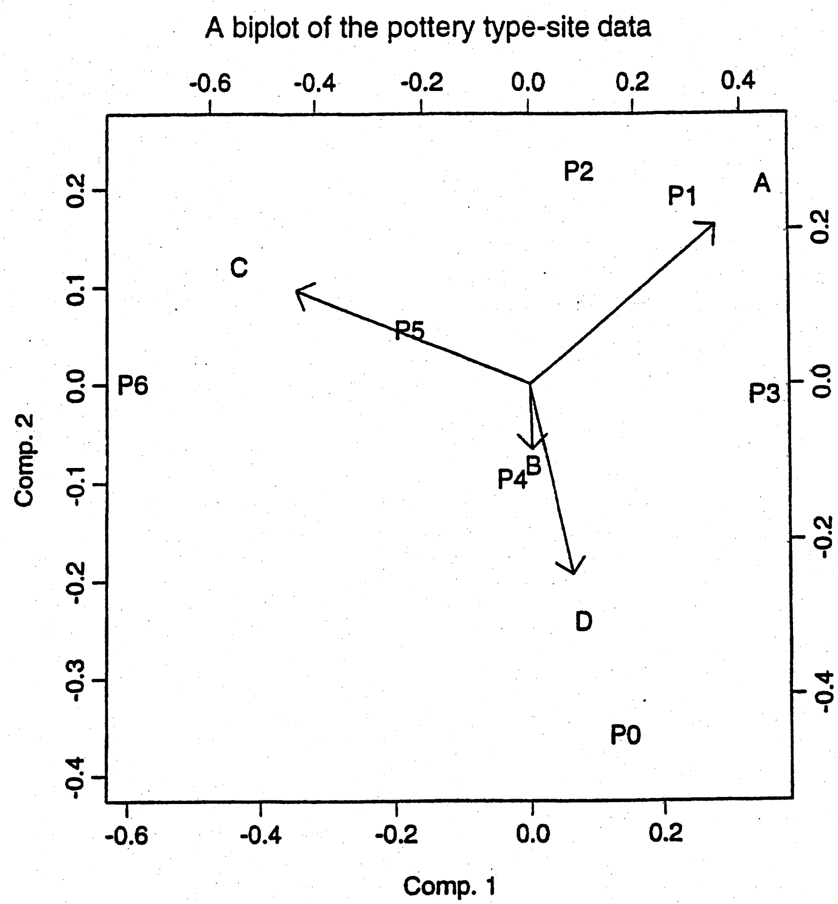
Cumulative inertia

0.5367 0.6329 0.7050 0.7506 0.7616 0.7662 0.7700

Cumulative proportion

0.6970 0.8219 0.9155 0.9747 0.9891 0.9950 1.0000

12.23. We construct biplot of the pottery type-site data, with row proportions as variables.




```
S
 0.0511 -0.0059 -0.0390 -0.0061
-0.0059  0.0084 -0.0051  0.0025
-0.0390 -0.0051  0.0628 -0.0187
-0.0061  0.0025 -0.0187  0.0223
```

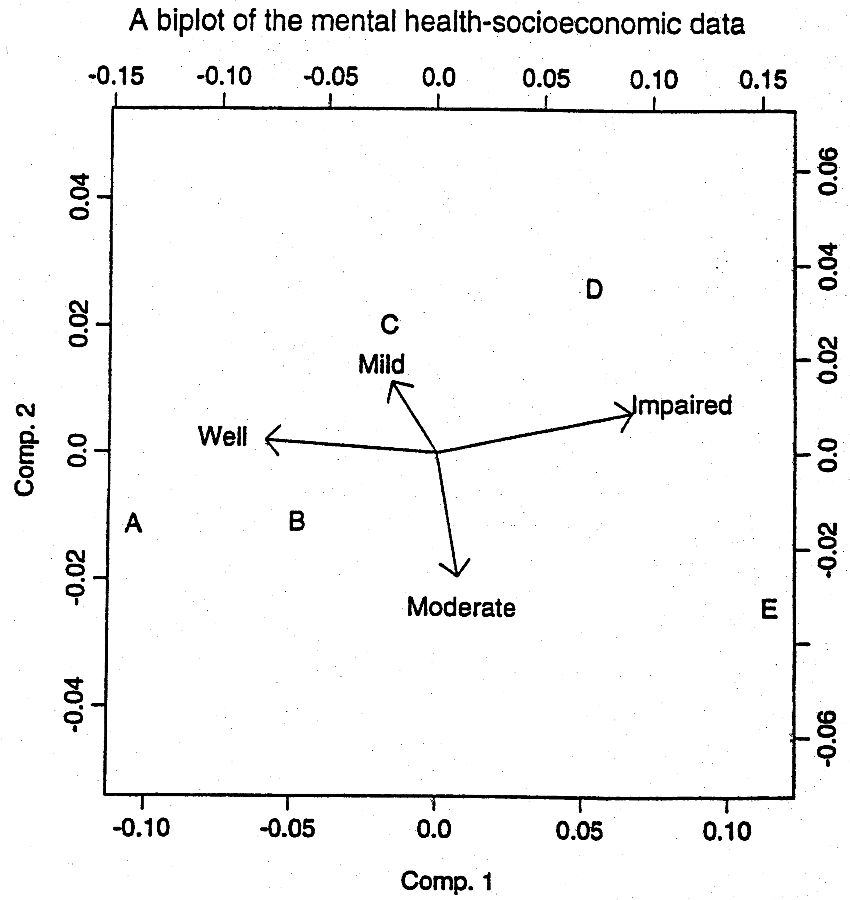
```
Eigenvectors of S
 0.6233  0.5853  0.1374 -0.5
 0.0064 -0.2385 -0.8325 -0.5
-0.7694  0.3464  0.1951 -0.5
 0.1396 -0.6932  0.5000 -0.5
```

```
Eigenvalues of S
0.0978 0.0376 0.0091 0.0000
```

	pc1	pc2	pc3	pc4
St. Dev.	0.3128	0.1940	0.0952	0
Prop. of Var.	0.6769	0.2604	0.0627	0
Cumulative Prop.	0.6769	0.9373	1.0000	1

As in the correspondence analysis.

12.24. We construct biplot of the mental health-socioeconomic data, with column proportions as variables.



S

0.003089	0.000809	-0.000413	-0.003485
0.000809	0.000329	-0.000284	-0.000853
-0.000413	-0.000284	0.000379	0.000318
-0.003485	-0.000853	0.000318	0.004021

Eigenvectors of S

-0.6487	0.0837	-0.5676	0.5
-0.1685	0.4764	0.7033	0.5
0.0794	-0.8320	0.2270	0.5
0.7379	0.2719	-0.3628	0.5

Eigenvalues of S

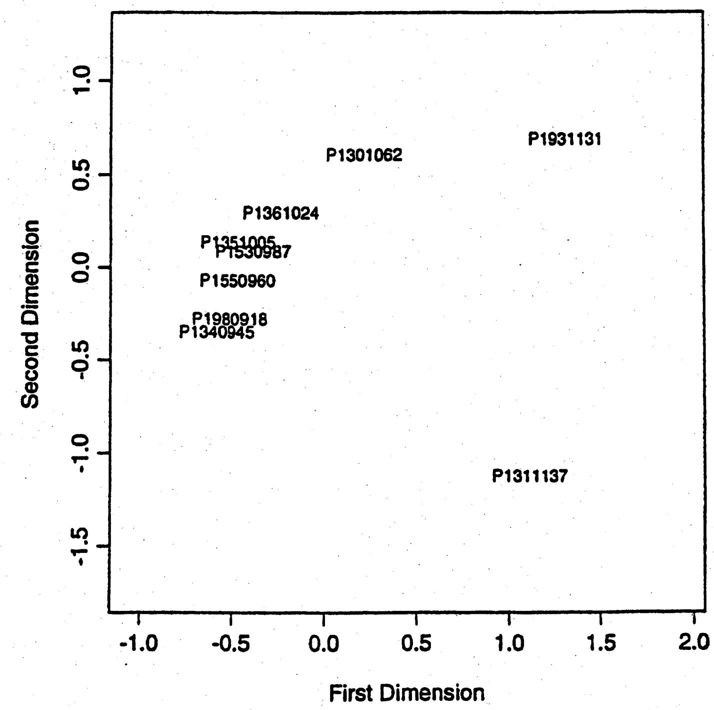
0.007314	0.000480	0.000024	0.000000

	pc1	pc2	pc3	pc4
St. Dev.	0.0855	0.0219	0.0049	0
Prop. of Var.	0.9355	0.0614	0.0031	0
Cumulative Prop.	0.9355	0.9969	1.0000	1

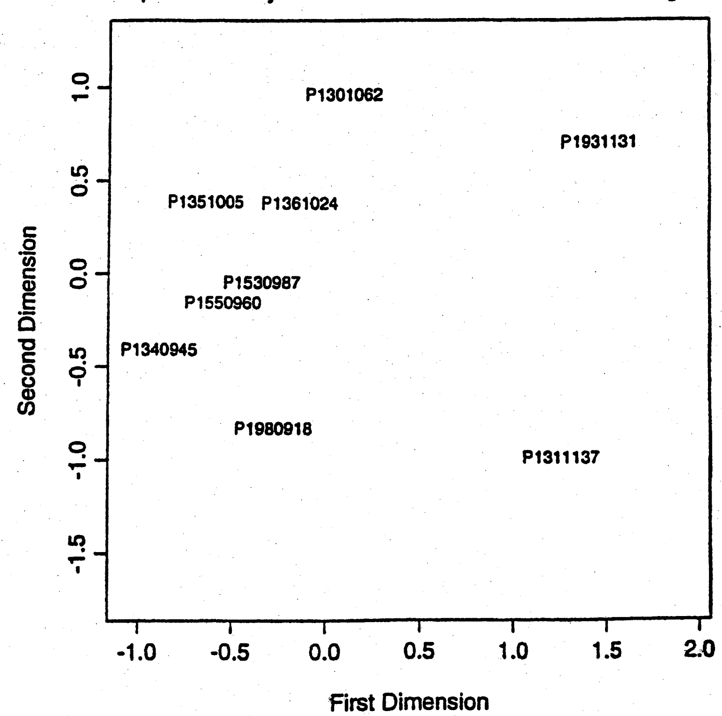
The biplot gives similar locations for health and socioeconomic status. A reflection about the 45 degree line would make them appear more alike.

12.25. A Procrustes analysis of archaeological data

A two-dimensional representation of archaeological sites produced by metric multidimensional scaling



A two-dimensional representation of archaeological sites produced by nonmetric multidimensional scaling



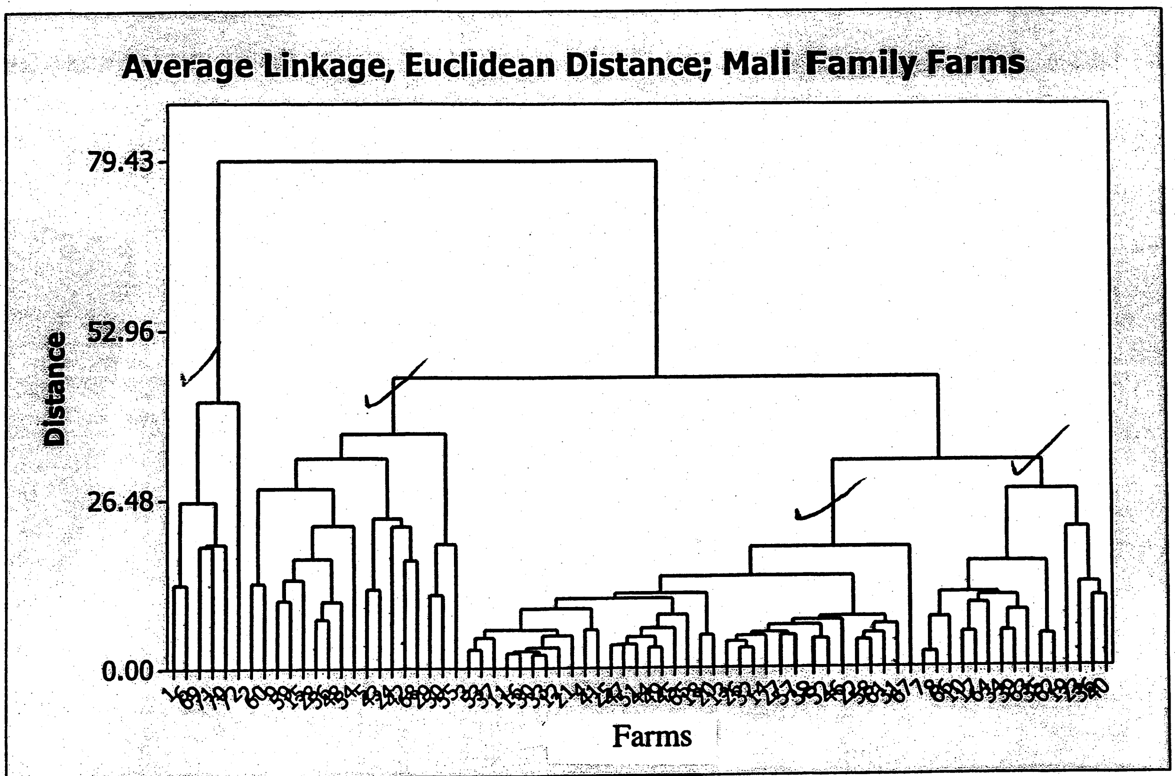
Site	Metric MDS		Nonmetric MDS	
P1980918	-0.512	-0.278	-0.276	-0.829
P1931131	1.318	0.692	1.469	0.703
P1550960	-0.470	-0.071	-0.545	-0.156
P1530987	-0.387	0.088	-0.338	-0.048
P1361024	-0.234	0.296	-0.137	0.379
P1351005	-0.469	0.137	-0.642	0.387
P1340945	-0.581	-0.349	-0.889	-0.409
P1311137	1.118	-1.122	1.262	-0.989
P1301062	0.216	0.608	0.096	0.963

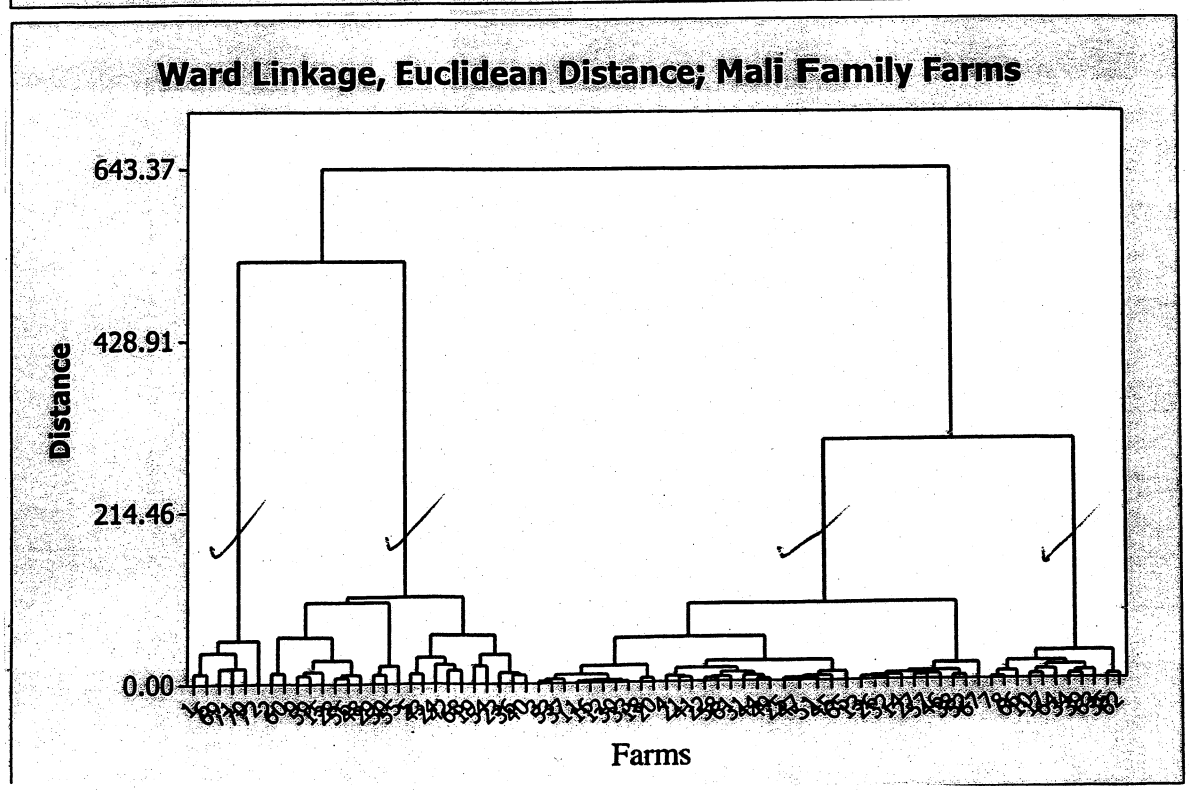
U		V	
-0.9893	-0.1459	-0.9977	-0.0679
-0.1459	0.9893	-0.0679	0.9977

Q		Lambda	
0.9969	0.0784	4.7819	0.000
-0.0784	0.9969	0.0000	2.715

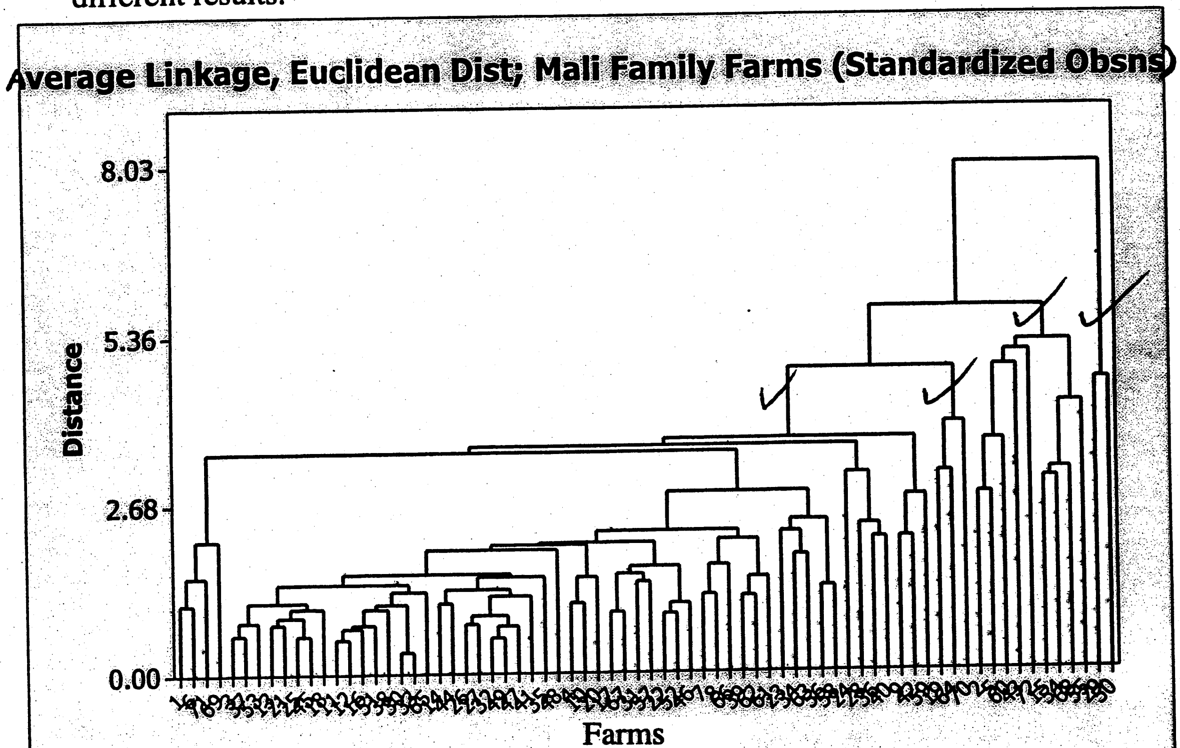
To better align the metric and nonmetric solutions, we multiply the nonmetric scaling solution by the orthogonal matrix $\hat{\mathbf{Q}}$. This corresponds to clockwise rotation of the nonmetric solution by 4.5 degrees. After rotation, the sum of squared distances, 0.803, is reduced to the Procrustes measure of fit $PR^2 = 0.756$.

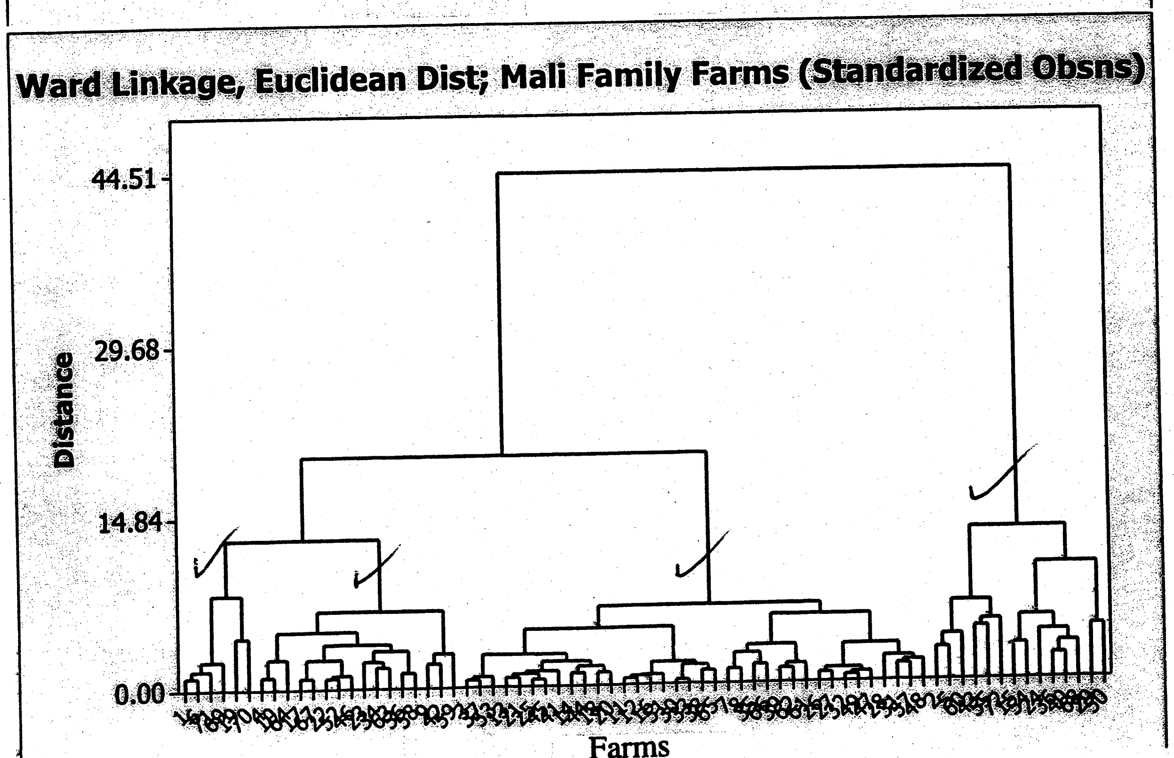
12.26 The dendrograms for clustering Mali Family Farms are given below for average linkage and Ward's method. The dendrograms are similar but a moderate number of distinct clusters is more apparent in the Ward's method dendrogram than the average linkage dendrogram. Both dendrograms suggest there may be as few as 4 clusters (indicated by the checkmarks in the figures) or perhaps as many as 7 or 8 clusters. Reading the "right" number of clusters from either dendrogram would depend on the use and require some subject matter knowledge.





12.27 If average linkage and Ward's method clustering is used with the standardized Mali Family Farm observations, the results are somewhat different from those using the original observations and different from one another. The dendrograms follow. There could be as few as 4 clusters (indicated by the checkmarks in the figures) or there could be as many as 8 or 9 clusters or more. The distinct clusters are more clearly delineated in the Ward's method dendrogram and if we focus attention on the 4 marked clusters, we see the two procedures produce quite different results.





12.28 The results for K = 5 and K = 6 clusters follow. The results seem reasonable and are similar to the results for Ward's method considered in Exercise 12.26. Note as the number of clusters increases from 5 to 6, cluster 1 in the K = 5 solution is partitioned into two clusters, 1 and 6, in the K = 6 solution, there is no change in the other clusters. Although not shown, K = 4 is a reasonable solution as well.

Data Display

Farm	ClustMemK=5	ClustMemK=6
1	1	1
2	2	2
3	3	3
4	3	3
5	5	5
6	1	6
7	4	4
8	4	4
9	2	2
10	4	4
11	3	3
12	3	3
13	3	3
14	3	3
15	2	2
16	3	3
17	4	4
18	3	3
19	5	5
20	3	3
21	3	3
22	3	3
23	5	5
24	5	5
25	3	3
26	3	3
27	3	3
28	3	3
29	2	2
30	3	3
31	3	3
32	3	3
33	3	3
34	4	4
35	4	4
36	5	5
37	3	3
38	3	3
39	4	4
40	5	5
41	3	3
42	5	5
43	5	5
44	3	3
45	3	3
46	3	3
47	3	3
48	2	2
49	3	3
50	2	2
51	3	3
52	3	3
53	3	3
54	2	2
55	2	2
56	3	3
57	3	3
58	3	3
59	2	2
60	2	2
61	3	3
62	4	4
63	4	4
64	4	4
65	3	3
66	4	4
67	4	4
68	2	2
69	1	1
70	1	1
71	1	1
72	1	6

Number of clusters: 5

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	6	2431.094	18.498	33.076
Cluster2	11	4440.330	19.511	24.647
Cluster3	35	3298.539	8.878	21.053
Cluster4	12	1129.083	9.072	16.024
Cluster5	8	1943.156	15.030	19.619

Number of clusters: 6

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	4	696.609	13.005	15.474
Cluster2	11	4440.330	19.511	24.647
Cluster3	35	3298.539	8.878	21.053
Cluster4	12	1129.083	9.072	16.024
Cluster5	8	1943.156	15.030	19.619
Cluster6	2	1005.125	22.418	22.418

✓ Identical for two choices of K

12.29 The results for K = 5 and K = 6 clusters follow. The results seem reasonable and are similar to the results for Ward's method considered in Exercise 12.27. Note as the number of clusters increases from 5 to 6, clusters 3 and 4 in the K = 5 solution lose 1 and 2 farms respectively to form cluster 6 in the K = 6 solution, there is no change in the other clusters. These results using standardized observations are somewhat different from the corresponding results using the original data. It makes a difference whether standardized or un-standardized observations are used.

Data Display

Farm	SdClusMemK=5	SdClusMemK=6
1	1	1
2	5	5
3	3	3
4	3	3
5	5	5
6	1	1
7	3	3
8	3	3
9	5	5
10	3	4
11	3	3
12	3	3
13	3	3
14	3	3
15	4	4
16	3	3
17	4	4
18	3	3
19	2	2
20	3	3
21	3	3
22	4	4
23	4	4
24	4	6
25	3	3
26	3	3
27	3	3
28	4	4
29	2	2
30	3	3
31	4	4
32	3	3
33	3	3
34	4	4
35	3	3
36	4	6
37	3	3
38	3	3
39	3	3
40	4	6
41	3	3
42	4	4
43	5	5
44	3	3
45	3	3
46	4	4
47	3	3
48	3	3
49	3	3
50	2	2
51	3	3
52	4	4
53	3	3
54	2	2
55	5	5
56	4	4
57	4	4
58	4	4
59	4	4
60	5	5
61	3	3
62	3	3
63	4	4
64	4	4
65	3	3
66	3	3
67	4	4
68	2	2
69	1	1
70	1	1
71	1	1
72	5	5

Number of clusters: 5

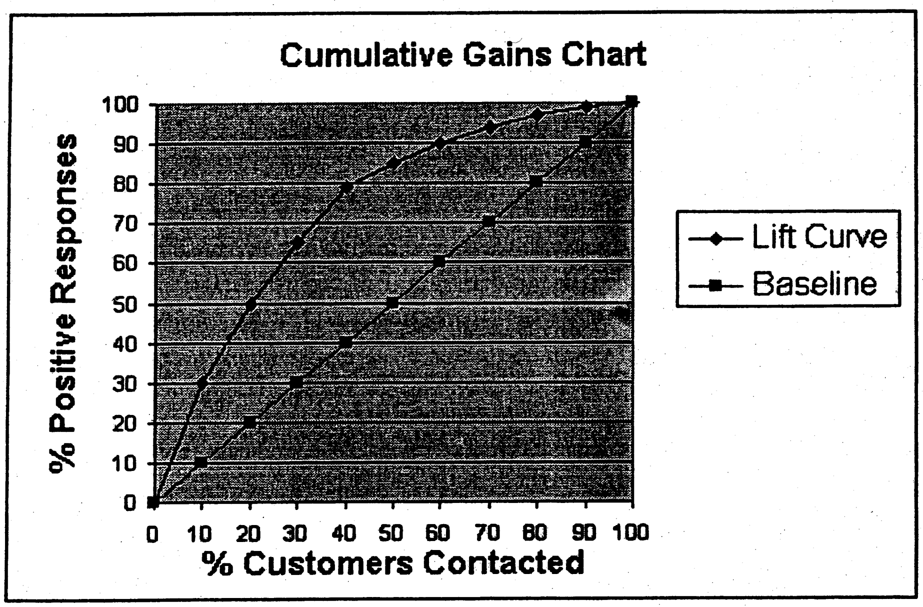
	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
✓Cluster1	5	14.050	1.568	2.703
✓Cluster2	5	56.727	3.288	4.259
Cluster3	35	55.318	1.211	1.993
Cluster4	20	84.099	1.954	3.172
✓Cluster5	7	63.071	2.970	3.482

Number of clusters: 6

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
✓Cluster1	5	14.050	1.568	2.703
✓Cluster2	5	56.727	3.288	4.259
Cluster3	34	51.228	1.183	1.951
Cluster4	18	65.501	1.806	3.195
✓Cluster5	7	63.071	2.970	3.482
Cluster6	3	7.960	1.604	1.954

✓ Identical for two choices of K

12.30 The cumulative lift (gains) chart is shown below. The y-axis shows the percentage of positive responses. This is the percentage of the total possible positive responses (20,000). The x-axis shows the percentage of customers contacted, which is a fraction of the 100,000 total customers. With no model, if we contact 10% of the customers we would expect 10%, or $2{,}000 = .1\times 20{,}000$, of the positive responses. Our response model predicts 6,000 or 30% of the positive responses if we contact the top 10,000 customers. Consequently, the y-values at $x = 10\%$ shown in the chart are 10% for baseline (no model) and 30% for the gain (lift) provided by the model. Continuing this argument for other choices of x (% customers contacted) and cumulating the results produces the lift (gains) chart shown. We see, for example, if we contact the top 40% of the customers determined by the model, we expect to get 80% of the positive responses.



12.31 **(a)** The Mclust function, which selects the best overall model according to the BIC criterion, selects a mixture with four multivariate normal components. The four estimated centers are:

$$\hat{\underset{\sim}{\mu}}_1 = \begin{bmatrix} 3.3188 \\ 6.7044 \\ 0.3526 \\ 0.1418 \\ 11.9742 \end{bmatrix} \quad \hat{\underset{\sim}{\mu}}_2 = \begin{bmatrix} 5.1806 \\ 5.2871 \\ 0.5910 \\ 0.1794 \\ 5.5369 \end{bmatrix}, \quad \hat{\underset{\sim}{\mu}}_3 = \begin{bmatrix} 7.2454 \\ 4.8099 \\ 0.3290 \\ 0.2431 \\ 3.2834 \end{bmatrix}, \quad \hat{\underset{\sim}{\mu}}_4 = \begin{bmatrix} 8.6893 \\ 4.1730 \\ 0.5158 \\ 0.2445 \\ 7.4846 \end{bmatrix}$$

and the estimated covariance matrices turn out to be restricted to be of the form $\eta_k\mathbf{D}$ where $\mathbf{D}$ is a diagonal matrix.

The estimated

$$\widehat{\mathbf{D}} = \text{diag}(11.2598, 2.7647, 0.3355, 0.0053, 18.0295)$$

and the estimated scale factors are $\hat{\eta}_1 = 0.0319$, $\hat{\eta}_2 = 0.3732$, $\hat{\eta}_3 = 0.0909$, $\hat{\eta}_4 = 0.1073$.

The estimated proportions are $\hat{p}_1 = 0.1059$, $\hat{p}_2 = 0.4986$, $\hat{p}_3 = 0.1322, \hat{p}_4 = 0.2633$.

This minimum BIC model has $BIC = -547.1408$.

(b) The model chosen above has 4 multivariate normal components. These four components are shown in the matrix scatter plot where the observations have been classified into one of the four populations.

The matrix scatter plot of the true classification, is given in the next figure.

Comparing the matrix scatter plot of the four group classification with the matrix scatter plot of the true classification, we see how the oil samples from the Upper sandstone are essentially split into two groups. This is clear from comparing the two scatter plots for (x_1, x_2).

We also repeat the analysis using the *me* function to select mixture distribution with $K = 3$ components. We further restrict the covariance matrices to satisfy $\Sigma_k = \eta_k\mathbf{D}$. The $K = 3$ groups selected by this function have estimated centers

$$\hat{\underset{\sim}{\mu}}_1 = \begin{bmatrix} 5.3395 \\ 5.2467 \\ 0.5485 \\ 0.1862 \\ 5.2465 \end{bmatrix}, \quad \hat{\underset{\sim}{\mu}}_2 = \begin{bmatrix} 8.5343 \\ 4.2762 \\ 0.4988 \\ 0.2453 \\ 6.6993 \end{bmatrix}, \quad \hat{\underset{\sim}{\mu}}_3 = \begin{bmatrix} 3.3228 \\ 6.7093 \\ 0.3511 \\ 0.1418 \\ 11.9780 \end{bmatrix},$$

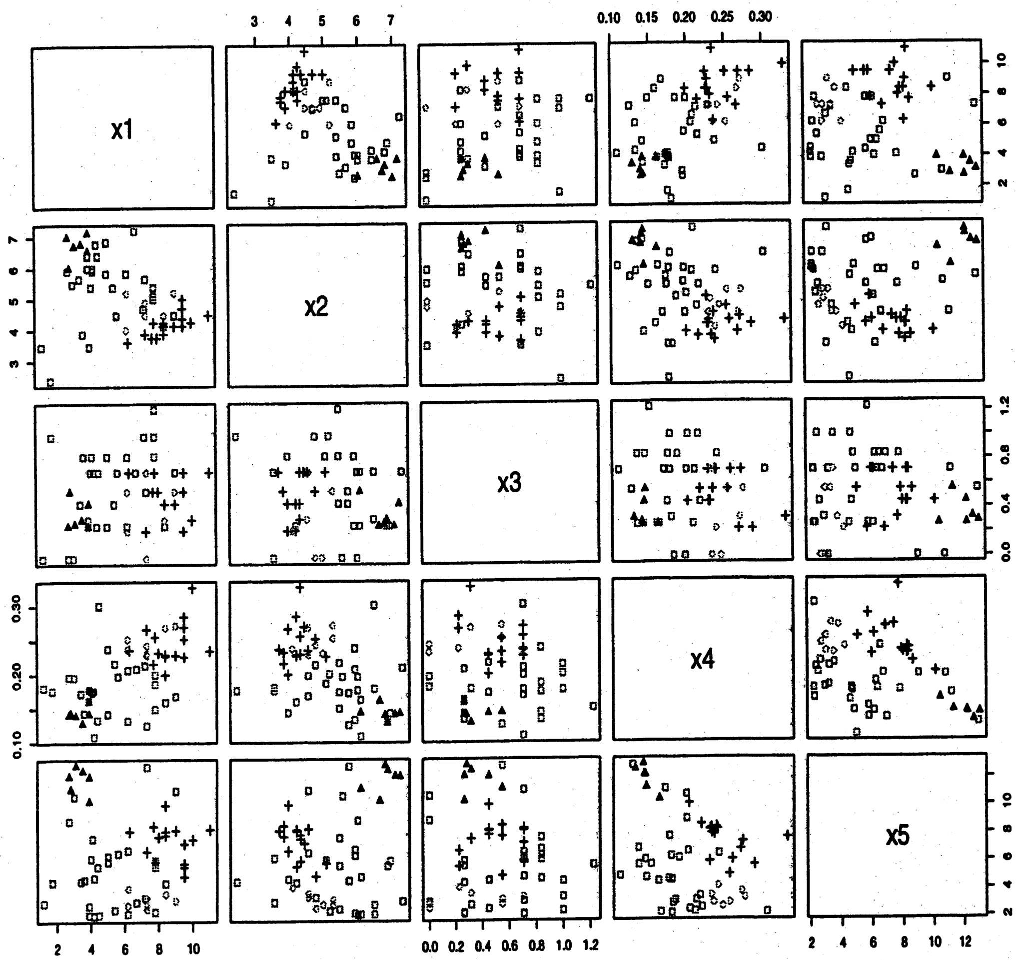


Figure 1: Classification into four groups using Mclust

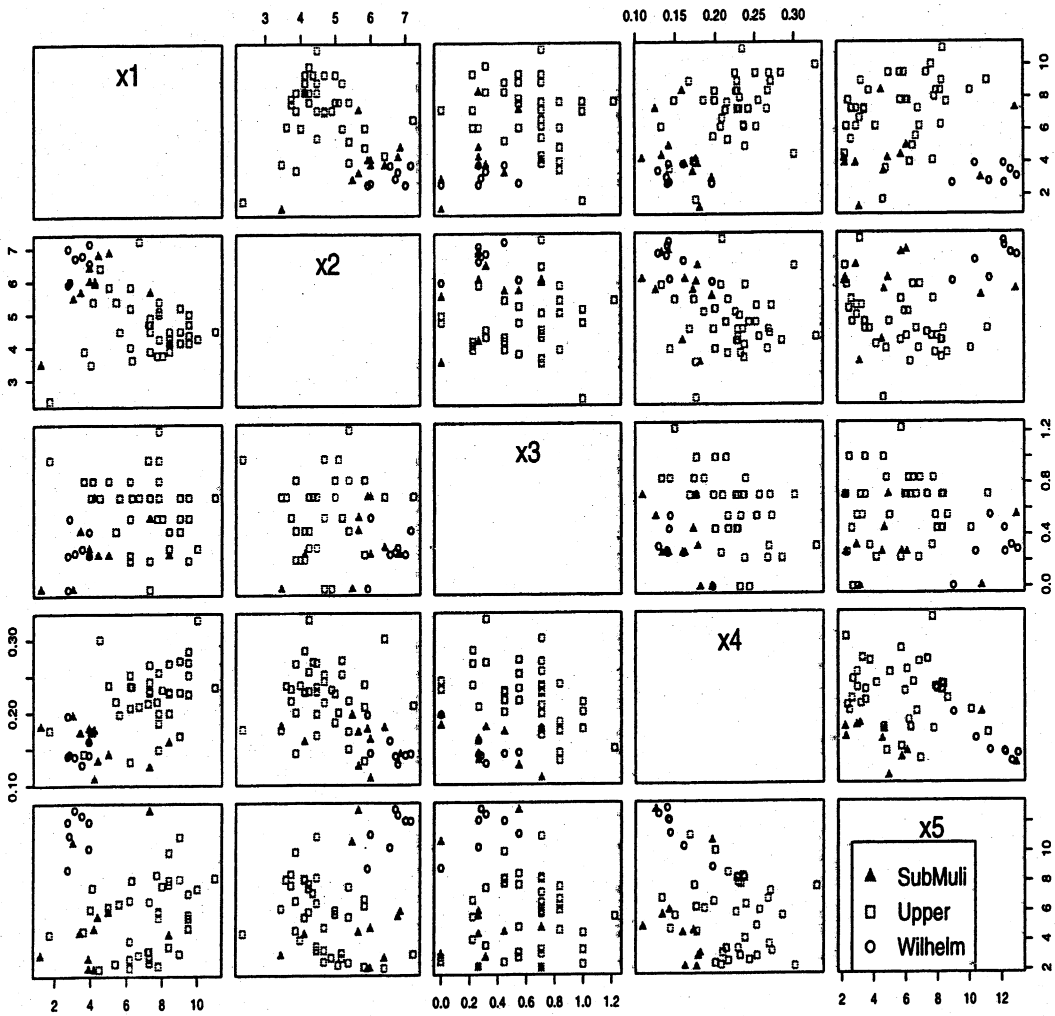


Figure 2: True classification into sandstone strata

the estimated diagonal matrix

$$\widehat{\mathbf{D}} = \operatorname{diag}(10.1535, 2.6295, 0.2969, 0.0052, 24.0955)$$

with estimated scale parameters $\widehat{\eta}_1 = 0.3702$, $\widehat{\eta}_2 = 0.1315$, $\widehat{\eta}_3 = 0.0314$, with resulting $BIC = -534.0949$.

The estimated proportions are $\widehat{p}_1 = 0.5651$, $\widehat{p}_2 = 0.3296$, $\widehat{p}_3 = 0.1052$.

If we use this method to classify the oil samples, the following samples are misclassified:

7	19	22	25	26	27	28	29	30	31
32	33	34	35	39	44	45	46	49	

and the misclassification error rate is 33.93%.