

ÍNDEX

1	Arbres	2
2	Corredors	3
3	Pardals	5
4	Dinasties	7
5	Coleòpters	11
6	Partits polítics	14
7	Idiomes	16
8	Ciutats	18
9	Biplot	19
10	Fruïtes	20
11	Adjectius	22
12	Assignatures	25
13	Flors	28
14	Noms catalans	32

1. ARBRES

1.1. Dades. Els arbres són 28 sureres i les variables medeixen els dipòsits de suro (en centigrams) a cada una de les quatre direccions cardinals: N, E, S, W.

N	E	S	W	N	E	S	W
72	66	76	77	91	79	100	75
60	53	66	63	56	68	47	50
56	57	64	58	79	65	70	61
41	29	36	38	81	80	68	58
32	32	35	36	78	55	67	60
30	35	34	26	46	38	37	38
39	39	31	27	39	35	34	37
42	43	31	25	32	30	67	32
37	40	31	25	60	50	48	54
33	29	27	36	35	37	39	39
32	30	34	28	39	36	39	31
63	45	74	63	50	34	37	40
54	46	60	52	43	37	39	50
47	51	52	43	48	54	57	43

1.2. Mitjanes, covariàncies i correlacions.

- Vector de mitjanes: (50.536, 46.179, 49.679, 45.179)
- Matriu de covariàncies:

$$S = \begin{pmatrix} 280.03 & 215.76 & 278.13 & 218.19 \\ 215.76 & 212.07 & 220.88 & 165.25 \\ 278.13 & 220.88 & 337.50 & 250.27 \\ 218.19 & 165.25 & 250.27 & 217.93 \end{pmatrix}$$

- Matriu de correlacions R i coeficient de dependència $\eta^2 = 1 - |R|$:

$$R = \begin{pmatrix} 1 & 0.885 & 0.905 & 0.883 \\ .885 & 1 & 0.826 & 0.769 \\ .905 & .826 & 1 & 0.923 \\ .883 & .769 & .923 & 1 \end{pmatrix} \quad \eta^2 = 0.994$$

1.3. Variables compostes. Són les següents:

Contrast eix N-S amb eix E-W: $Y_1 = N + S - E - W$
 Contrast N amb S: $Y_2 = N - S$
 Contrast E-W: $Y_3 = E - W$

	Y_1	Y_2	Y_3
Mitjanes:	8.857	0.857	1.000
Variàncies:	124.1	61.27	99.5

1.4. Variables normalitzades.

$$Z_1 = (N + S - E - W)/2 \quad Z_2 = (N - S)/\sqrt{2} \quad Z_3 = (E - W)/\sqrt{2}$$

	Z_1	Z_2	Z_3
Mitjanes:	4.42	0.606	0.707
Variàncies:	31.03	30.63	49.75

1.5. Interpretació. Quan normalitzem les variables aconseguim que tinguin variàncies més homogènies. La principal direcció de variabilitat apareix al fer la comparació de l'eix N-S amb l'eix E-W.

2. CORREDORS

2.1. Dades. Temps parcials en minuts que 12 corredors triguen en recórrer 16 quilòmetres:

corredor	qm 4	qm 8	qm 12	qm16
1	10	10	13	12
2	12	12	14	15
3	11	10	14	13
4	9	9	11	11
5	8	8	9	8
6	8	9	10	9
7	10	10	8	9
8	11	12	10	9
9	14	13	11	11
10	12	12	12	10
11	13	13	11	11
12	14	15	14	13

2.2. Mètode. Anàlisi de components principals sobre les 4 columnes de la matriu de dades:

- X_1 = temps qms 1 a 4
- X_2 = temps qms 5 a 8
- X_3 = temps qms 9 a 12
- X_4 = temps qms 13 a 16

2.3. Resultats.

- Vector de mitjanes: (11.00, 11.08, 11.41, 10.92)

- Matrius de covariàncies i correlacions:

$$S = \begin{pmatrix} 4.364 & 4.091 & 2.091 & 2.273 \\ & 4.265 & 1.871 & 1.917 \\ & & 4.083 & 3.765 \\ & & & 4.265 \end{pmatrix} \quad R = \begin{pmatrix} 1 & .9483 & .4953 & .5268 \\ & 1 & .4484 & .4494 \\ & & 1 & .9022 \\ & & & 1 \end{pmatrix}$$

- Vectors i valors propis de S :

	t_1	t_2	t_3	t_4
	.5275	.4538	-.2018	-.6893
	.5000	.5176	.2093	.6621
	.4769	-.5147	.6905	-.1760
	.4943	-.5112	-.6624	.2357
λ	12.26	4.098	.4273	.1910
%	72.22	24.13	2.52	1.15
acum	72.22	96.35	98.85	100

- Components principals primera i segona:

$$\begin{aligned} Y_1 &= 0.527X_1 + 0.500X_2 + 0.477X_3 + 0.494X_4 & \text{var}(Y_1) &= 12.26 \\ Y_2 &= 0.453X_1 + 0.517X_2 - 0.514X_3 - 0.511X_4 & \text{var}(Y_2) &= 4.098 \end{aligned}$$

- Transformació per components principals:

$$\mathbf{Y} = \mathbf{XT}$$

on \mathbf{X} és la matriu de dades, \mathbf{T} és la matriu amb els vectors propis, \mathbf{Y} és la matriu amb els valors de les components principals sobre els 12 individus (coordenades principals).

2.4. Interpretació.

1. La primera component principal és gairebé proporcional a la suma dels temps parcials. Per tant, podem interpretar Y_1 com el temps que triguen o $-Y_1$ com la *rapidesa* en fer la cursa.
2. La segona component principal té coeficients positius a X_1, X_2 i coeficients negatius a X_3, X_4 . Un corredor amb valors alts a Y_2 vol dir que ha sigut lent al principi i més ràpid al final. Un corredor amb valors baixos a Y_2 vol dir que ha sigut ràpid al principi i més lent al final. Podem interpretar aquesta component com la *forma* de correr.

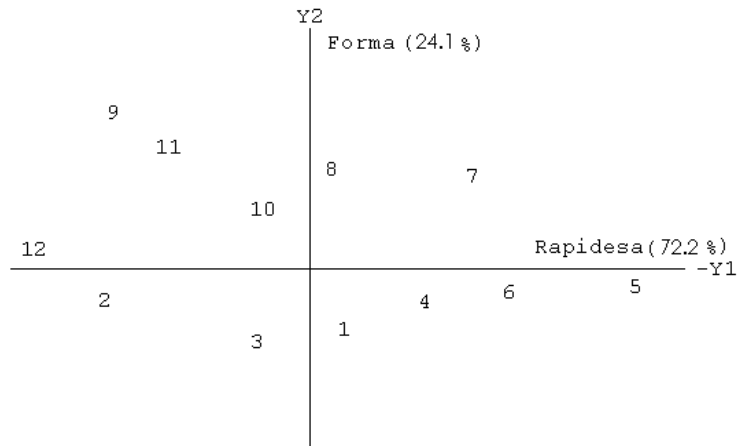


Figura 1: Representació de 12 corredors per anàlisi de components principals.

3. PARDALS

3.1. Dades. Mesures de 5 variables biomètriques sobre 49 pardals femelles, que varen ser recollides quasi moribundes després d'un temporal. Les 21 primeres sobrevisqueren, i les altres 28 no sobrevisqueren.

Sobreviuen					No sobreviuen				
X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5
					155	240	31.4	18.0	20.7
					156	240	31.5	18.2	20.6
					160	242	32.6	18.8	21.7
156	245	31.6	18.5	20.5	152	232	30.3	17.2	19.8
154	240	30.4	17.9	19.6	160	250	31.7	18.8	22.5
153	240	31.0	18.4	20.6	155	237	31.0	18.5	20.0
153	236	30.9	17.7	20.2	157	245	32.2	19.5	21.4
155	243	31.5	18.6	20.3	165	245	33.1	19.8	22.7
163	247	32.0	19.0	20.9	153	231	30.1	17.3	19.8
157	238	30.9	18.4	20.2	162	239	30.3	18.0	23.1
155	239	32.8	18.6	21.2	162	243	31.6	18.8	21.3
164	248	32.7	19.1	21.1	159	245	31.8	18.5	21.7
158	238	31.0	18.8	22.0	159	247	30.9	18.1	19.0
158	240	31.3	18.6	22.0	155	243	30.9	18.5	21.3
160	244	31.1	18.6	20.5	162	252	31.9	19.1	22.2
161	246	32.3	19.3	21.8	152	230	30.4	17.3	18.6
157	245	32.0	19.1	20.0	159	242	30.8	18.2	20.5
157	235	31.5	18.1	19.8	155	238	31.2	17.9	19.3
156	237	30.9	18.0	20.3	163	249	33.4	19.5	22.8
158	244	31.4	18.5	21.6	163	242	31.0	18.1	20.7
153	238	30.5	18.2	20.9	156	237	31.7	18.2	20.3
155	236	30.3	18.5	20.1	159	238	31.5	18.4	20.3
163	246	32.5	18.6	21.9	161	245	32.1	19.1	20.8
159	236	31.5	18.0	21.5	155	235	30.7	17.7	19.6
					162	247	31.9	19.1	20.4
					153	237	30.6	18.6	20.4
					162	245	32.5	18.5	21.1
					164	248	32.3	18.8	20.9

Les mesures (en mm.) són:

X_1 = long. total, X_2 = extensió ala, X_3 = long. bec i cap,

X_4 = long. húmer, X_5 = long. quilla estèrnum.

3.2. Mètode. Anàlisi de components principals sobre les 5 columnes de la matriu de dades de $n = 49$ files.

3.3. Resultats.

- Mitjanes de les variables:

158.0 241.3 31.46 18.47 20.83

- Matriu de covariàncies:

$$S = \begin{pmatrix} 13.35 & 13.61 & 1.922 & 1.331 & 2.192 \\ & 25.68 & 2.714 & 2.198 & 2.658 \\ & & .6316 & .3423 & .4146 \\ & & & .3184 & .3394 \\ & & & & .9828 \end{pmatrix}$$

- Tres primers vectors i valors propis:

	\mathbf{t}_1	\mathbf{t}_2	\mathbf{t}_3
	.5365	.8281	-.1565
	.8290	-.5505	-.0577
	.0965	.0335	.2375
	.0743	-.0146	.2032
	.1003	.0992	.9351
λ	35.33	4.622	.6309
%	86.22	11.28	1.540
acum	86.22	97.51	99.05

- Dues primeres components principals:

$$Y_1 = 0.54X_1 + 0.83X_2 + 0.09X_3 + 0.07X_4 + 0.10X_5$$

$$Y_2 = 0.83X_1 - 0.55X_2 + 0.03X_3 - 0.14X_4 + 0.10X_5$$

- Matriu de correlacions R i coeficient de dependència $\eta^2 = 1 - |R|$:

$$R = \begin{pmatrix} 1 & .7350 & .6618 & .6453 & .6051 \\ & 1 & .6737 & .7685 & .5290 \\ & & 1 & .7632 & .5263 \\ & & & 1 & .6066 \\ & & & & 1 \end{pmatrix} \quad \eta^2 = 0.9631$$

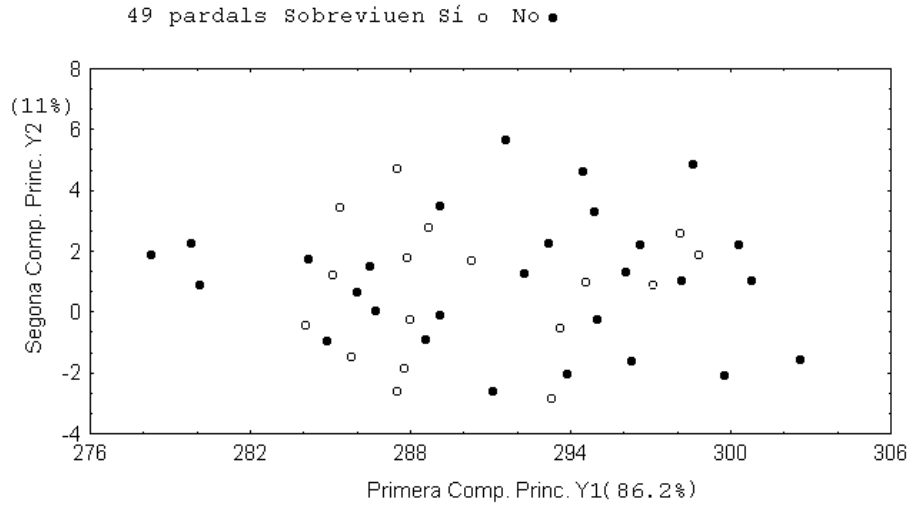


Figura 2: Representació de 49 pardals al llarg de dues dimensions de grandària i forma. Els pardals assenyalats amb ● són els que no sobreviuen.

3.4. Interpretació.

1. La primera component principal Y_1 té tots els coeficients positius. La interpretarem com un factor de *grandària*. Explica el 86.22 % de la variabilitat.
2. La segona component principal Y_2 té coeficients positius i negatius. La interpretarem com un factor de *forma*. Explica el 11.28 % de la variabilitat.
3. Grandària i forma són dimensions incorrelacionades i expliquen el 97.51% de la variabilitat de les dades.
4. Els pardals extrems tenen menys possibilitats de sobreviure.

4. DINASTIES

4.1. Dades. Mesures de 4 variables biomètriques sobre 150 cranis d'homes de l'Antic Egipte, corresponents als següents períodes:

- | | | |
|---|------------------------------|------------|
| 1 | Primeres dinasties (4000 aC) | $n_1 = 30$ |
| 2 | Dinasties tardanes (3300 aC) | $n_2 = 30$ |
| 3 | Dinasties 12 i 13 (1850 aC) | $n_3 = 30$ |
| 4 | Període Tolemaic (200 aC) | $n_4 = 30$ |
| 5 | Període Romà (150 dC) | $n_5 = 30$ |

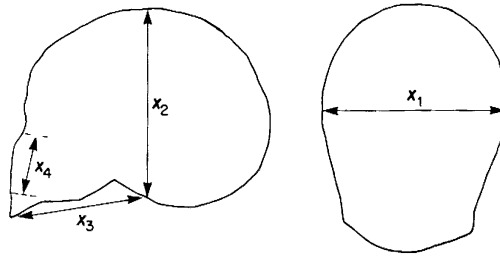


Figura 3: Mesures biomètriques sobre cranis.

Les variables (mesures en mm.) són:

X_1 = amplada, X_2 =alçada, X_3 = longitud base, X_4 = longitud nasal.

Primeres dinast. Dinasties tardanes Dinasties 12 i 13 Període tolemaic Període romà

X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
131	138	89	49	124	138	101	48	137	141	96	52	137	134	107	54	137	123	91	50
125	131	92	48	133	134	97	48	129	133	93	47	141	128	95	53	136	131	95	49
131	132	99	50	138	134	98	45	132	138	87	48	141	130	87	49	128	126	91	57
119	132	96	44	148	129	104	51	130	134	106	50	135	131	99	51	130	134	92	52
136	143	100	54	126	124	95	45	134	134	96	45	133	120	91	46	138	127	86	47
138	137	89	56	135	136	98	52	140	133	98	50	131	135	90	50	126	138	101	52
139	130	108	48	132	145	100	54	138	138	95	47	140	137	94	60	136	138	97	58
125	136	93	48	133	130	102	48	136	145	99	55	139	130	90	48	126	126	92	45
131	134	102	51	131	134	96	50	136	131	92	46	140	134	90	51	132	132	99	55
134	134	99	51	133	125	94	46	126	136	95	56	138	140	100	52	139	135	92	54
129	138	95	50	133	136	103	53	137	129	100	53	132	133	90	53	143	120	95	51
134	121	95	53	131	139	98	51	137	139	97	50	134	134	97	54	141	136	101	54
126	129	109	51	131	136	99	56	136	126	101	50	135	135	99	50	135	135	95	56
132	136	100	50	138	134	98	49	137	133	90	49	133	136	95	52	137	134	93	53
141	140	100	51	130	136	104	53	129	142	104	47	136	130	99	55	142	135	96	52
131	134	97	54	131	128	98	45	135	138	102	55	134	137	93	52	139	134	95	47
135	137	103	50	138	129	107	53	129	135	92	50	131	141	99	55	138	125	99	51
132	133	93	53	123	131	101	51	134	125	90	60	129	135	95	47	137	135	96	54
139	136	96	50	130	129	105	47	138	134	96	51	136	128	93	54	133	125	92	50
132	131	101	49	134	130	93	54	136	135	94	53	131	125	88	48	145	129	89	47
126	133	102	51	137	136	106	49	132	130	91	52	139	130	94	53	138	136	92	46
135	135	103	47	126	131	100	48	133	131	100	50	144	124	86	50	131	129	97	44
134	124	93	53	135	136	97	52	138	137	94	51	141	131	97	53	143	126	88	54
128	134	103	50	129	126	91	50	130	127	99	45	130	131	98	53	134	124	91	55
130	130	104	49	134	139	101	49	136	133	91	49	133	128	92	51	132	127	97	52
138	135	100	55	131	134	90	53	134	123	95	52	138	126	97	54	137	125	85	57
128	132	93	53	132	130	104	50	136	137	101	54	131	142	95	53	129	128	81	52
127	129	106	48	130	132	93	52	133	131	96	49	136	138	94	55	140	135	103	48
131	136	114	54	135	132	98	54	138	133	100	55	132	136	92	52	147	129	87	48
124	138	101	46	130	128	101	51	138	133	91	46	135	130	100	51	136	133	97	51

4.2. Mètode. Volem estudiar si hi ha diferències entre les 5 poblacions i representar-les separades per la distància de Mahalanobis. Farem ús de l'Anàlisi Canònica de Poblacions.

4.3. Resultats.

- Mitjanes de les variables:

Pobl.	X_1	X_2	X_3	X_4
Prim. d.	131.37	133.60	99.17	50.53
D. tard.	132.37	132.70	99.07	50.23
12&13	134.47	133.80	96.03	50.57
Tolem.	135.50	132.30	94.53	51.97
Romà	136.17	130.33	93.50	51.37
Comuna	133.97	132.55	96.46	50.93

- Matriu de covariàncies comuna:

$$S = \begin{pmatrix} 21.11 & .0368 & .0791 & 2.009 \\ & 23.485 & 5.200 & 2.845 \\ & & 24.18 & 1.133 \\ & & & 10.15 \end{pmatrix}$$

- Test de Bartlett per a l'homogeneïtat de les matrius de covariàncies. $\chi^2 = 45.66$, amb 40 g.l. (valor crític 55.76, nivell 5%). No significatiu.
- Matriu de dispersió entre grups:

$$B = \begin{pmatrix} 502.8 & -228.1 & -626.6 & 135.4 \\ & 229.9 & 292.3 & -66.07 \\ & & 803.3 & -180.7 \\ & & & 61.20 \end{pmatrix}$$

Segueix una distribució Wishart $W_4(4, \Sigma)$

- Matriu de dispersió dintre de grups:

$$W = \begin{pmatrix} 3061. & 5.333 & 11.47 & 291.3 \\ & 3405. & 754.0 & 412.5 \\ & & 3506. & 164.3 \\ & & & 1472. \end{pmatrix}$$

Segueix una distribució Wishart $W_4(145, \Sigma)$.

- Matriu de dispersió total:

$$T = \begin{pmatrix} 3564. & -222.8 & -615.1 & 426.7 \\ & 3635. & 1046. & 346.4 \\ & & 4309. & -16.40 \\ & & & 1533. \end{pmatrix}$$

- Test de comparació de mitjanes. Valor lambda de Wilks: $\Lambda = |W| / |B + W| = 0.6632$. Segueix una $\Lambda(4, 145, 4)$. Aproximació asimptòtica a una $F = 3.90$ amb 16, 434 g.ll. Hi ha diferències significatives.
- Transformació canònica, valors propis, percentatge acumulat (esquerra) i correlacions entre variables observables i canòniques (dreta):

	\mathbf{v}_1	\mathbf{v}_2	Y_1	Y_2
	.1267	.0387	X_1	.615
	-.0370	.2101	X_2	-.286
	-.1451	-.0681	X_3	-.130
	.0828	-.0773	X_4	-.058
λ	2.054	.1885		
%	88.22	96.32		

- Significació dels eixos canònics. Els valors propis l_i de B respecte de W són: .4251, .0390, .0157, .0020. La prova de significació dels valors propis dóna:

Eix	l_i	χ^2	g.ll.	Eixos	Wilks	χ^2	g.ll.
1	.4251	51.18	7	1 a 4	.664	59.24	16
2	.0390	5.53	5	2 a 4	.946	8.06	9
3	.0157	2.25	3	3 a 4	.983	2.53	4
4	.0020	0.28	1	4	.998	0.28	1

- Coordenades canòniques i radis de la representació canònica (coeficient de confiança del 90%):

Pobla.	Y_1	Y_2	radi
Prim. d.	-.795	-.033	.519
D. tard.	-.645	-.153	.519
12&13	.047	.340	.519
Tolem.	.567	.058	.519
Romà	.825	-.212	.519

4.4. Interpretació. Només la primera dimensió canònica és significativa. La podem interpretar com una dimensió temporal, que indica una certa evolució entre els diferents cranis al llarg del temps. Aquesta evolució es manifesta principalment en un augment de la amplada del crani (X_1) i una disminució de la llargada de la mandíbula (X_3).

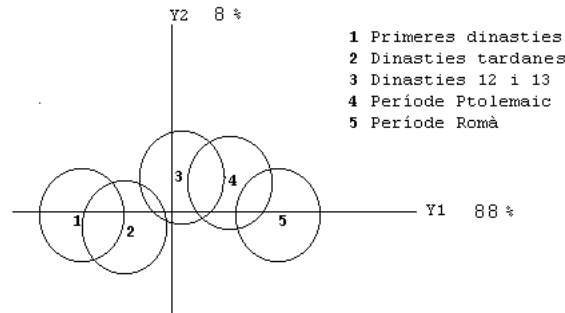


Figura 4: Representació canònica de 5 dinasties de l'Antic Egipte. El primer eix canònic es pot interpretar com una dimensió temporal.

5. COLEÒPTERS

5.1. Dades. Mesures de 5 variables biomètriques sobre 275 coleòpters del gènere *Timarcha* de 5 espècies trobades a 8 localitats:

1	<i>T. sinustocollis</i> (Campellas, Pirineos)	$n_1 = 40$
2	<i>T. sinustocollis</i> (Planollas, Pirineos)	$n_2 = 40$
3	<i>T. indet</i> (vall de Llauset, Pirineos, Osa)	$n_3 = 20$
4	<i>T. monserratis</i> (Collformic, Barcelona)	$n_4 = 40$
5	<i>T. monserratis</i> (Collfsuspina, Barcelona)	$n_5 = 40$
6	<i>T. catalaunensis</i> (La Garriga, Barcelona)	$n_6 = 40$
7	<i>T. balearica</i> (Mahón, Balears)	$n_7 = 15$
8	<i>T. pimeliodes</i> (Palermo, Sicília)	$n_8 = 40$

Les mesures (en mm.) són:

X_1 = long. pronoto, X_2 =diam. màxim pronoto, X_3 = base pronoto, X_4 = long. èlites, X_5 = diam. màxim èlites.

5.2. Mètode. Volem estudiar si hi ha diferències entre les 8 espècies i representar-les mitjançant la distància de Mahalanobis. Farem ús de l'Anàlisi Canònica de Poblacions.

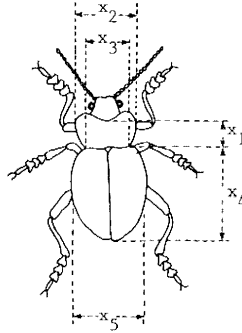


Figura 5: Mesures biomètriques sobre coleòpters.

5.3. Resultats.

- Matriu de covariàncies comuna:

$$S = \begin{pmatrix} 3.277 & 3.249 & 2.867 & 5.551 & 4.281 \\ & 7.174 & 6.282 & 9.210 & 7.380 \\ & & 6.210 & 8.282 & 6.685 \\ & & & 20.30 & 13.34 \\ & & & & 13.27 \end{pmatrix}$$

- Test de Bartlett per a l'homogeneïtat de les matrius de covariàncies. $\text{Khi-quadrat} = 229.284$, amb 105 g.l. Significatiu al 5%.
- Matriu de dispersió entre grups:

$$B = \begin{pmatrix} 6268.9 & 11386.0 & 8039.25 & 22924.9 & 17419.3 \\ & 21249.3 & 15370.2 & 42795.6 & 32502.0 \\ & & 11528.1 & 31009.3 & 23475.6 \\ & & & 86629.8 & 65626.8 \\ & & & & 49890.1 \end{pmatrix} \sim W_5(7, \Sigma)$$

- Matriu de dispersió dintre de grups:

$$W = \begin{pmatrix} 874.8 & 867.5 & 765.4 & 1482.1 & 1142.9 \\ & 1915.3 & 1677.4 & 2458.9 & 1970.3 \\ & & 1658.1 & 2211.1 & 1784.8 \\ & & & 5419.3 & 3562.6 \\ & & & & 3541.9 \end{pmatrix} \sim W_5(267, \Sigma)$$

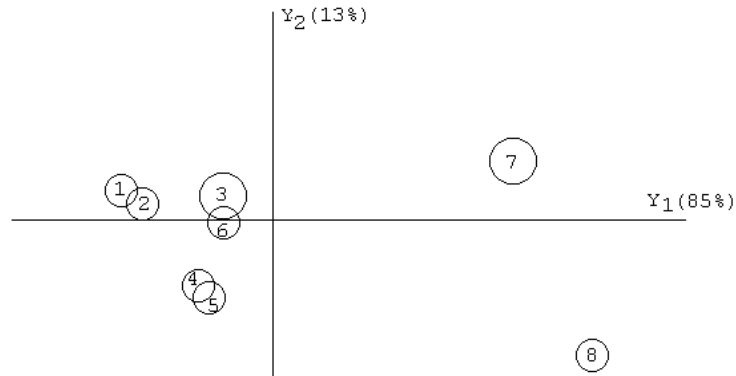


Figura 6: Representació canònica de 8 espècies de coleòpters.

- Matriu de dispersió total:

$$T = B + W$$

- Test de comparació de mitjanes. Valor lambda de Wilks:

$$\Lambda = |W| / |B + W| = 0.0102.$$

Segueix una $\Lambda(5, 267, 7)$. Aproximació asimptòtica a una $F = 62.5$ amb 35 i 1108 g.l. Hi ha diferències molt significatives.

- Transformació canònica, valors propis i percentatge acumulat de variància:

	v_1	v_2
	-.0292	.2896
	.5553	.7040
	-.6428	-.9326
	.1259	-.1326
	.1125	.0059
λ	158.64	24.53
%	85.03	98.18

5.4. Interpretació. D'acord amb la Fig. 6, les poblacions 1 i 2 pertanyen clarament a la mateixa espècie, així com la 4 i 5. Les poblacions 3 i 6 són espècies molt pròximes i les 7 i 8 es diferencien molt de les altres espècies.

6. PARTITS POLÍTICS

6.1. Dades. Matriu de distàncies entre 6 partits polítics: PP, PSOE, CU, IU, ERC, IC-V, obtinguda observant 11 característiques sociològiques i econòmiques sobre constitució de l'Estat, poder judicial i executiu, financiació autonòmica, intervenció econòmica, avortament, monarquia, etc., i aplicant la similaritat de Jaccard (les distàncies estan elevades al quadrat):

	PP	PSOE	CU	IU	ERC	IC-V
PP	0					
PSOE	1.400	0				
CU	1.714	2.000	0			
IU	1.273	0.444	1.800	0		
ERC	2.000	1.250	1.600	1.111	0	
IC-V	1.818	1.111	1.714	0.667	0.667	0

6.2. Mètode. Anàlisi de Coordenades Principals sobre la matriu de distàncies a fi d'obtenir una representació dels 6 partits polítics.

6.3. Resultats. Matriu B :

$$B = \begin{pmatrix} 0.7961 & -0.0705 & -0.0089 & -0.0829 & -0.3353 & -0.2985 \\ -0.0705 & 0.4628 & -0.3186 & 0.1650 & -0.1269 & -0.1117 \\ -0.0089 & -0.3186 & 0.9000 & -0.2944 & -0.0834 & -0.1946 \\ -0.0829 & 0.1650 & -0.2944 & 0.3111 & -0.1333 & 0.0345 \\ -0.3353 & -0.1269 & -0.0834 & -0.1333 & 0.5333 & 0.1456 \\ -0.2985 & -0.1117 & -0.1946 & 0.0345 & 0.1456 & 0.4248 \end{pmatrix}$$

- Valors propis: 1.2722, 1.1229, 0.5199, 0.3642, 0.1490, 0.0000
- Coordenades Principals:

	X_1	X_2	X_3
PP	0.5319	-0.6050	0.3826
PSOE	-0.3061	-0.3767	-0.3764
CU	0.7550	0.4932	-0.2904
IU	-0.3026	-0.2612	-0.1846
ERC	-0.2949	0.4808	0.2767
IC-V	-0.3834	0.2689	0.1921
<hr/>			
v. propi	1.2722	1.1229	0.5199
% acum	37.11	69.86	85.03

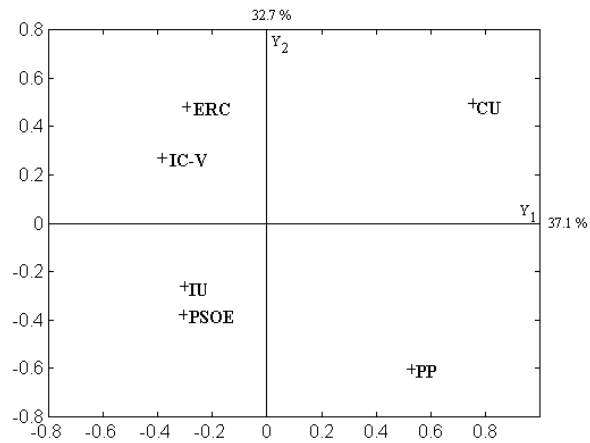


Figura 7: Representació per Anàlisi de Coordenades Principals de 6 partits polítics en relació a 11 característiques polítiques i sociològiques.

6.4. Interpretació. El primer eix principal (Fig. 7) explica el 37% de la variabilitat, i el podem entendre com una dimensió que ordena els partits segons la ideologia Esquerra-Dreta. El segon eix principal explica el 32% de la variabilitat i es podria interpretar com una dimensió Nacionalisme-Centralisme.

7. IDIOMES

7.1. Dades. Matriu de dissimilaritats entre 11 idiomes europeus:

Anglès, Norueg, Danès, Holandès, Alemany, Francès, Italià, Espanyol, Polac, Hongarès i Finlandès.

Cada dissimilaritat és el nombre de primeres lletres diferents de com s'escriuen els números 1 a 10. Per exemple, entre Anglès i Norueg hi ha dues diferències: (1=one-en, 8=eight-ate); entre Francès i Norueg hi ha sis primeres lletres diferents i entre Espanyol i Francès n'hi ha dues (4=cuatro-quatre, 8=ocho-huit).

	Angl	Noru	Dane	Hola	Alem	Fran	Esp	Ital	Polo	Hon	Fin
Angl	0										
Noru	2	0									
Dane	2	1	0								
Hola	7	5	6	0							
Alem	6	4	5	5	0						
Fran	6	6	6	9	7	0					
Espa	6	6	5	9	7	2	0				
Ital	6	6	5	9	7	1	1	0			
Polo	7	7	6	10	8	5	3	4	0		
Hong	9	8	8	8	9	10	10	10	10	0	
Finla	9	9	9	9	9	9	9	9	9	8	0

7.2. Mètode. Anàlisi cluster sobre la matriu de dissimilaritats aplicant el mètode del màxim (*complete linkage*) i del mínim (*single linkage*). La matriu inicial es transforma i es va aproximant a una matriu ultramètrica. La representació d'aquesta matriu ultramètrica és el dendrograma o arbre ultramètric.

7.3. Resultats. Les Figures 8 i 9 donen una representació en forma d'arbre ultramètric. L'índex de la jerarquia és a l'esquerra de la representació. Els resultats són semblants i les agrupacions reflecteixen els grups d'idiomes que són més similars.

7.4. Interpretació. Les llengües d'origen llatí s'agrupen, amb una certa similitat amb el Polonès. Els idiomes anglosaxons també s'agrupen. El Finlandès i l'Hongarès s'aparten de les altres llengües.

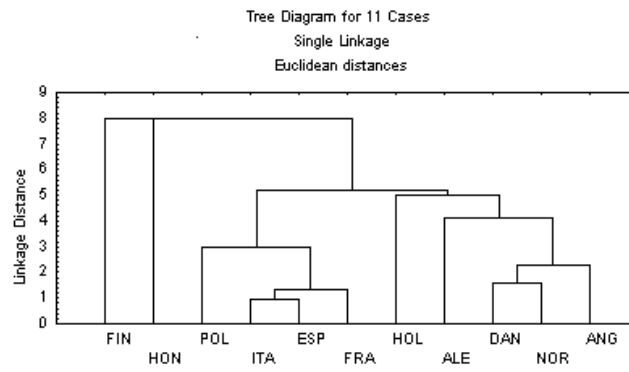


Figura 8: Dendrograma representant 11 idiomes europeus pel mètode del mínim.

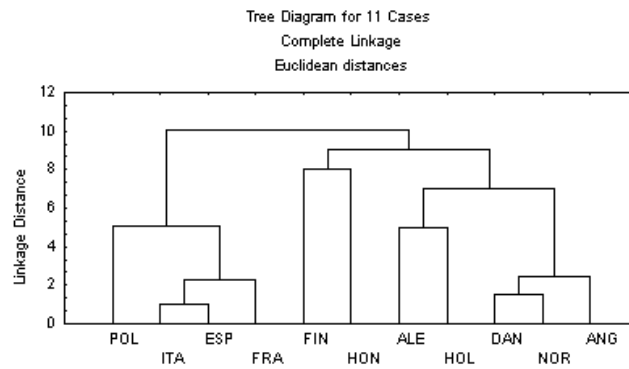


Figura 9: Dendrograma representant 11 idiomes europeus pel mètode del màxim.

8. CIUTATS

8.1. Dades. Distàncies per avió en milles entre 10 ciutats dels Estats Units.

	Atl	Chi	Den	Hou	LA	Mia	NY	SF	Sea	Was
Atlanta	0									
Chicago	587	0								
Denver	1212	920	0							
Houston	701	940	879	0						
Los Angeles	1936	1745	831	1734	0					
Miami	604	1188	1726	968	2339	0				
New York	748	713	1631	1420	2451	1092	0			
San Francisco	2139	1858	949	1645	347	2594	2571	0		
Seattle	2182	1737	1021	1891	959	2734	2408	678	0	
Washington	543	597	1494	1220	2300	923	205	2442	2329	0

8.2. Mètode. Anàlisi de coordenades principals sobre la matriu de distàncies entre ciutats.

8.3. Resultats. Valors i vectors propis de la matriu B . Les dues primeres contenen les primeres coordenades principals i són també els dos primers vectors propis de B .

	Valors propis	Ciutats	Coorden. 1	Coorden. 2
1	9582144	Atlanta	718.7	-143.0
2	1686820	Chicago	382.0	340.8
3	8157	Denver	-481.6	25.3
4	1432	Houston	161.5	-527.8
5	507	Los Angeles	-1203.7	-390.1
6	25.1	Miami	1133.5	-581.9
7	0.0	New York	1072.2	519.0
8	-897	San Francisco	-1420.6	-112.6
9	-5467	Seattle	-1341.7	579.7
10	-35478	Washington	979.6	335.5

8.4. Interpretació. La matriu B té valors propis negatius, per tant la matriu de distàncies no és euclidiana. Aixó es deu a que els avions no volen en línia recta. Atès que els valors propis positius predominen sobre els negatius, podem fer la representació prenent les dues primeres coordenades principals.

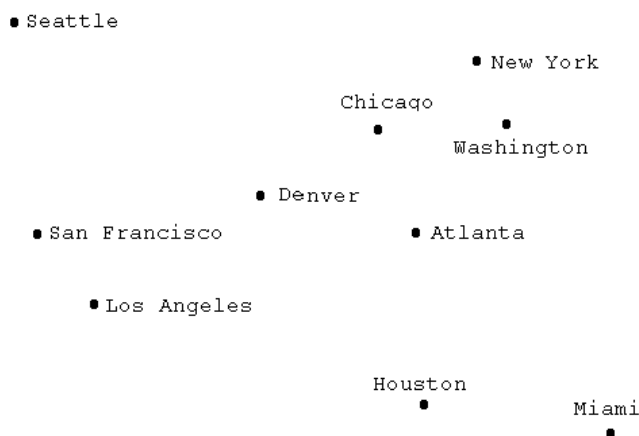


Figura 10: Representació per coordenades principals de 10 ciutats dels Estats Units a partir de les seves distàncies per avió.

9. BILOT

9.1. Definició i obtenció. Un biplot és una representació, dins d'un mateix gràfic, de les files (individus) i les columnes (variables) d'una matriu de dades $\mathbf{X}(n \times p)$. La transformació per components principals $\mathbf{Y} = \mathbf{X}\mathbf{T}$ permet representar les files. Per fer el mateix amb les columnes, podem entendre una variable X_j com el conjunt de punts de coordenades

$$\mathbf{x}_j(\alpha_j) = (0, \dots, \alpha_j, \dots, 0) \quad m_j \leq \alpha_j \leq M_j,$$

on α_j és un paràmetre que varia entre el mínim i el màxim valor de X_j . Aleshores la representació de X_j és simplement l'eix $\mathbf{x}_j(\alpha)\mathbf{T}$.

Seguint aquest procediment, és fàcil veure que la representació dels eixos és el feix de segments

$$(\alpha_1 \mathbf{v}_1, \dots, \alpha_p \mathbf{v}_p)$$

on $\mathbf{v}_1, \dots, \mathbf{v}_p$ són les files de \mathbf{T} .

9.2. Exemple. Aquest exemple completa la representació dels corredors. Amb les dades dels 12 corredors, obtenim la representació biplot de la Figura 11. Convé representar les variables com a vectors amb origen el centre de la configuració.

9.3. Interpretació. Els corredors 9, 11, 12 tenen valors alts a les variables X_1, X_2 . Els corredors 6,7 tenen valors baixos a les variables X_3, X_4 , etc. Les variables X_1, X_2 estan molt correlacionades, així com les variables X_3, X_4 .

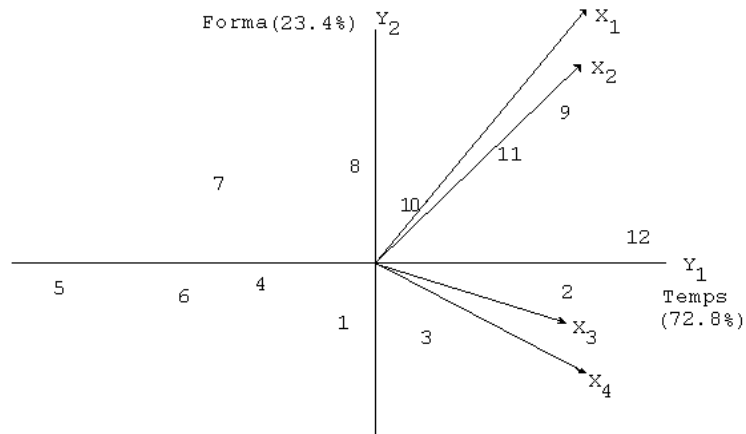


Figura 11: Representació biplot (variables i individus) de 12 corredors.

10. FRUITES

10.1. Introducció. Algunes fruites comunes (pruna, cirera, préssec, albercoc, pera, poma, nespra) pertanyen a la família de les Rosàcies, gènere *Prunus* (*P. domèstica* = pruna, *P. ávium*=cirera, *P. pérsica*=préssec, *P. ameníaca* =albercoc), gènere *Pyrus* (*P. commúnis*=pera, *P. málus*=poma) i al gènere *Méspilus* (*M. germànica* = nespra). Per exemple, la pera és l'espècie *Pyrus commúnis*, del gènere *Pyrus* de la família Rosàcies.

Els botànics, seguint criteris naturalistes, classifiquen les fruites així:

Família	Gènere	Espècie		
{	Rosàcies	{	<i>Prunus...</i>	<i>P. domèstica</i> (pruna)
			<i>P. ávium</i> (cirera)	
			<i>P. pérsica</i> (préssec)	
			<i>P. ameníaca</i> (albercoc)	
	{	Pyrus...	<i>P. commúnis</i> (pera)	
			<i>P. málus</i> (poma)	
	{	Méspilus...	<i>M. germànica</i> (nespra)	

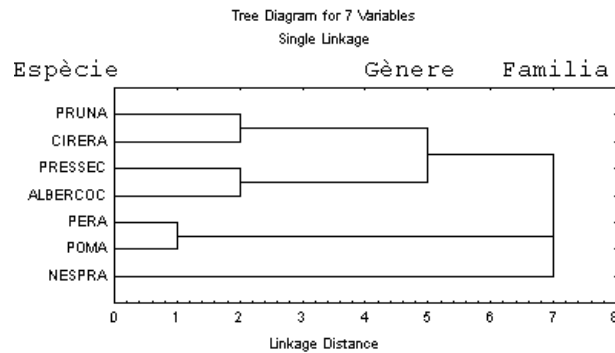


Figura 12: Representació jeràrquica de set fruites comunes.

10.2. Dades. La següent matriu de dissimilaritats s’ha obtingut preguntant a estudiants d’estadística:

	Pruna	Cirera	Prèssec	Alberc	Pera	Poma	Nespra
Pruna	0						
Cirera	2	0					
Prèssec	5	5	0				
Albercoc	6	6	2	0			
Pera	8	8	7	7	0		
Poma	8	8	7	7	1	0	
Nespra	9	9	10	9	8	7	0

10.3. Mètode. Aplicarem el mètode del mínim (*single linkage*) de classificació jeràrquica. Aquest mètode parteix d’una matriu de distàncies (o de similaritats) entre les fruites, que va agrupant segons el grau de proximitat entre fruites o grups de fruites.

10.4. Resultats. La classificació seguint el mètode del mínim (Fig. 12) proporciona uns resultats semblants a la classificació tradicional.

11. ADJECTIUS

11.1. Dades. La següent matriu proporciona les distàncies entre 23 adjectius del castellà:

Bajo, Corto, Diminuto, Menudo, Pequeño, Enorme, Inmenso, Voluminoso, Alto, Delgado, Elevado, Fino, Largo, Ancho, Angosto, Estrecho, Grande, Grueso, Profundo, Hueco, Denso, Pesado, Ligero. La distància s'obté de dues maneres:

a) Cada distància d_{ij} és la mitjana sobre 90 individus que varen puntuar la dissimilaritat entre cada parella d'adjectius i, j , des de 0 (molt semblant) fins a 4 (totalment diferent). Aixó s'indica a la meitat superior de la matriu.

b) Els 90 individus agrupaven els adjectius en grups. Cada similaritat s_{ij} és el nombre de vegades que els adjectius i, j estaven en el mateix grup i la distància és $90 - s_{ij}$. Aixó s'indica a la meitat inferior de la matriu.

	Baj	Cor	Dim	Men	Peq	Eno	Inm	Vol	Alt	Deg	Ele	Fin	Lar	Anc	Ang	Est	Gra	Gru	Pro	Hue	Den	Pes	Lig
Bajo	0	2.30	2.32	2.32	1.52	3.50	3.43	3.38	3.71	3.33	3.57	3.31	3.31	3.17	2.87	3.14	3.38	2.88	3.07	3.41	3.43	3.35	3.27
Corto	60	0	1.94	2.06	1.46	3.54	3.64	3.46	3.53	2.98	3.51	2.87	3.51	3.24	2.85	2.62	3.46	3.23	3.37	3.24	3.14	3.25	2.93
Diminuto	74	70	0	1.10	0.93	3.67	3.72	3.54	3.60	2.38	3.48	1.86	3.44	3.41	2.44	2.13	3.56	3.53	3.50	3.34	3.23	3.56	2.34
Menudo	29	76	42	0	1.01	3.73	3.56	3.58	3.37	1.83	3.42	1.71	3.24	3.40	2.80	2.26	3.50	3.34	3.47	3.36	3.30	3.24	1.85
Pequeno	70	62	16	39	0	3.74	3.72	3.56	3.61	2.71	3.37	2.23	3.44	3.26	2.20	2.08	3.72	3.34	3.41	3.36	3.20	3.40	2.25
Enorme	90	90	87	89	87	0	0.37	0.97	1.91	3.43	1.96	3.47	1.92	2.47	3.43	3.41	0.90	2.72	2.64	3.43	2.94	2.31	3.43
Inmenso	90	90	88	90	88	22	0	1.60	2.02	3.43	2.10	3.40	2.28	2.18	3.56	3.46	1.14	2.70	2.41	3.25	3.05	2.65	3.48
Voluminoso	89	89	89	87	89	66	63	0	2.72	3.61	2.45	3.60	2.94	2.35	3.48	3.52	1.30	1.82	3.02	3.42	2.55	2.27	3.47
Alto	80	84	88	89	87	85	83	87	0	3.04	0.82	3.15	2.63	3.23	3.36	3.21	1.83	3.18	2.96	3.48	3.22	2.98	3.41
Delgado	83	80	80	64	80	90	90	89	83	0	2.97	1.15	2.76	3.48	1.62	1.38	3.32	3.63	3.32	3.38	3.36	3.51	2.47
Elevado	84	87	88	89	88	84	84	86	17	85	0	3.12	2.60	3.20	3.36	3.25	2.00	3.27	3.13	3.46	3.34	3.24	3.27
Fino	84	81	74	53	75	90	90	89	83	21	86	0	2.83	3.40	1.96	2.01	3.35	3.62	3.41	3.38	3.26	3.45	2.02
Largo	84	80	89	89	88	87	85	85	74	79	75	87	0	3.24	3.04	3.08	2.46	3.37	2.80	3.42	3.28	3.32	3.41
Ancho	85	83	89	89	88	86	84	76	82	83	84	87	73	0	3.48	3.53	1.03	2.76	2.82	3.27	2.97	3.18	3.32
Angosto	82	74	77	78	79	90	89	88	85	53	86	58	82	84	0	0.68	3.33	3.55	3.37	3.34	3.21	3.38	2.91
Estrecho	81	74	82	81	84	89	90	89	85	54	85	63	81	83	23	0	1.95	1.94	3.26	3.44	2.80	2.35	3.31
Grande	87	88	84	86	82	37	49	62	77	87	78	88	83	80	89	89	0	2.85	2.81	3.46	3.11	3.10	3.40
Grueso	87	86	89	86	87	81	86	64	85	82	86	86	84	63	87	86	72	0	3.23	3.36	2.44	2.35	3.47
Profundo	82	86	89	88	89	86	86	83	87	88	86	89	87	85	85	86	87	85	0	2.57	2.77	3.23	3.43
Hueco	82	83	88	89	88	90	90	88	87	85	84	87	85	86	84	84	88	87	66	0	3.33	3.41	2.84
Denso	89	89	89	87	89	87	86	77	88	87	89	88	87	82	89	88	85	72	79	87	0	3.35	3.48
Pesado	90	90	90	89	88	88	88	75	87	89	89	88	84	90	90	85	58	89	90	56	0	3.51	
Ligero	86	87	83	69	83	90	90	90	89	72	89	71	90	90	83	80	90	89	90	87	84	81	0

11.2. Mètode. Multidimensional scaling no mètric sobre la matriu de distàncies (meitat superior) per esbrinar si hi ha dimensions semàntiques que ordenin els adjectius. Els passos del mètode són:

- a) La distància original δ_{ij} s'ajusta a una disparitat \hat{d}_{ij} per regressió monòtona.
- b) Fixada una dimensió, s'aproxima \hat{d}_{ij} a una distància Euclidiana d_{ij}
- c) Es calcula la mesura de "stress"

$$[\sum (d_{ij} - \hat{d}_{ij})^2 / \sum d_{ij}^2]^{1/2}$$

- d) Es representen les $n(n - 1)/2$ distàncies d_{ij} vs les \hat{d}_{ij} , per visualitzar les relacions de monotomia.

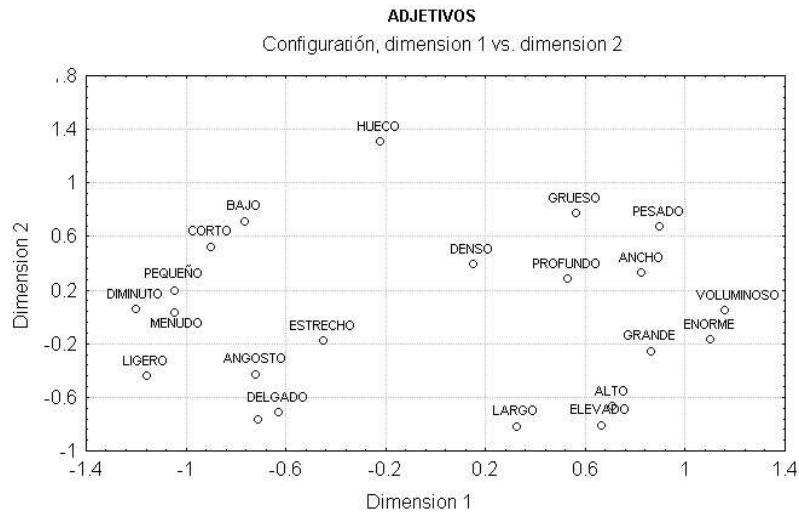


Figura 13: Representació MDS de 23 adjectius del castellà.

La configuració en dues dimensions (Fig. 13) és una millor aproximació en dimensió 2 a les distàncies originals (transformades monotònicament) en el sentit de que minimitza el “stress”. Té un “stress” del 19%.

11.3. Interpretacio. S’aprecien diversos gradients de valoració dels adjectius:

1. Diminuto \longleftrightarrow Enorme
2. Bajo-Corto \longleftrightarrow Alto-Largo
3. Delgado \longleftrightarrow Grueso
4. Ligero \longleftrightarrow Pesado.
5. Hueco (constitueix un adjectiu diferenciat).

La representació en el estudi original considera 6 dimensions, que representa separatament, amb un stress del 5%, però la interpretació no és gaire diferent. Per a aquesta representació s’obté el gràfic de la Fig. 14, que relaciona distàncies de la representació amb disparitats.

11.4. Anàlisi cluster. Seguidament s’aplica una anàlisi cluster sobre la matriu de distàncies (meitat inferior) pel mètode del mínim (single linkage) i del màxim (complete linkage). Els resultats són bastant semblants (Figures 15 i 16), indicant que hi ha una bona estructura jeràrquica. Hi ha una divisió principal, que agrupa els adjectius de pes i extensió espacial, seguint la dicotomia “gran quantitat” vs “petita quantitat”.

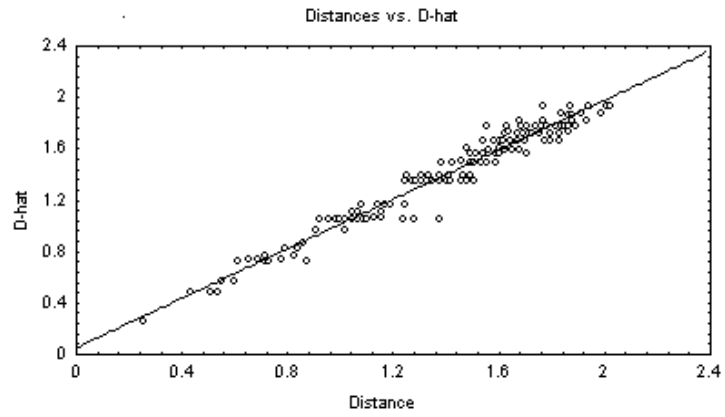


Figura 14: Distàncies originals (transformades) entre adjectius vs distàncies de la representació prenent 6 dimensions. El stress és 5%.

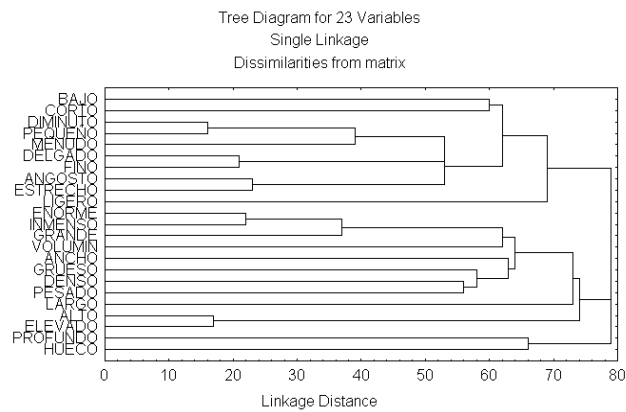


Figura 15: Representació de 23 adjectius pel mètode del mínim.

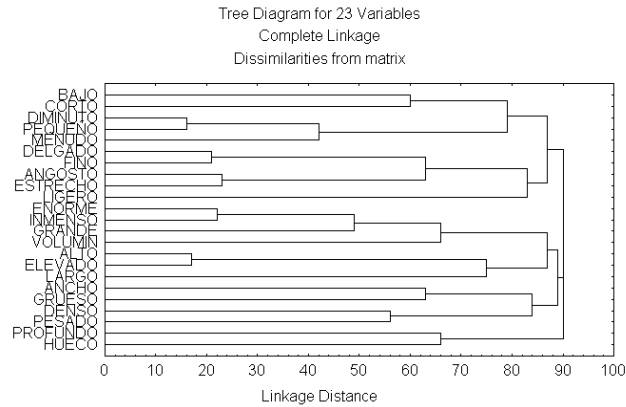


Figura 16: Representació de 23 adjectius pel mètode del màxim.

12. ASSIGNATURES

12.1. Dades. Estudi de 21 assignatures troncal de la carrera de Matemàtiques. S'han considerat les següents variables:

- $X_1 =$ nivell abstracció
- $X_2 =$ aplicabilitat
- $X_3 =$ grau numèric
- $X_4 =$ grau dificultat

Les dades són les valoracions donades per 23 alumnes d'últim any de carrera, segons una escala de 1 a 10. El valor considerat ha estat la moda de les respostes dels alumnes.

Assignatura	X_1	X_2	X_3	X_4
1.Àlgebra Lineal	4	8	4	3
2.Anàlisi I	3	9	4	4
3.Informàtica	1	6	9	1
4.Geometria Lineal	3	6	3	3
5.Anàlisi II	2	8	7	3
6.Mètodes Numèrics	3	6	10	4
7.Geometria Projectiva	6	6	1	6
8.Anàlisi III	3	8	4	4
9.Probabilitats	3	7	7	5
10.Topologia	7	6	1	7
11.Geometria de C. i S.	5	6	5	5
12.Anàlisi IV	2	7	6	6
13.Estadística	1	6	10	1
14.Algebra I	10	7	7	7
15.Funcions Anal.	5	4	5	6
16.ÀlgebraII	10	4	2	10
17.Càlcul Numeric	5	3	10	7
18.Topologia Algeb.	6	3	1	5
19.Analisi Funcional	8	4	2	9
20.Equacions Diferenc.	7	3	8	8
21.Geometria Diferenc.	10	3	5	10

12.2. Resultats. Els resultats de l'anàlisi de components principals són:

1. Vector de mitjanes i matriu de covariàncies:

$$\bar{\mathbf{x}} = (4.95, 5.71, 5.28, 5.43)$$

$$S = \begin{pmatrix} 38.5 & 11.7 & -9.65 & -0.50 & 11.9 \\ 11.7 & 8.247 & -2.91 & -3.68 & 6.52 \\ -9.65 & -2.91 & 3.51 & .185 & -2.87 \\ -0.50 & -3.68 & .1857 & 9.21 & -2.73 \\ 11.9 & 6.521 & -2.871 & -2.73 & 6.66 \end{pmatrix}$$

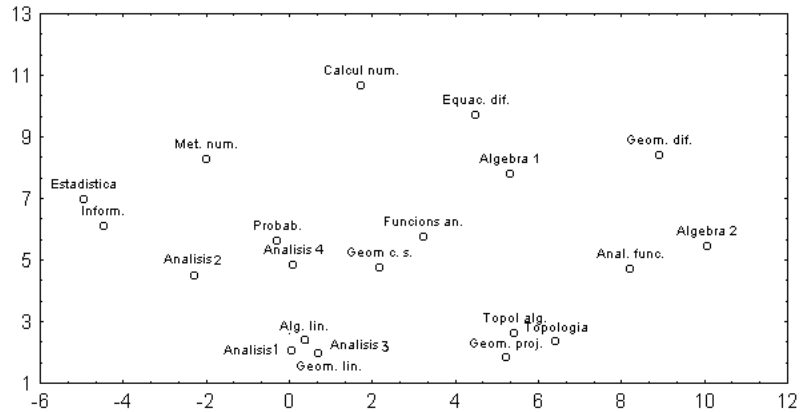


Figura 17: Representació per components principals de 21 assignatures troncs de la llicenciatura de matemàtiques.

2. Dos primers vectors i valors propis de S :

	t_1	t_2
	.9410	-.2988
	.3427	.4238
	-.2403	.0813
	-.0668	-.8974
	.3340	.2859
V. propi	49.48	11.64
%	74.96	17.50
Acum	74.96	92.47

3. Matriu de correlacions:

$$R = \begin{pmatrix} 1 & -.5413 & -.4227 & .8801 \\ -.5413 & 1 & .0326 & -.5936 \\ -.4227 & .0326 & 1 & -.3483 \\ .8801 & -.5936 & -.3483 & 1 \end{pmatrix}$$

12.3. Interpretació. D'acord amb la Fig. 17, la primera dimensió principal ordena les assignatures per grau de dificultat. La segona dimensió no té una interpretació clara, però una rotació d'eixos revelaria una dimensió que separaria assignatures amb contingut numèric d'assignatures més abstractes.

13. FLORS

13.1. Dades. Mesures de 4 variables biomètriques quantitatives sobre 3 espècies de flors del gènere Iris (dades de R.A. Fisher):

1. Iris setosa, 2. Iris versicolor, 3. Iris virginica.

Les variables i les dades són:

X_1 = llargada del sèpal, X_2 = amplada del sèpal,
 X_3 = llargada del pètal, X_4 = amplada del pètal.

X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3.0	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3	6.3	2.9	5.6	1.8
5.0	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3.0	5.8	2.2
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3.0	6.6	2.1
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
5.0	3.4	1.5	0.2	4.9	2.4	3.3	1.0	7.3	2.9	6.3	1.8
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
5.4	3.7	1.5	0.2	5.0	2.0	3.5	1.0	6.5	3.2	5.1	2.0
4.8	3.4	1.6	0.2	5.9	3.0	4.2	1.5	6.4	2.7	5.3	1.9
4.8	3.0	1.4	0.1	6.0	2.2	4.0	1.0	6.8	3.0	5.5	2.1
4.3	3.0	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5.0	2.0
5.8	4.0	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.4	3.9	1.3	0.4	5.6	3.0	4.5	1.5	6.5	3.0	5.5	1.8
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1.0	7.7	3.8	6.7	2.2
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6.0	2.2	5.0	1.5
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
5.1	3.7	1.5	0.4	6.1	2.8	4.0	1.3	5.6	2.8	4.9	2.0
4.6	3.6	1.0	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2.0
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
5.0	3.0	1.6	0.2	6.6	3.0	4.4	1.4	7.2	3.2	6.0	1.8
5.0	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
5.2	3.5	1.5	0.2	6.7	3.0	5.0	1.7	6.1	3.0	4.9	1.8
5.2	3.4	1.4	0.2	6.0	2.9	4.5	1.5	6.4	2.8	5.6	2.1
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1.0	7.2	3.0	5.8	1.6
4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	2.8	6.1	1.9
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1.0	7.9	3.8	6.4	2.0
5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
5.5	4.2	1.4	0.2	6.0	2.7	5.1	1.6	6.3	2.8	5.1	1.5
4.9	3.1	1.5	0.2	5.4	3.0	4.5	1.5	6.1	2.6	5.6	1.4
5.0	3.2	1.2	0.2	6.0	3.4	4.5	1.6	7.7	3.0	6.1	2.3
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3.0	1.3	0.2	5.6	3.0	4.1	1.3	6.0	3.0	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4.0	1.3	6.9	3.1	5.4	2.1
5.0	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3.0	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4.0	1.2	5.8	2.7	5.1	1.9
5.0	3.5	1.6	0.6	5.0	2.3	3.3	1.0	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3.0	1.4	0.3	5.7	3.0	4.2	1.2	6.7	3.0	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5.0	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3.0	5.2	2.0
5.3	3.7	1.5	0.2	5.1	2.5	3.0	1.1	6.2	3.4	5.4	2.3
5.0	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3.0	5.1	1.8

13.2. Mètode. Donades les tres noves flors (individus)

Individu	X_1	X_2	X_3	X_4
x_1	4.6	3.6	1.0	0.2
x_2	6.8	2.8	4.8	1.4
x_3	7.2	3.2	6.0	1.8

volem esbrinar a quines de les poblacions anteriors pertanyen. Per això necessitem una regla de decisió que ens permeti assignar aquests individus a alguna de les 3 poblacions. Farem ús de l'Anàlisi Discriminant.

13.3. Resultats.

- Mitjanes de les variables:

Població	X_1	X_2	X_3	X_4
I. setosa	5.0060	3.4280	1.4620	0.2460
I. versicolor	5.9360	2.7700	4.2600	1.3260
I. virginica	6.5880	2.9740	5.5520	2.0260
Comuna	5.8433	3.0573	3.7580	1.1993

- Discriminador lineal

L'estimació de la matriu de covariàncies comuna és

$$S = \begin{pmatrix} .2650 & .0927 & .1675 & .0384 \\ & .1154 & .05524 & .0327 \\ & & .18519 & .0426 \\ & & & .0418 \end{pmatrix}$$

Les distàncies de Mahalanobis (al quadrat) entre les tres poblacions són:

	I.setosa	I.versicolor	I.virginica
I.setosa	0	89.864	179.38
I.versicolor		0	17.201
I.virginica			0

Les funcions discriminants lineals són:

$$\begin{aligned} L_{12}(x) &= \frac{1}{2} [M^2(x, \bar{x}_2) - M^2(x, \bar{x}_1)], \\ L_{13}(x) &= \frac{1}{2} [M^2(x, \bar{x}_3) - M^2(x, \bar{x}_1)], \\ L_{23}(x) &= L_{13}(x) - L_{12}(x), L_{21}(x) = -L_{12}(x), \\ L_{31}(x) &= -L_{13}(x), L_{32}(x) = -L_{23}(x). \end{aligned}$$

on \bar{x}_i són les mitjanes de les variables en la població i , és a dir, l'individu mig de la població i , $M^2(x, \bar{x}_i)$ és la distància de Mahalanobis al quadrat entre el nou individu x i l'individu mig \bar{x}_i , per $i = 1, 2, 3$. quadrat entre el nou individu x i l'individu mig \bar{x}_i , per $i = 1, 2, 3$. La regla de decissió consisteix en assignar l'individu x a la població i si

$$L_{ij}(x) > 0 \quad \forall j \neq i.$$

Per estimar la probabilitat de classificació errònia pce és útil llevar cada individu, classificar-lo partint dels altres i observar si surt ben classificat (mètode *leaving-on-out*). El resultat d'aquest procediment dóna:

		Població assignada		
		1	2	3
Població original	1	50	0	0
	2	0	48	2
	3	0	1	49

Hi ha només 3 individus mal classificats i la pce estimada seria $3/150 = 0.02$. La classificació per als nous individus és:

Ind.	L_{12}	L_{13}	L_{21}	L_{23}	L_{31}	L_{32}	Pobl.
x_1	55.681	103.87	-55.681	48.186	-103.87	-48.186	1
x_2	-51.107	-44.759	51.107	6.3484	44.759	-6.3484	2
x_3	-76.866	-82.785	76.866	-5.9195	82.785	5.9195	3

- Discriminador quadràtic

El test d'homogeneïtat de covariàncies (test de Bartlett) dóna una $\chi^2 = 140.94$ amb 20 graus de llibertat. El valor crític de taules per a un nivell de significació del 5% és 31.410. Per tant, les diferències entre les matrius de covariàncies són significatives. Així doncs, el discriminador quadràtic podria resultar més adient.

Els discriminadors quadràtics són:

$$Q_{12}(x) = \frac{1}{2}x'(S_2^{-1} - S_1^{-1})x + x'(S_1^{-1}\bar{x}_1 - S_2^{-1}\bar{x}_2) + \frac{1}{2}\bar{x}_2'S_2^{-1}\bar{x}_2 - \frac{1}{2}\bar{x}_1'S_1^{-1}\bar{x}_1 + \frac{1}{2}\log|S_2| - \frac{1}{2}\log|S_1|$$

$$Q_{13}(x) = \frac{1}{2}x'(S_3^{-1} - S_1^{-1})x + x'(S_1^{-1}\bar{x}_1 - S_3^{-1}\bar{x}_3) + \frac{1}{2}\bar{x}_3'S_3^{-1}\bar{x}_3 - \frac{1}{2}\bar{x}_1'S_1^{-1}\bar{x}_1 + \frac{1}{2}\log|S_3| - \frac{1}{2}\log|S_1|$$

$$Q_{23}(x) = Q_{13}(x) - Q_{12}(x)$$

$$Q_{21}(x) = -Q_{12}(x), \quad Q_{31}(x) = -Q_{13}(x), \quad Q_{32}(x) = -Q_{23}(x).$$

La regla de decisió consisteix en assignar l'individu x a la població i si

$$Q_{ij}(x) > 0 \quad \forall j \neq i, \quad i, j = 1, 2, 3.$$

La taula de classificacions és:

		Població assignada		
		1	2	3
Població original	1	50	0	0
	2	0	47	3
	3	0	1	49

Ara s'han classificat malament 4 individus de les dades originals. La classificació per als nous individus és:

Ind.	Q_{12}	Q_{13}	Q_{21}	Q_{23}	Q_{31}	Q_{32}	Pobl.
x_1	61.177	94.767	-61.177	33.589	-94.767	-33.589	1
x_2	-224.46	-217.98	224.46	6.4782	217.98	-6.4782	2
x_3	-382.98	-387.88	382.98	-4.9060	387.88	4.9060	3

- Discriminador basat en distàncies:

Utilitzant com a matriu de la mètrica la matriu de distàncies de Mahalanobis, les funcions de proximitat són

$$\phi_i^2(x) = (x - \bar{x}_i)' S^{-1} (x - \bar{x}_i), \quad i = 1, 2, 3.$$

La regla de decisió consisteix en assignar l'individu x a la població més pròxima, és a dir, si $\phi_i^2(x) = \min_{1 \leq j \leq 3} \{\phi_j^2(x)\}$, s'assigna x a la població i . Observeu que hauríem d'obtenir els mateixos resultats que utilitzant el discriminador lineal, ja que

$$L_{ij}(x) = \frac{1}{2}(\phi_j^2(x) - \phi_i^2(x)), \quad i, j = 1, 2, 3.$$

La classificació per als nous individus és:

Ind.	ϕ_1^2	ϕ_2^2	ϕ_3^2	Pobl.
x_1	2.2864	113.65	210.02	1
x_2	105.94	3.7243	16.422	2
x_3	171.10	17.364	5.5252	3

14. NOMS CATALANS

14.1. Dades. Es consideren les 41 comarques catalanes i per a cadascuna es tabulen els valors de les següents variables:

$$\begin{aligned} X_1 &= \log(\text{percentatge de vots a CU}), & X_2 &= \log(\text{percentatge de vots a PSC}), \\ X_3 &= \log(\text{percentatge de vots a PP}), & X_4 &= \log(\text{percentatge de vots a ERC}), \\ Y_1 &= \log(\text{quocient Juan/Joan}), & Y_2 &= \log(\text{quocient Juana/Joanna}), \end{aligned}$$

on el “quocient Juan/Joan” significa el resultat de dividir el nombre d’homes que es diuen Juan pel nombre d’homes que es diuen Joan. Valors positius de les variables Y_1, Y_2 a una comarca indiquen predomini del nom castellà sobre el nom català.

Comarq.	CU	PSC	PP	ERC	Juan	Joan	Juana	Joanna
1.A. Camp.	44.6	29.6	6.2	16.1	684	605	143	38
2.A. Empo.	47.3	30.7	7.9	10.8	1628	1264	358	101
3.A. Pene.	47.4	31.8	5.6	10.7	1502	1370	281	90
4.A. Urgell	49.5	24.7	6.4	17.3	370	346	56	39
5.A. Ribag.	42.1	41.1	5.9	8.9	29	30	9	4
6.Anoia	44.8	33.9	6.6	8.7	1759	975	433	115
7.Bages	47.9	30	4.9	12.2	2766	1970	559	145
8.B. Camp	40.8	33.3	10	12	2025	1081	600	138
9.B. Ebre	44.2	31.3	12.1	9.5	1634	484	329	138
10.B. Emp.	48.2	32.4	5.1	11	1562	1423	334	153
11.B. Llob.	48.1	27.6	9.4	5.6	10 398	2687	3103	325
12. B. Pene.	39.7	40.5	9.1	7.9	957	577	236	33
13.Barç.	32	41.2	12.2	7.1	27 841	10 198	9287	1598
14.Bergu.	51.2	25.8	4.4	14.7	830	590	108	33
15.Cerda.	51.1	25.9	5.5	13.9	190	228	50	12
16.Conca B.	49.9	20.9	5.9	17.9	247	492	49	45
17.Garraf	37.9	39	8.5	7.8	1474	477	618	154
18.Garrig.	50	24.1	6.4	17.5	191	269	21	33
19.Garrot.	56.1	23.4	4.3	13.3	950	1168	100	91
20.Giron.	42.8	31.7	6.6	14.7	1978	1861	430	191
21.Mares.	43	32.9	8.9	9.2	5234	3053	1507	280
22.Monts.	49.4	31.5	8.1	8	907	314	229	82
23.Nogue.	53.7	24.3	7	12.2	557	487	92	37
24.Osona	56.7	18.5	3.9	16	1794	2548	222	100
25.P. Jussà	50	30.5	4.9	12.4	154	115	27	14
26.P. Sobirà	51.1	30.8	4.8	10.9	61	121	9	15
27.P. Urg.	52.4	25.8	6.6	12.6	393	299	58	20
28.Pla Est.	57.1	15.7	4.5	20	159	869	32	52
29.Prior.	45.9	27.7	6.2	16.9	173	149	37	16
30.R. Ebre	48.9	31.3	6.8	10.4	407	185	98	29
31.Ripoll.	55.4	25.8	3.3	12.8	603	457	75	17
32.Segar.	53.67	21.16	6.87	15.58	222	320	27	15
33.Segrià	42.77	35.33	9.66	8.91	2049	951	625	202
34.Selva	49.2	29	6.2	11.4	1750	1680	340	152
35.Solso.	57.8	17.5	5.8	15.9	95	401	20	12
36.Tarra.	34.53	38.76	13.89	8.81	2546	940	852	117
37.Ter. A.	49	25.1	14.2	9.3	164	125	55	20
38.Urgell	54.18	22.5	6.9	13.86	144	656	45	56
39.Val. A.	44.49	38.3	12.59	2.67	97	19	37	2
40.Vall. Oc.	33.68	42.62	8.42	7.1	11 801	4482	3110	416
41.Vall Or.	40.72	37.96	7.51	7.63	4956	2636	1227	233

14.2. Mètode. Anàlisi de correlació canònica per relacionar els vots a 4 partits (eleccions autonòmiques de 1999) amb el grau de catalanisme dels noms propis Joan i Joanna.

14.3. Resultats.

- Matriu de correlacions

	X_1	X_2	X_3	X_4	Y_1	Y_2
X_1	1	-.8520	-.6536	-.5478	-.6404	-.5907
X_2		1	.5127	-.7101	.7555	.6393
X_3			1	-.6265	.5912	.5146
X_4				1	-.7528	-.7448
Y_1					1	.8027
Y_2						1

- Els valors propis ó arrels de l'equació:

$$|\mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21} - \lambda\mathbf{R}_{11}| = 0,$$

són $\lambda_1 = 0.7017$, $\lambda_2 = 0.1701$.

- Només hi ha dues correlacions canòniques:

$$r_1 = 0.8377, \quad r_2 = 0.4125.$$

- Variables canòniques:

$$\begin{aligned} U_1 &= +0.083X_1 - 0.372X_2 - 0.1130X_3 + 0.555X_4, & (r_1 = 0.8377), \\ V_1 &= +0.706Y_1 + 0.339Y_2, \\ U_2 &= +1.928X_1 + 2.4031.546X_2 + 1.127X_3 + 1.546X_4, & (r_2 = 0.4125). \\ V_2 &= +1.521Y_1 - 1.642Y_2, \end{aligned}$$

14.4. Interpretació. Les primeres variables canòniques U_1, V_1 , que podem escriure de forma convencional com

$$\begin{aligned} U_1 &= +0.083CU - 0.372PSC - 0.1130PP + 0.555ERC, \\ V_1 &= +0.706(\text{Juan/Joan}) + 0.339(\text{Juana/Joanna}), \end{aligned}$$

ens indiquen que les regions més catalanes, en el sentit que el noms castellans Juan i Juana no predominen tant sobre els catalans Joan i Joanna, tendeixen a votar més a CU i ERC. Les regions que voten més al PSC i al PP són, en general, més castellanitzades. Les segones variables canòniques indicarien que els vots a tots els partits tenen a veure amb el contrast entre el nom d'home i el de dona.