Tarea 2

Christian Badillo Luis Nuñez Luz Maria Santana Sealtiel Pichardo

Tabla de contenidos

1 Parte 4

Vemos los datos.

Tabla 1: Primeras diez observaciones.

Al2O3	Fe2O3	MgO	CaO	Na2O	K2O	TiO2	MnO	BaO	kiln
18.8	9.52	2.00	0.79	0.40	3.20	1.01	0.077	0.015	1
16.9	7.33	1.65	0.84	0.40	3.05	0.99	0.067	0.018	1
18.2	7.64	1.82	0.77	0.40	3.07	0.98	0.087	0.014	1
16.9	7.29	1.56	0.76	0.40	3.05	1.00	0.063	0.019	1
17.8	7.24	1.83	0.92	0.43	3.12	0.93	0.061	0.019	1
18.8	7.45	2.06	0.87	0.25	3.26	0.98	0.072	0.017	1
16.5	7.05	1.81	1.73	0.33	3.20	0.95	0.066	0.019	1
18.0	7.42	2.06	1.00	0.28	3.37	0.96	0.072	0.017	1
15.8	7.15	1.62	0.71	0.38	3.25	0.93	0.062	0.017	1
14.6	6.87	1.67	0.76	0.33	3.06	0.91	0.055	0.012	1

Dado que las escalas de los datos son distintas, se normalizaron los datos con la función `scale` de

R base, después se calculo la distancia euclidiana para las 45 observaciones.

```
data.centered <- scale(data)
dist.matrix <- as.matrix(dist(data.centered,
                              method = "euclidean"), 45, 45)
```

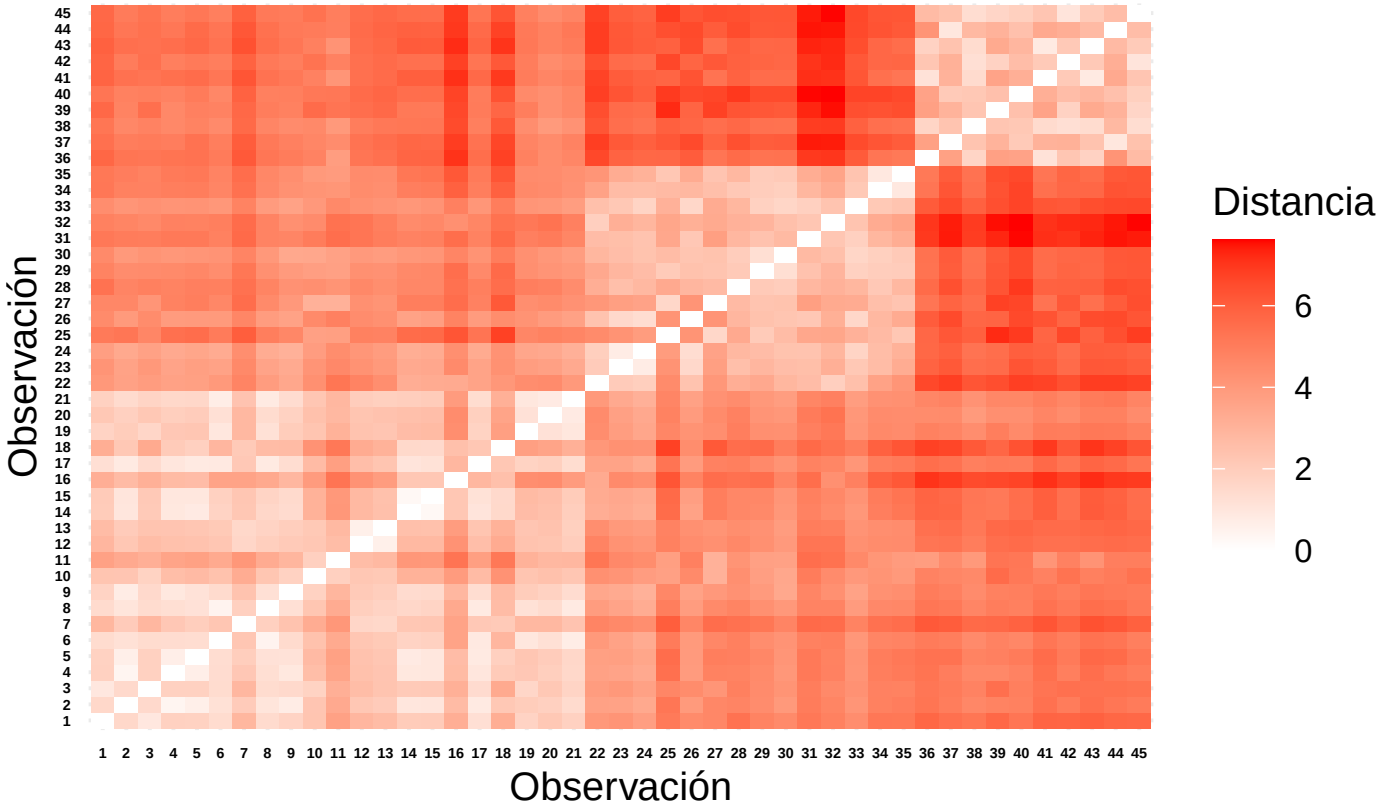


Figura 1: Mapa de Calor de la Matrix de Distancias.

En el mapa de calor se puede visibilizar una estructura de 3 grupos marcada, por lo cual se espera que cualquier análisis que contemple la existencia de 3 grupos debería de ajustarse bien.

1.1 Análisis de Conglomerados.

Se filtran los datos para solo tomar en cuenta la composición química de las vasijas y después se procede a realizar el análisis jerárquico de conglomerados usando liga sencilla, liga completa y el método de Ward.

```
data.chem <- data.centered %>%
    as.data.frame() %>%
    select(!c(kiln))

dist.chem <- data.chem %>%
```

```
    dist() %>%
    as.matrix(ncols =45, nrows = 45)

link.complete <- agnes(x = dist.chem, diss = T, method = "complete")
link.single <- agnes(x = dist.chem, diss = T, method = "single")
cluster.ward <- agnes(x = dist.chem, diss = T, method = "ward")
```

Usando la hipótesis de que existen 3 grupos se predice que el mejor corte se vera reflejado con $k = 3$ para los distintos métodos. Se visualizarán los grupos formados usando las primeras dos componentes principales con la ayuda del paquete `factoextra` de `R`.

1.1.1 Liga Sencilla.

1.1.1.1 Dos Grupos.

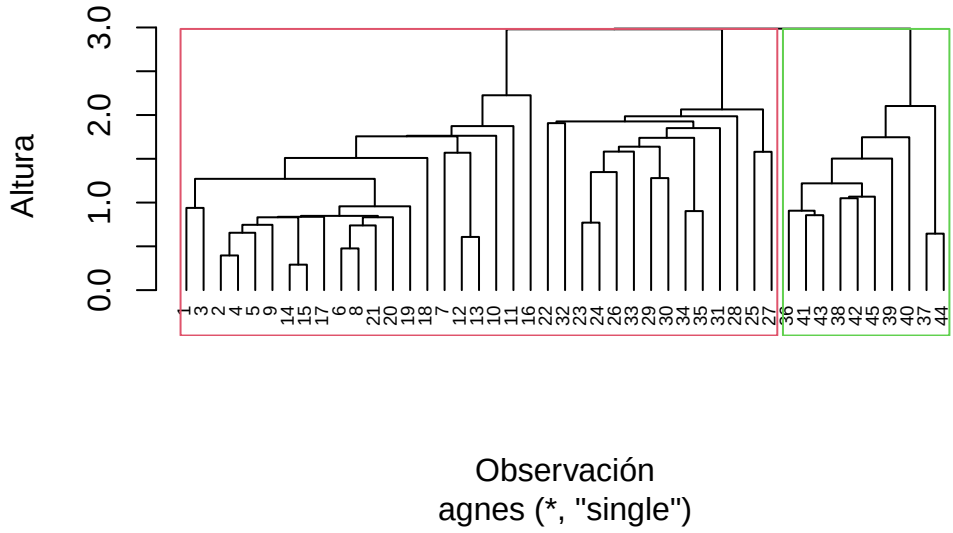


Figura 2: Dendograma: 2 grupos (liga sencilla).

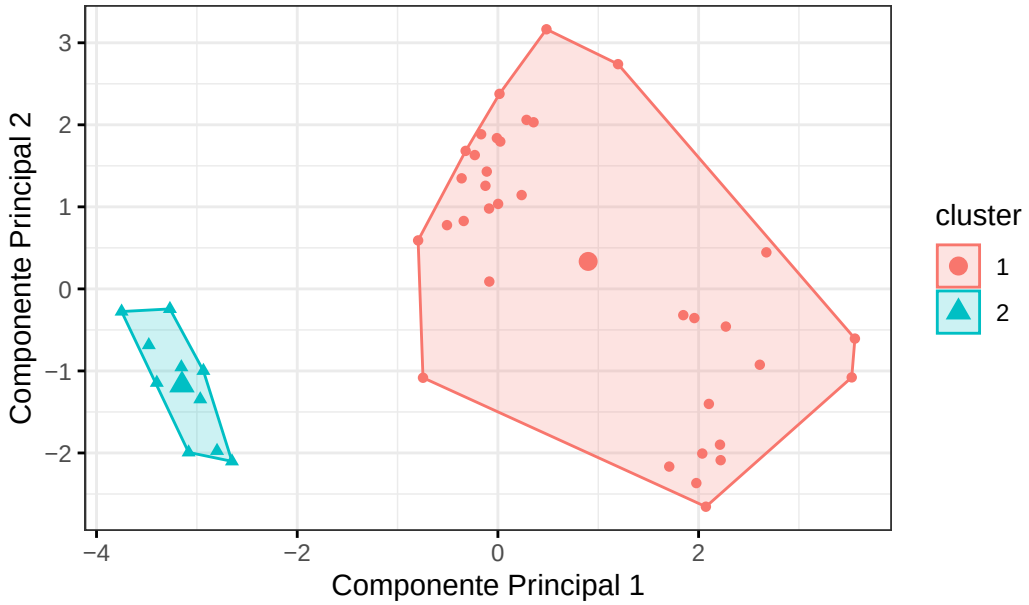


Figura 3: Conglomerados: 2 grupos (liga sencilla).

1.1.1.2 Tres Grupos.

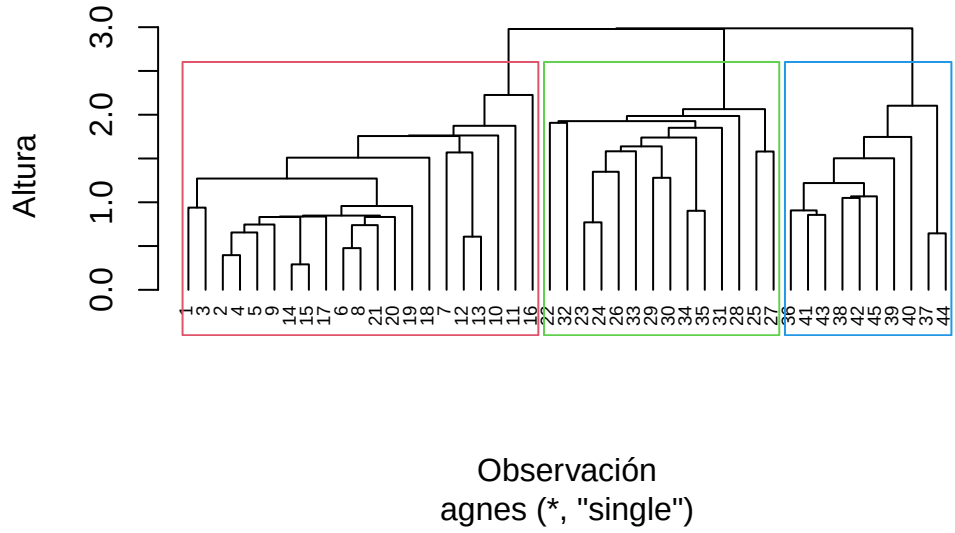


Figura 4: Dendograma: 3 grupos (liga sencilla).

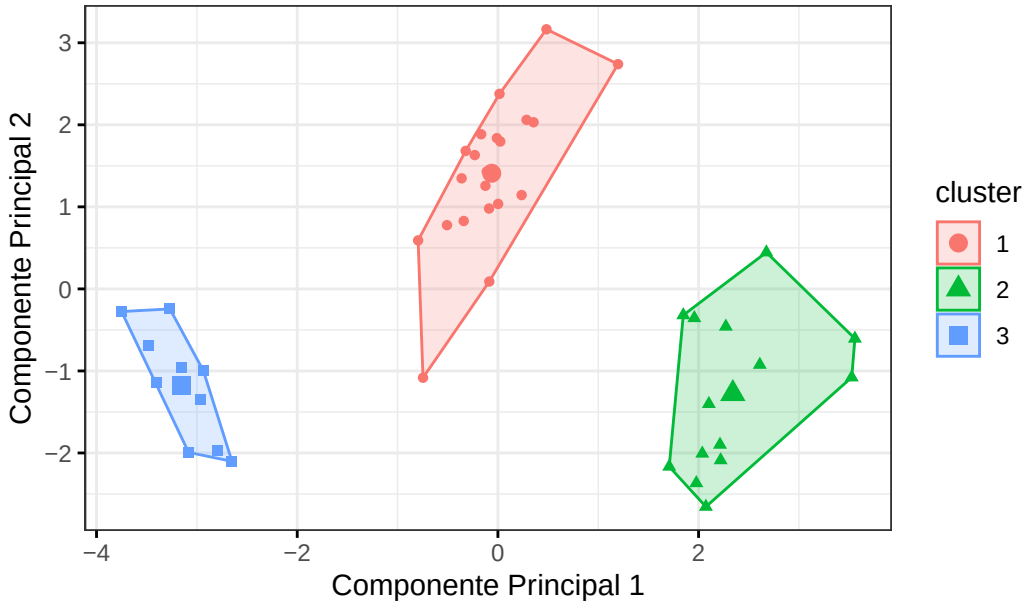


Figura 5: Conglomerados: 3 grupos (liga sencilla).

1.1.1.3 Cuatro Grupos.

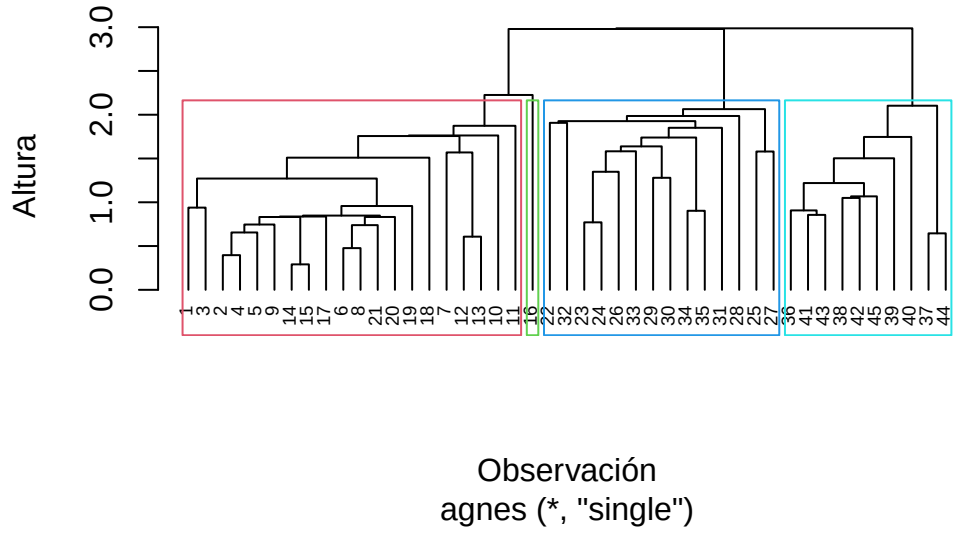


Figura 6: Dendograma: 4 grupos (liga sencilla).

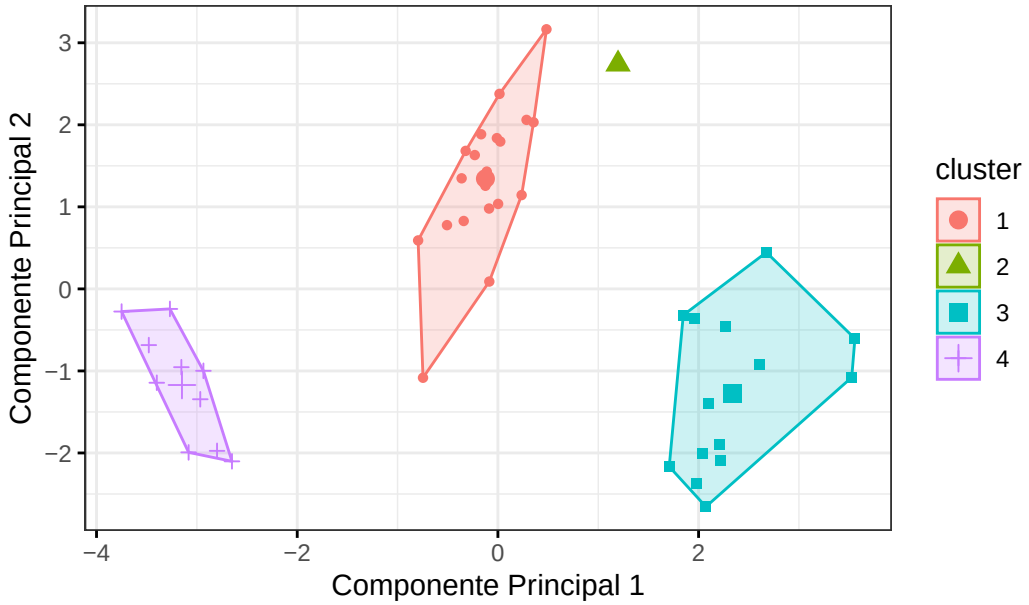


Figura 7: Conglomerados: 4 grupos (liga sencilla).

1.1.1.4 Cinco Grupos.

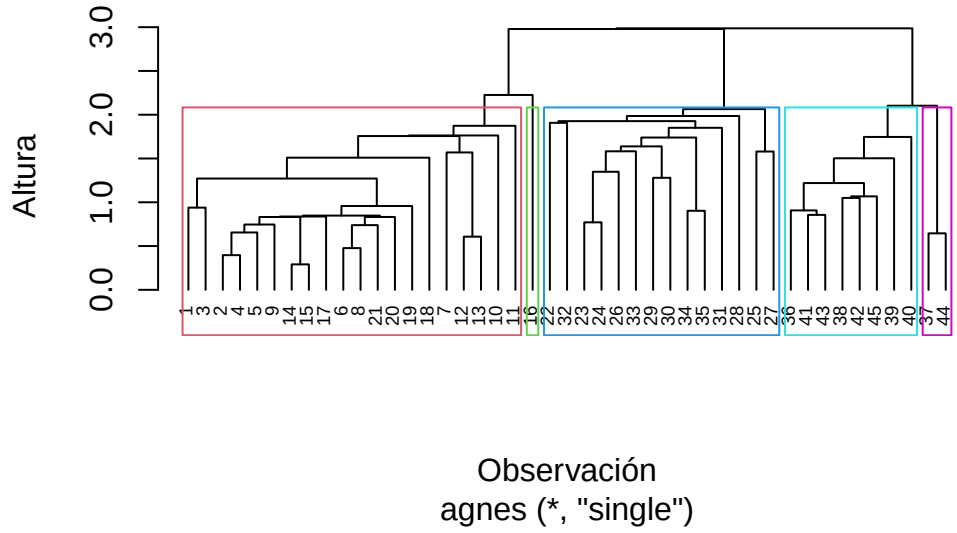


Figura 8: Dendograma: 5 grupos (liga sencilla).

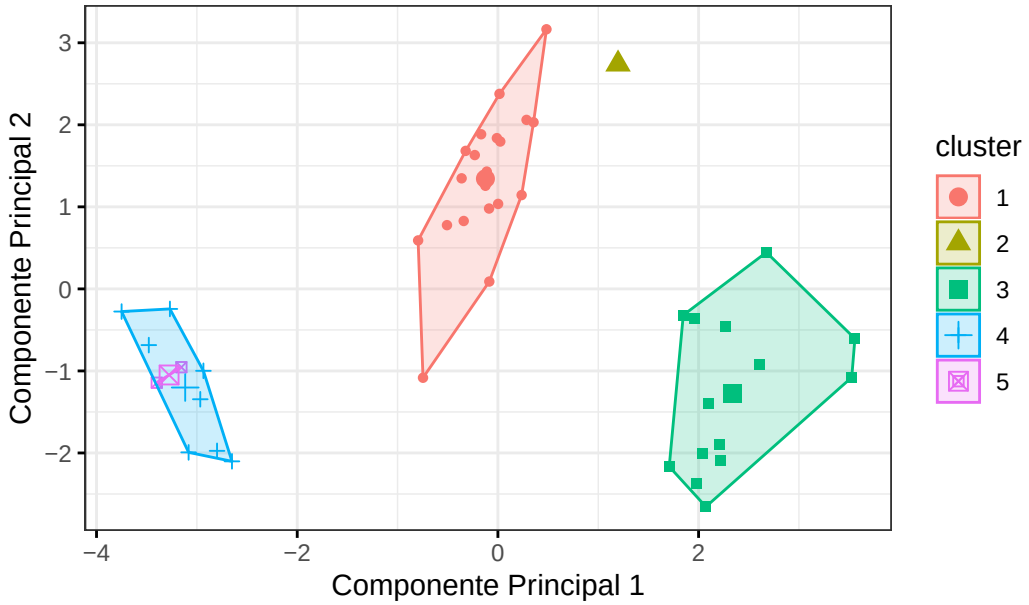


Figura 9: Conglomerados: 5 grupos (liga sencilla).

1.1.1.5 Seis Grupos.

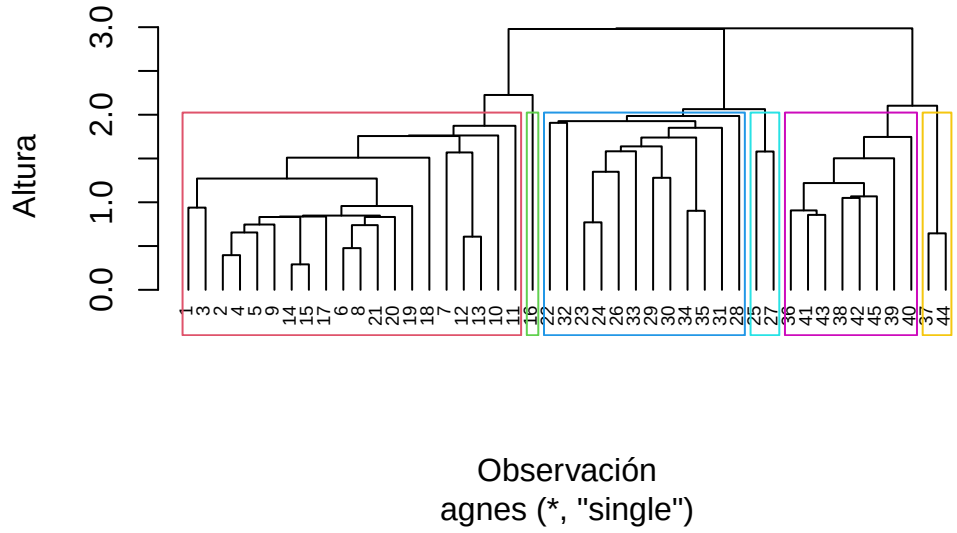


Figura 10: Dendograma: 6 grupos (liga sencilla).

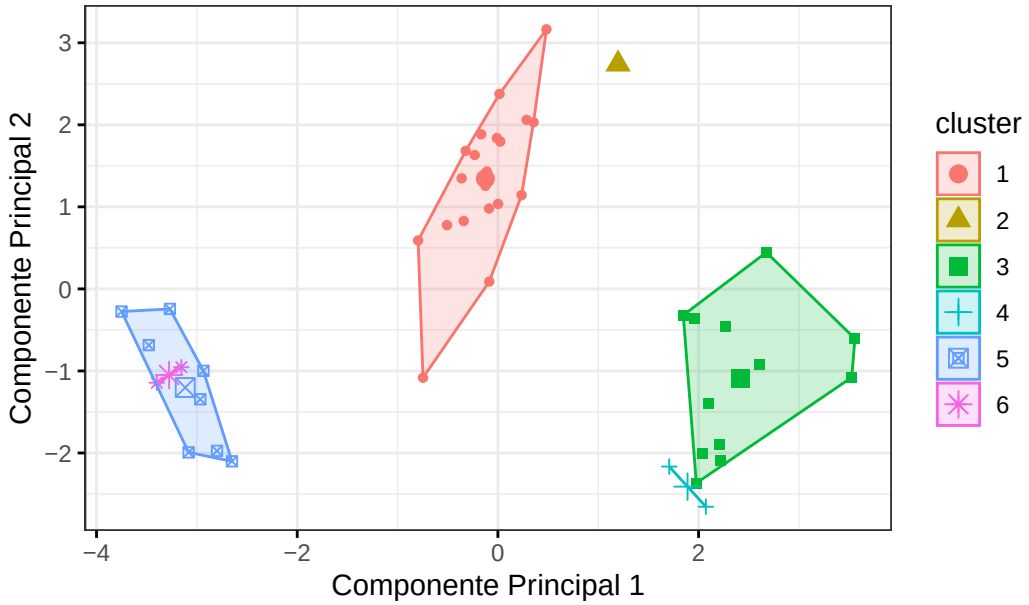


Figura 11: Conglomerados: 6 grupos (liga sencilla).

1.1.1.6 Siete Grupos.

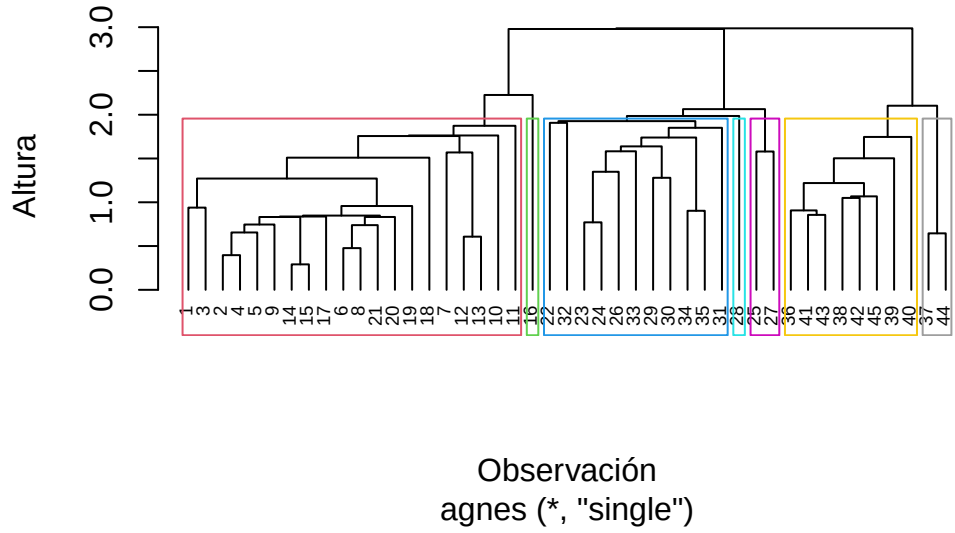


Figura 12: Dendograma: 7 grupos (liga sencilla).

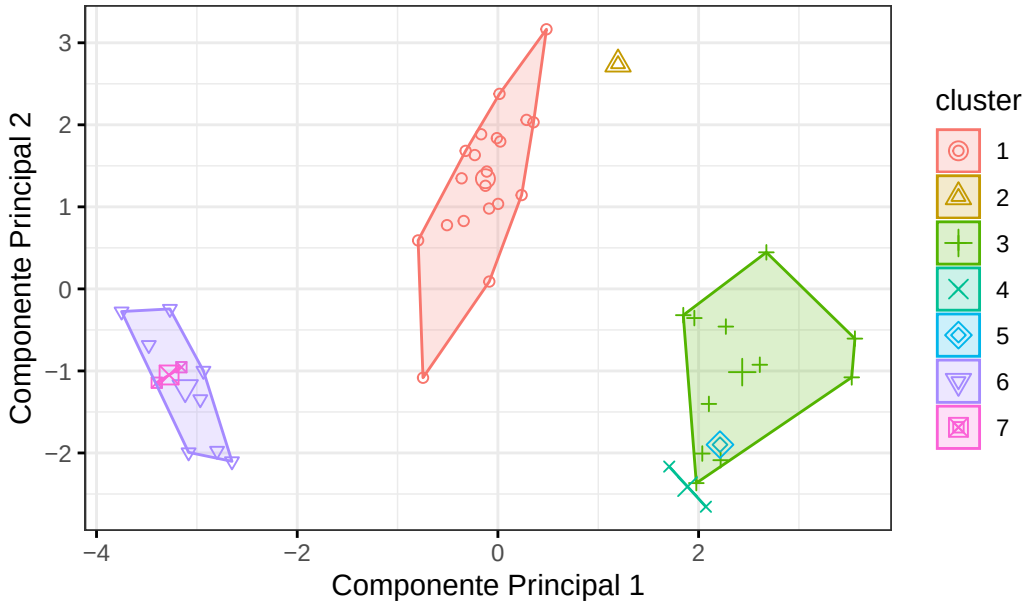


Figura 13: Conglomerados: 7 grupos (liga sencilla).

1.1.1.7 Ocho Grupos.

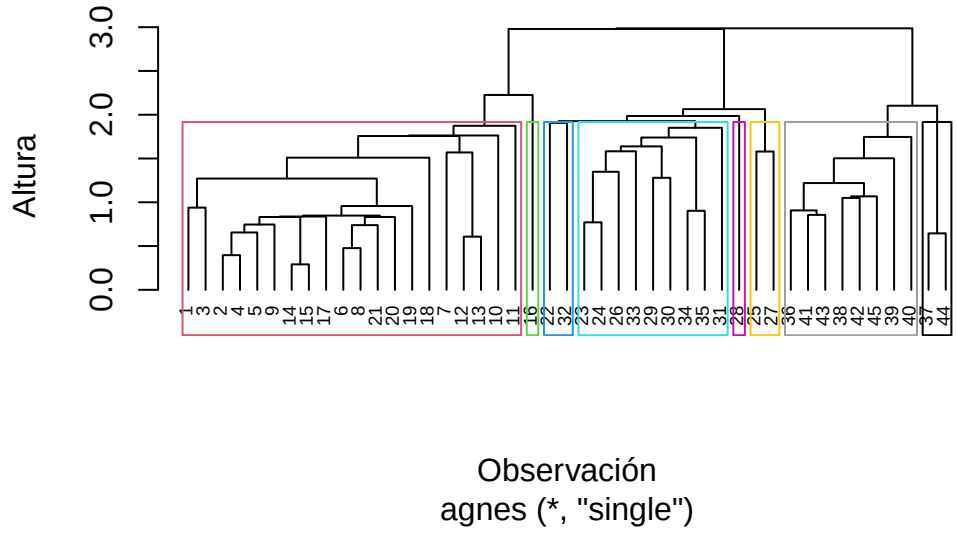


Figura 14: Dendograma: 8 grupos (liga sencilla).

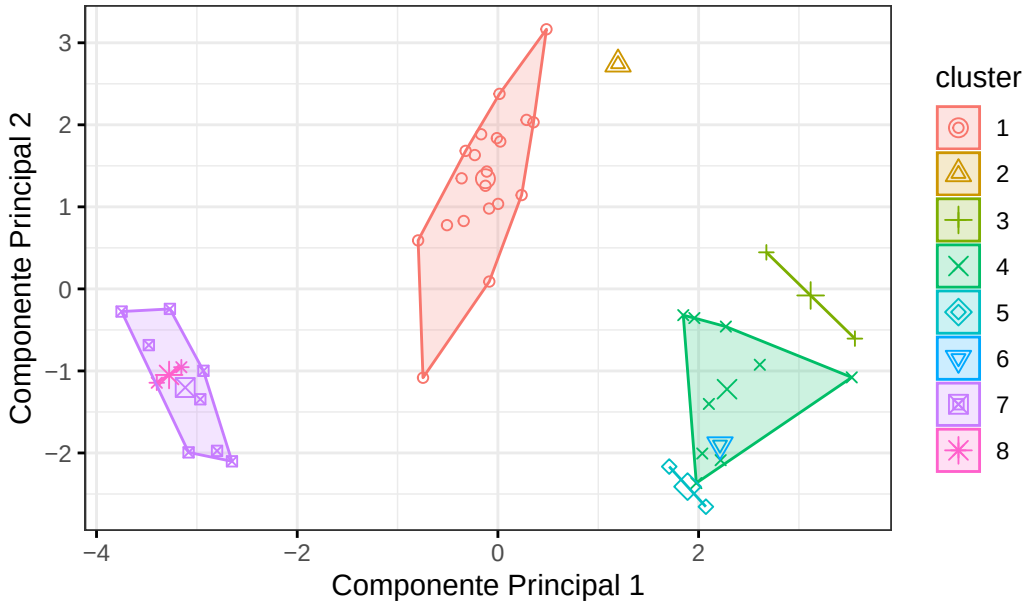


Figura 15: Conglomerados: 8 grupos (liga sencilla).

Se puede observar que usando la liga sencilla, el mejor agrupamiendo se da para $k = 3$ y $k = 4$, dado que las demás tienden a crear una espacie de subgrupo dentro de otro, al menos en la proyección observada en el plano de la primera y segunda componente principal.

1.1.2 Liga Completa.

1.1.2.1 Dos Grupos.

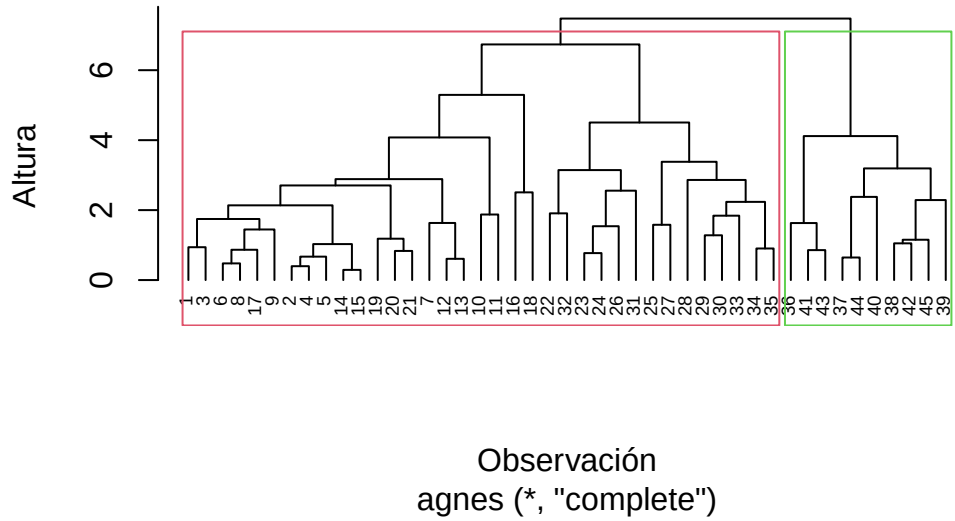


Figura 16: Dendograma: 2 grupos (liga completa).

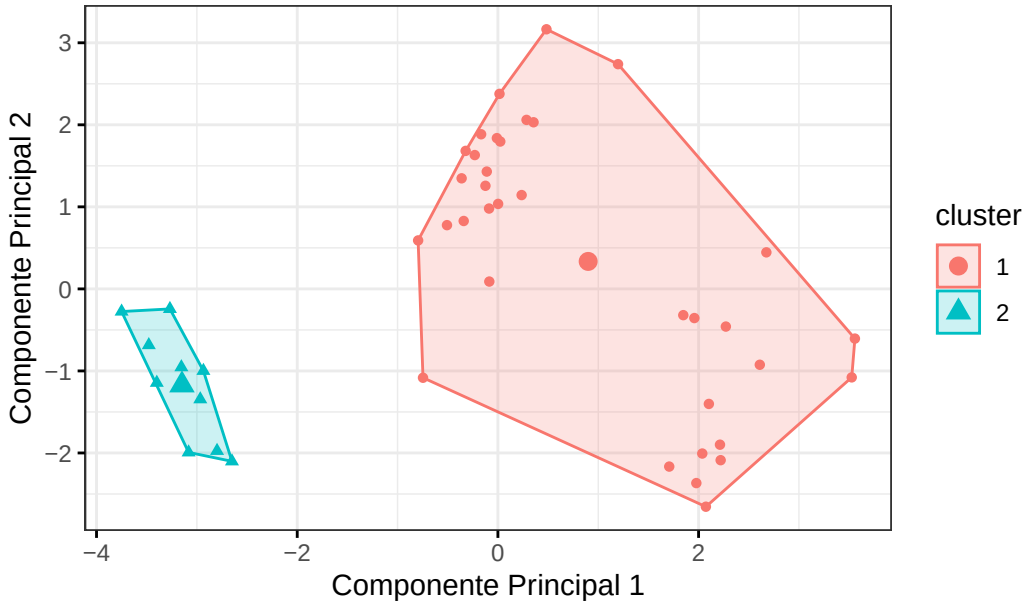


Figura 17: Conglomerados: 2 grupos (liga completa).

1.1.2.2 Tres Grupos.

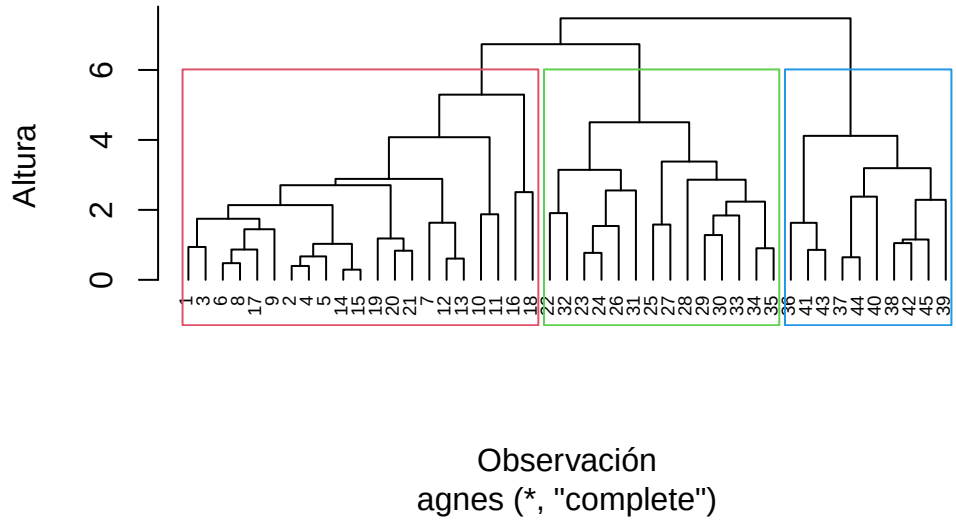


Figura 18: Dendograma: 3 grupos (liga completa).

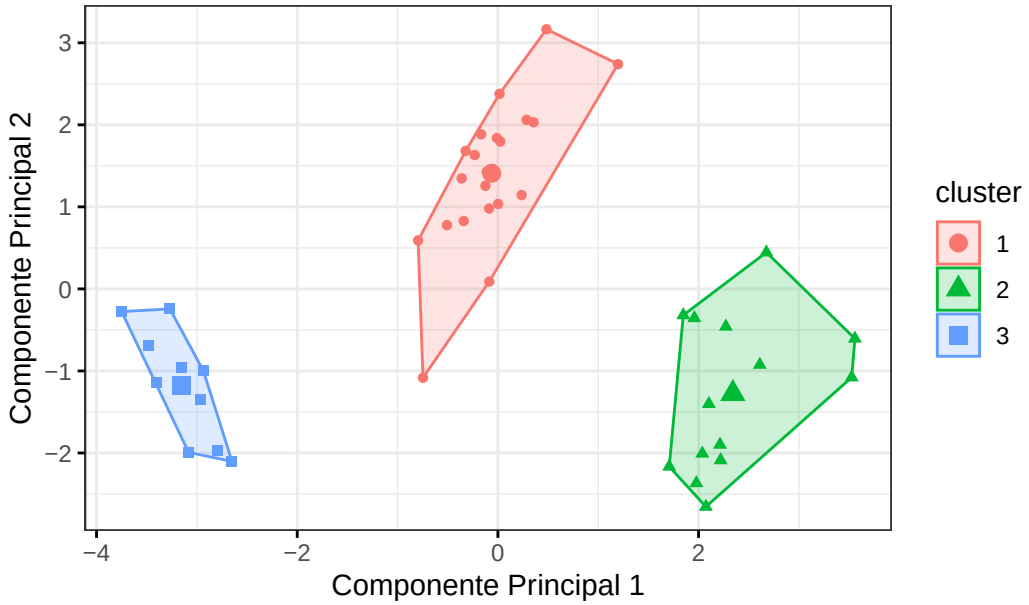


Figura 19: Conglomerados: 3 grupos (liga completa).

1.1.2.3 Cuatro Grupos.

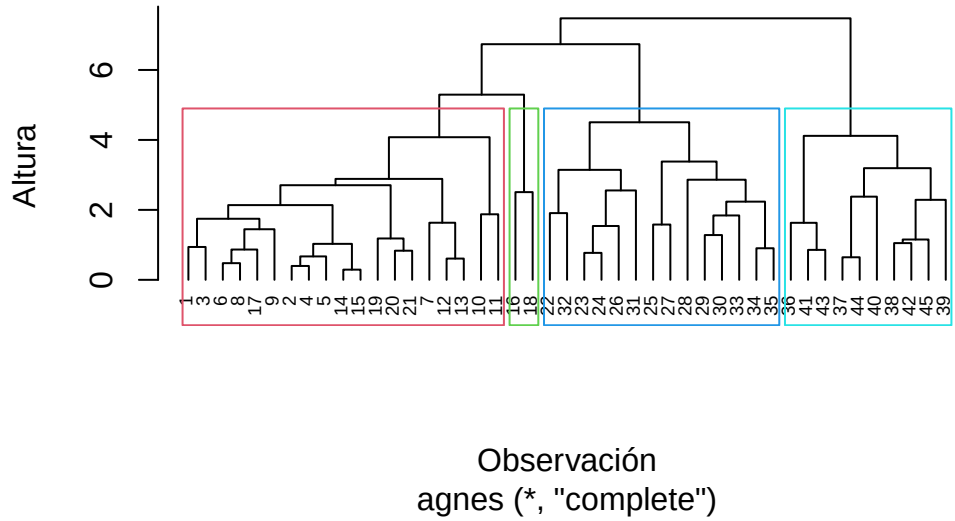


Figura 20: Dendograma: 4 grupos (liga completa).

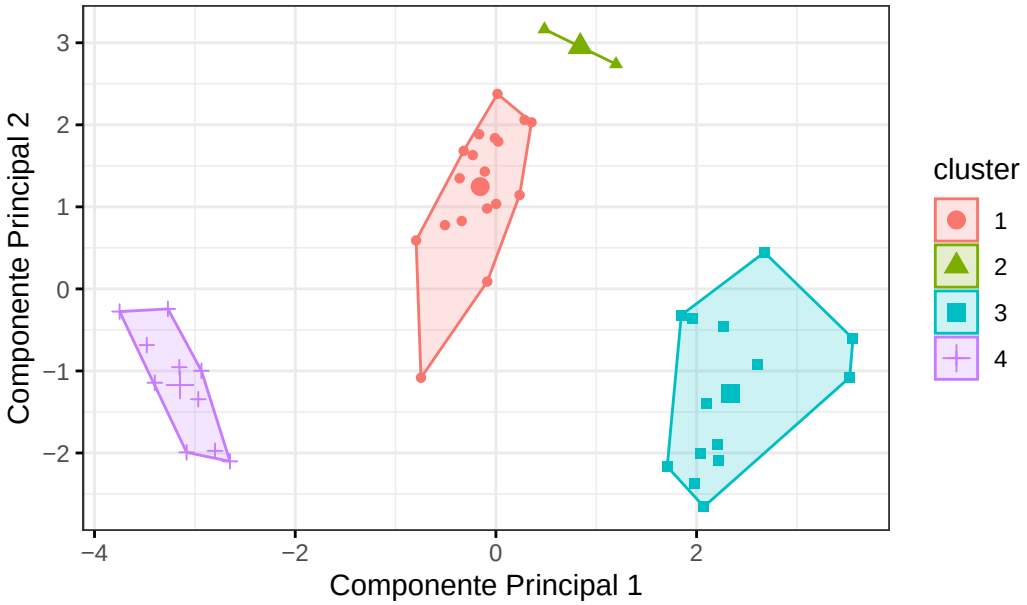


Figura 21: Conglomerados: 4 grupos (liga completa).

1.1.2.4 Cinco Grupos.

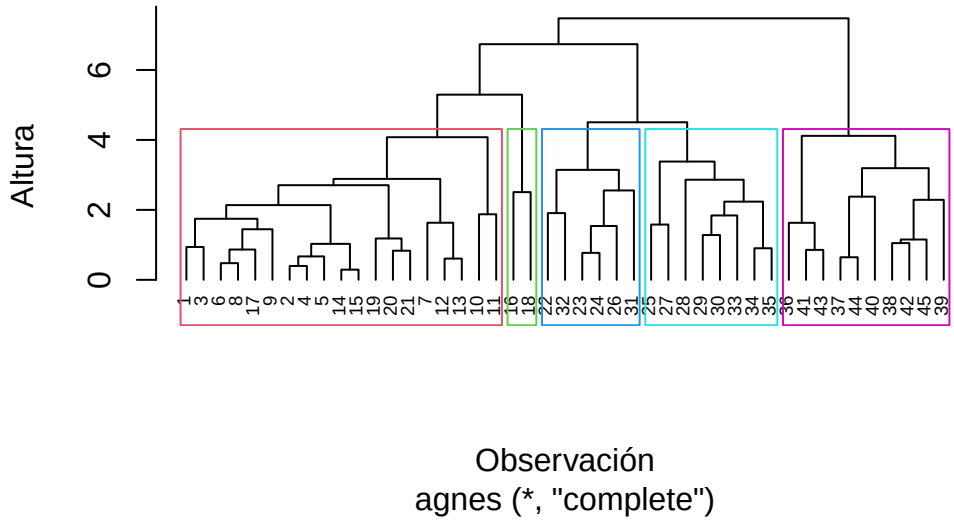


Figura 22: Dendograma: 5 grupos (liga completa).

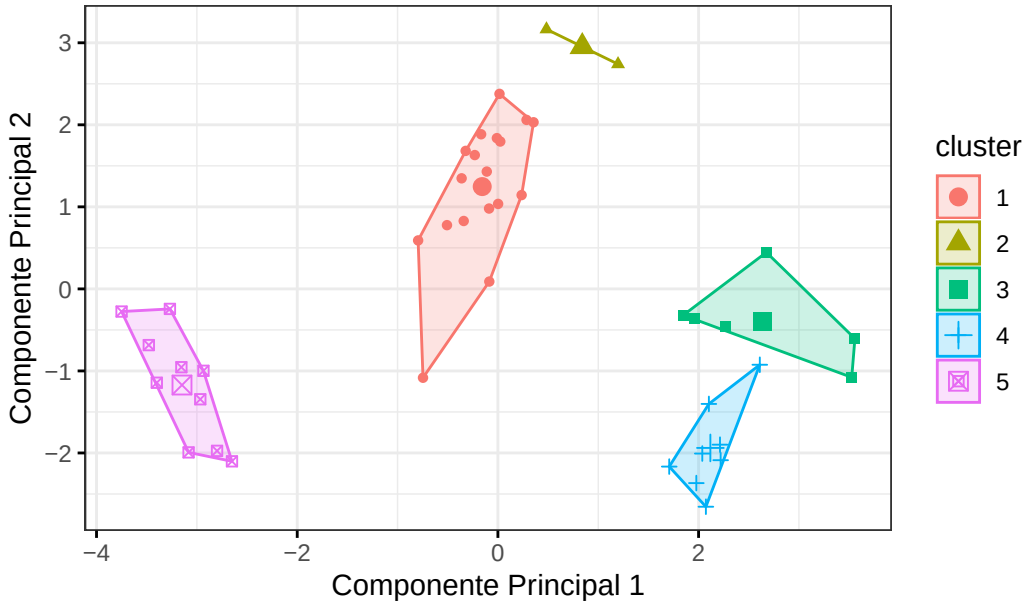


Figura 23: Conglomerados: 5 grupos (liga completa).

1.1.2.5 Seis Grupos.

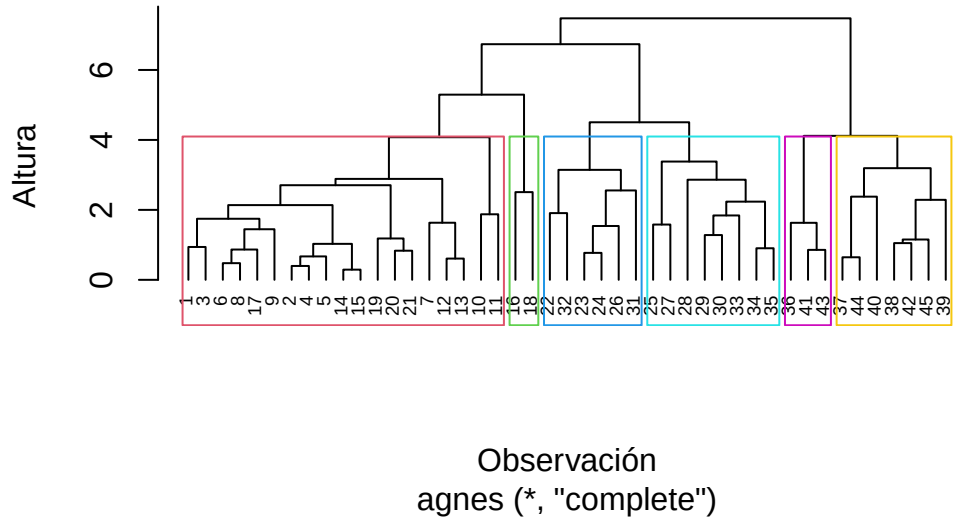


Figura 24: Dendograma: 6 grupos (liga completa).

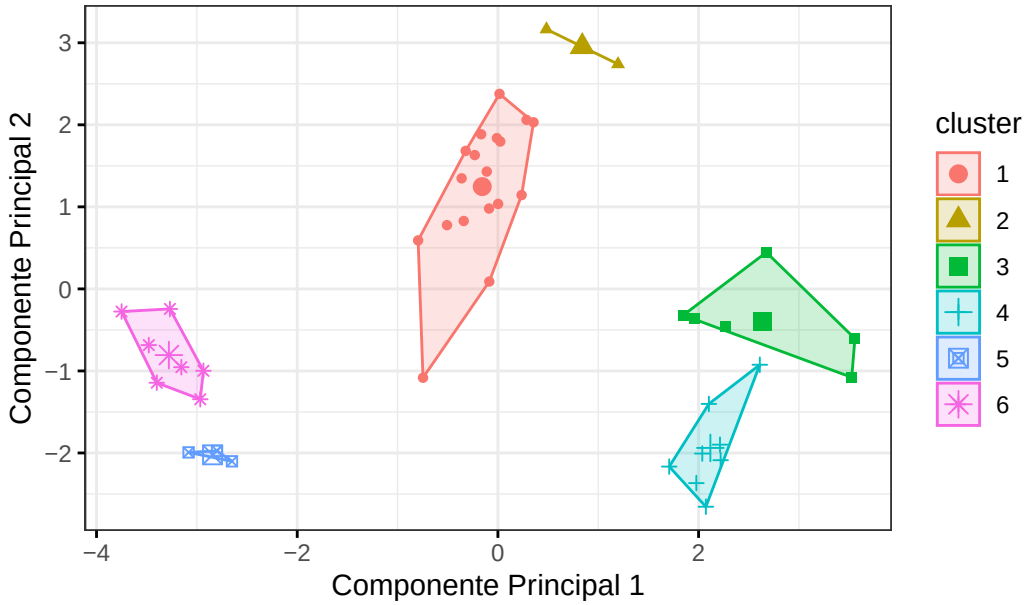


Figura 25: Conglomerados: 6 grupos (liga completa).

1.1.2.6 Siete Grupos.

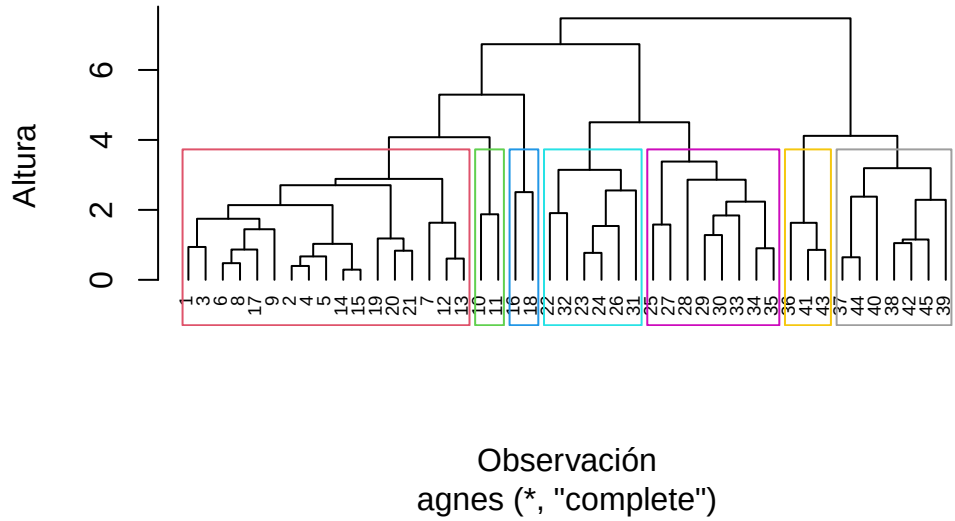


Figura 26: Dendograma: 7 grupos (liga completa).

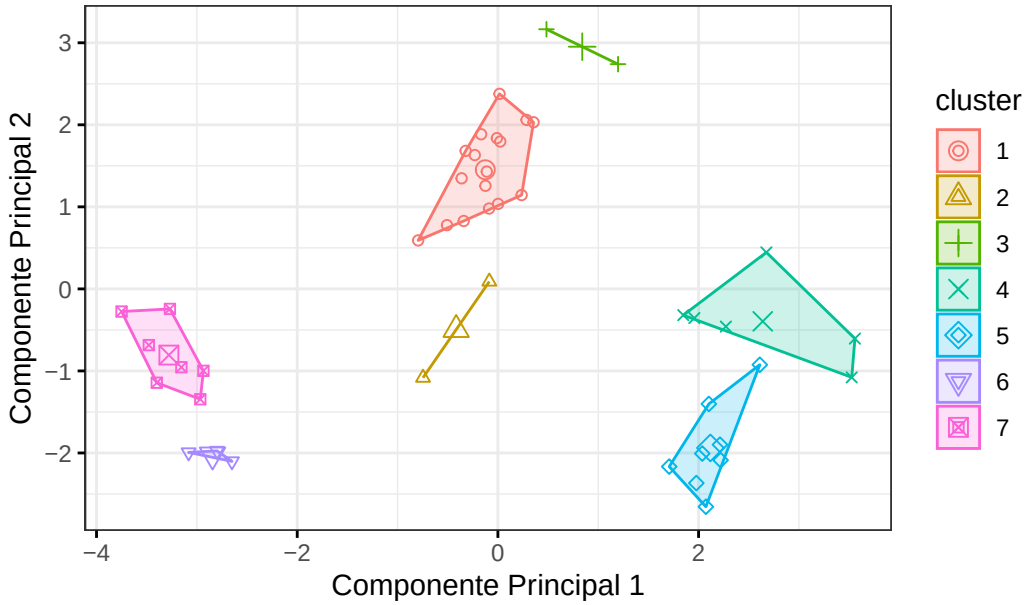


Figura 27: Conglomerados: 7 grupos (liga completa).

1.1.2.7 Ocho Grupos.

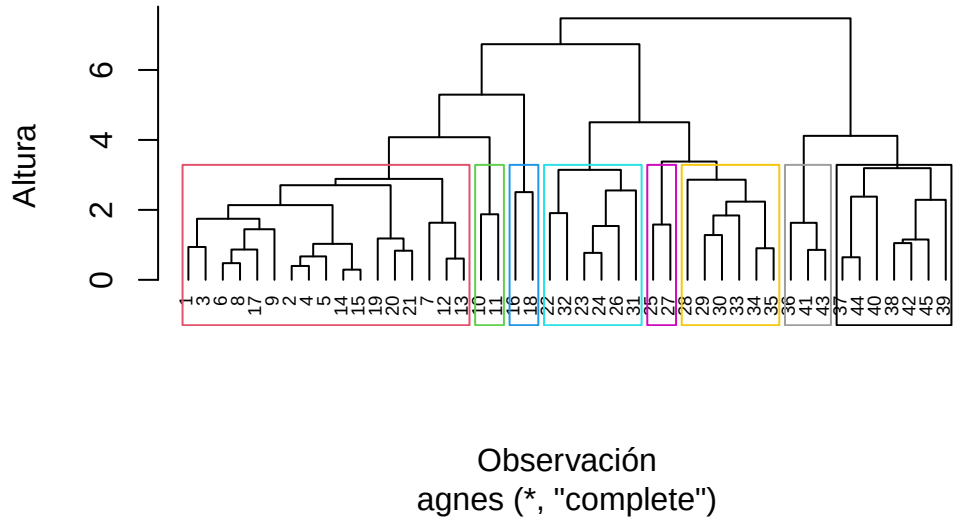


Figura 28: Dendograma: 8 grupos (liga completa).

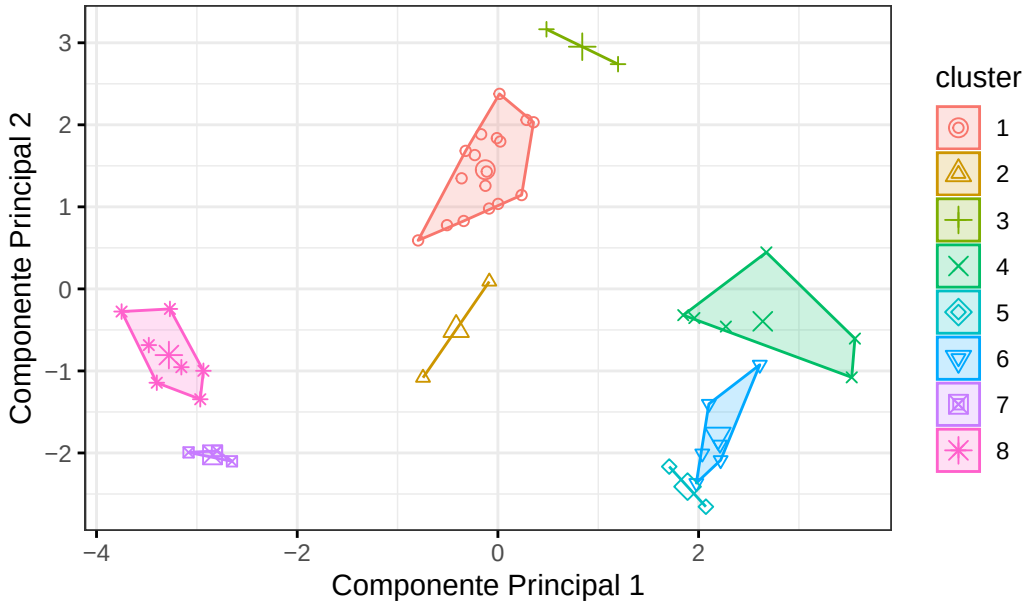


Figura 29: Conglomerados: 8 grupos (liga completa).

Usando la liga completa parece que una buena selección para k podría ser 3 o 7, dado que son el número de grupos que parece diferenciar mejor a las observaciones, al menos usando la liga completa.

1.1.3 Métdo de Ward.

1.1.3.1 Dos Grupos.

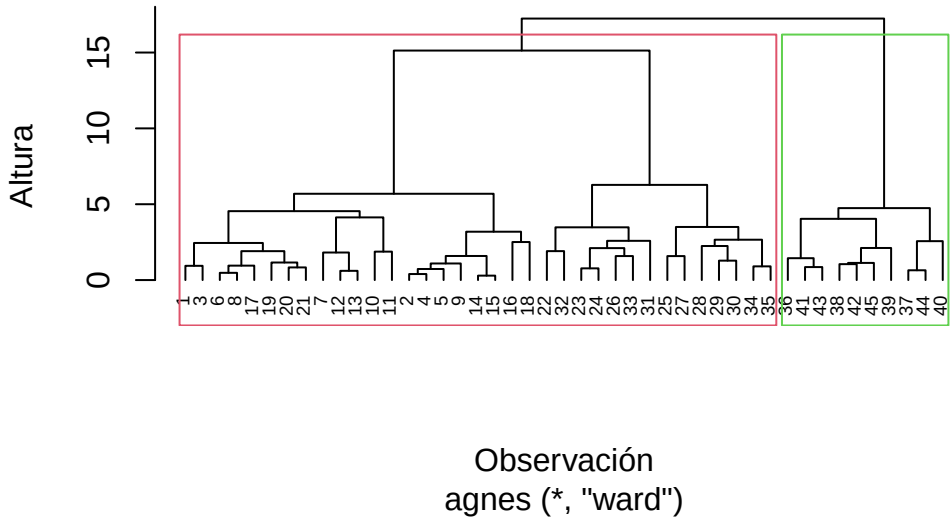


Figura 30: Dendograma: 2 grupos (métdo de Ward).

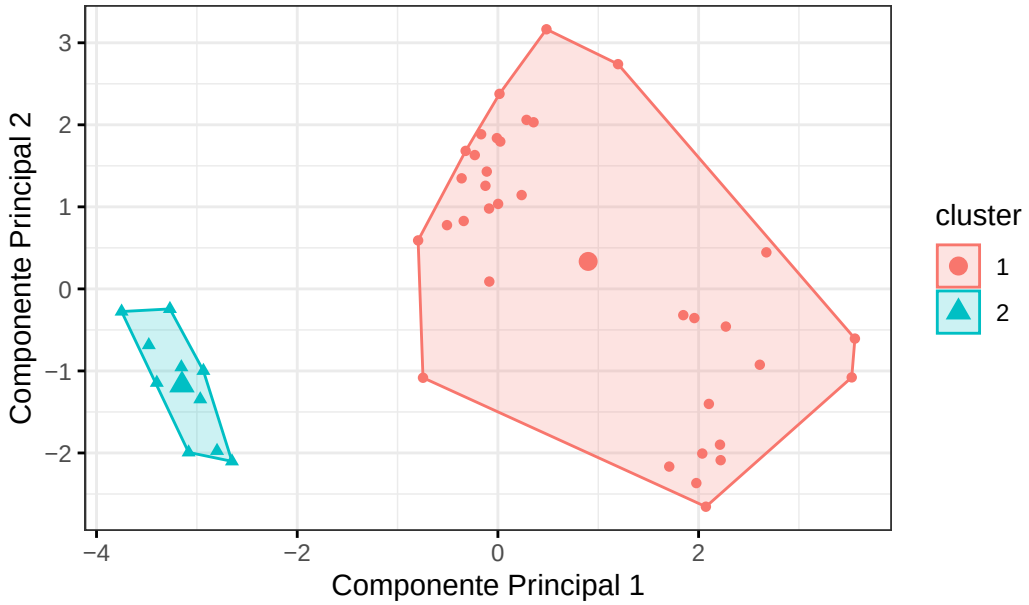


Figura 31: Conglomerados: 2 grupos (métdo de Ward).

1.1.3.2 Tres Grupos.

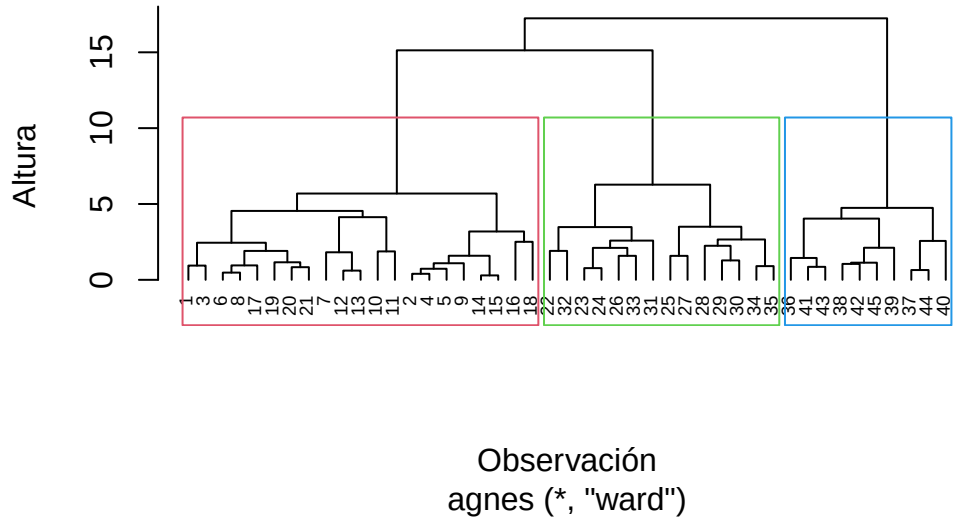


Figura 32: Dendograma: 3 grupos (métdo de Ward).

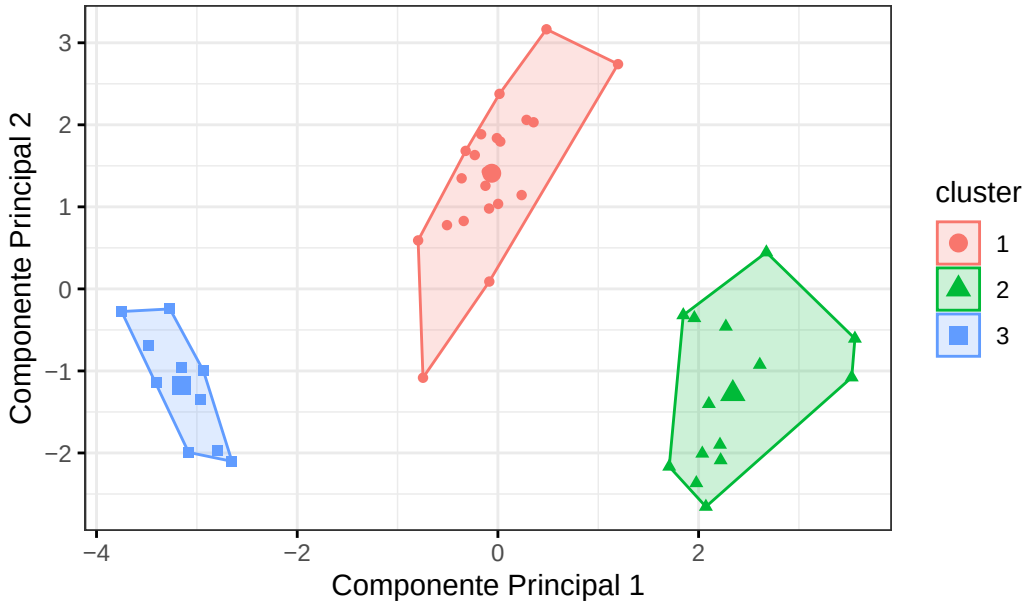


Figura 33: Conglomerados: 3 grupos (métdo de Ward).

1.1.3.3 Cuatro Grupos.

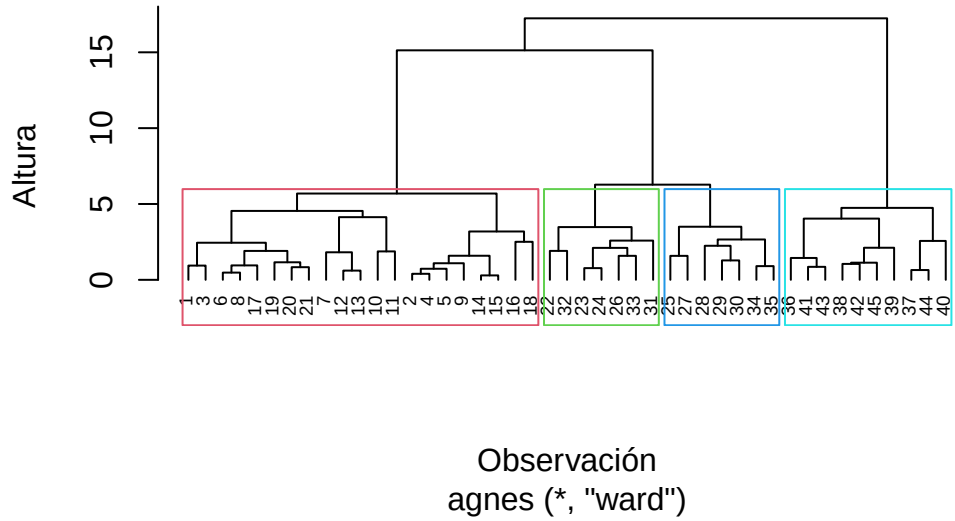


Figura 34: Dendograma: 4 grupos (métdo de Ward).

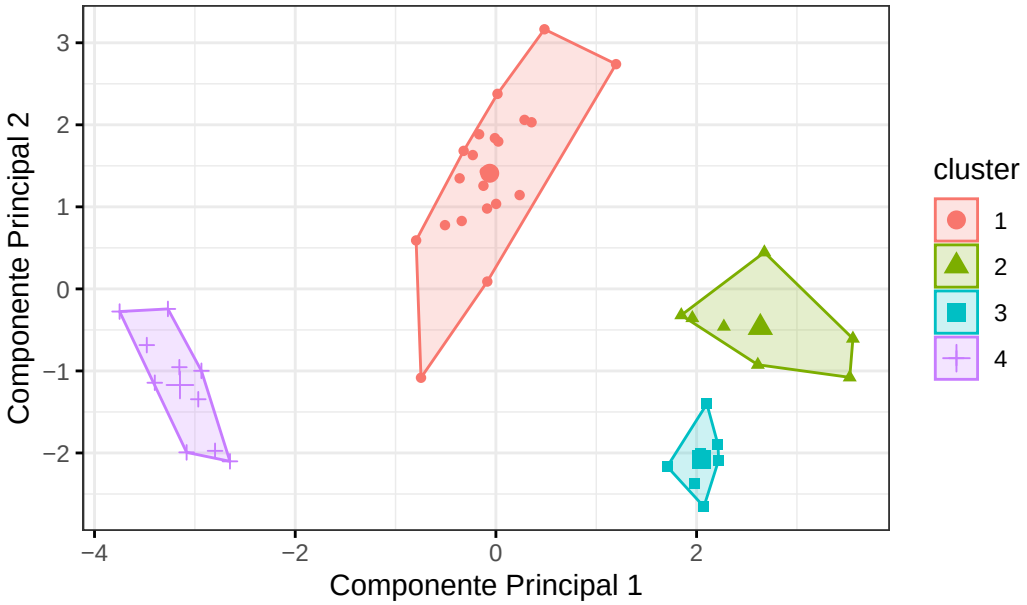


Figura 35: Conglomerados: 4 grupos (métdo de Ward).

1.1.3.4 Cinco Grupos.

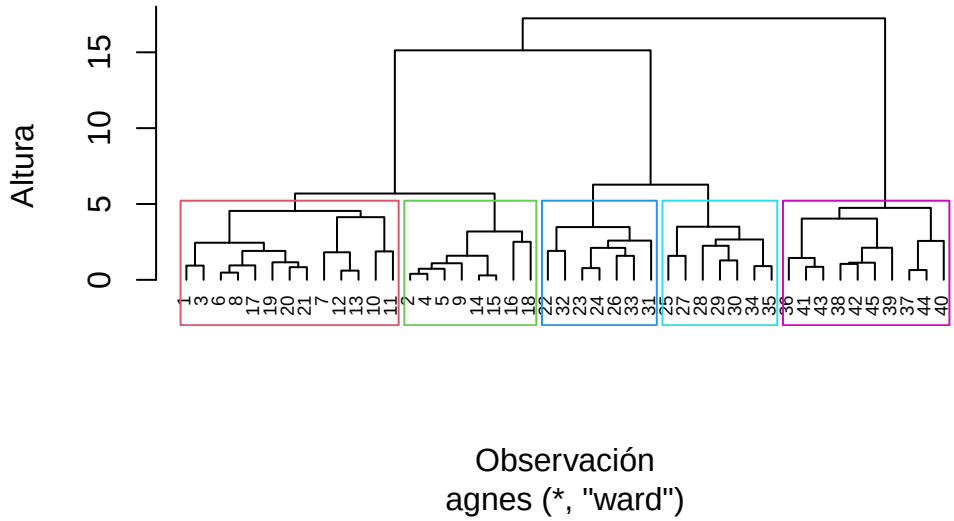


Figura 36: Dendograma: 5 grupos (métdo de Ward).

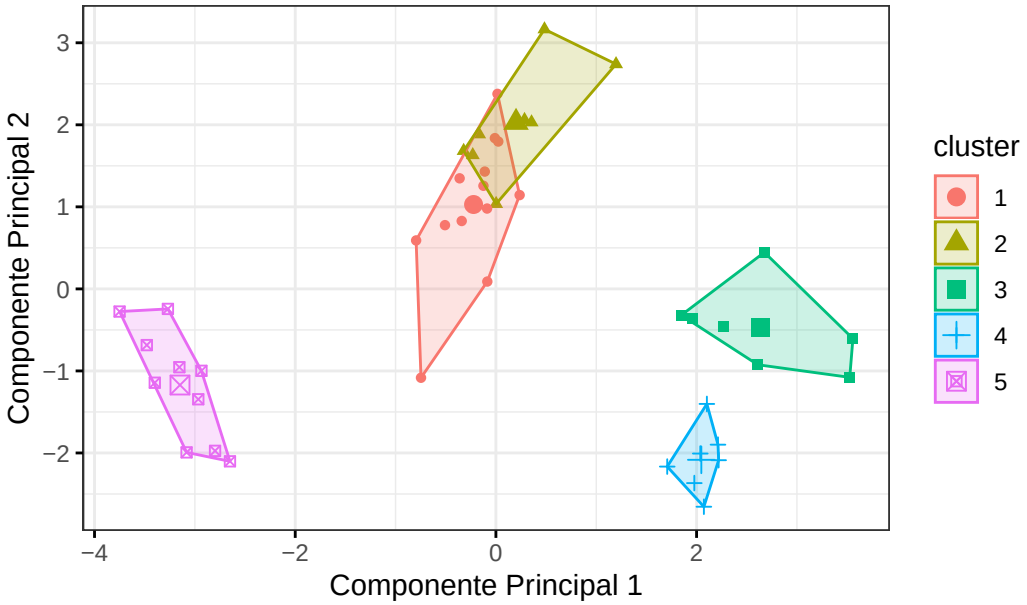


Figura 37: Conglomerados: 5 grupos (métdo de Ward).

1.1.3.5 Seis Grupos.

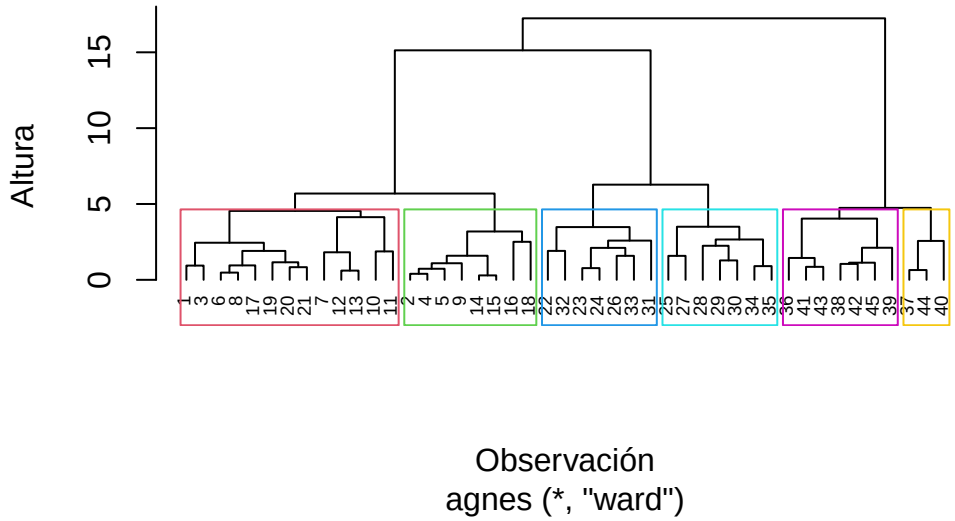


Figura 38: Dendograma: 6 grupos (métdo de Ward).

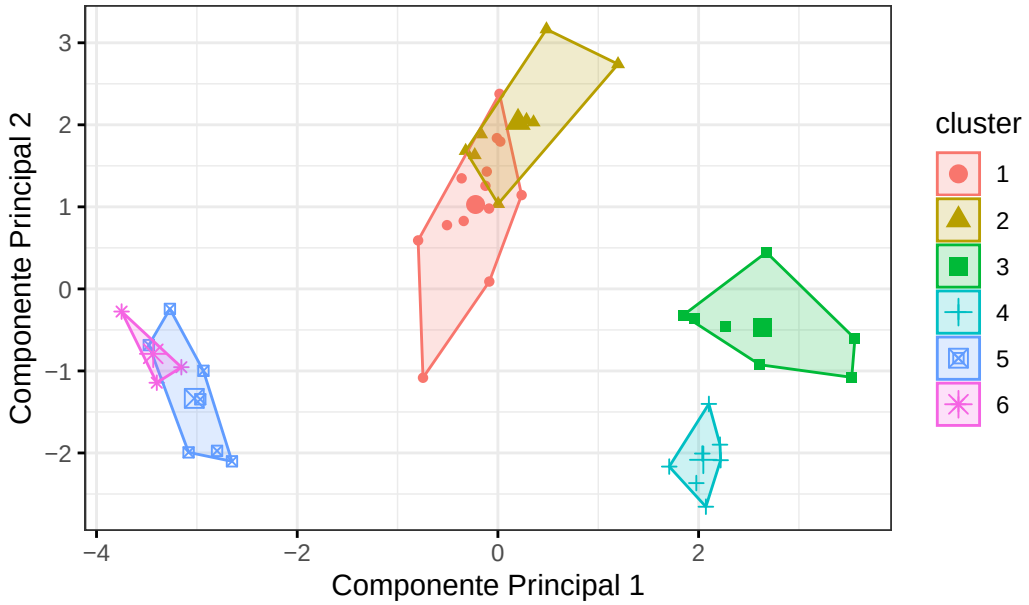


Figura 39: Conglomerados: 6 grupos (métdo de Ward).

1.1.3.6 Siete Grupos.

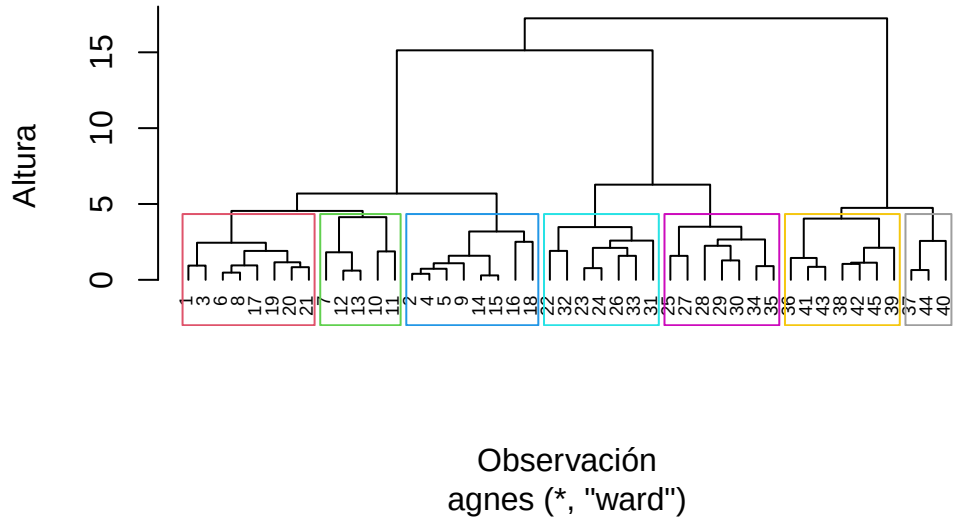


Figura 40: Dendograma: 7 grupos (métdo de Ward).

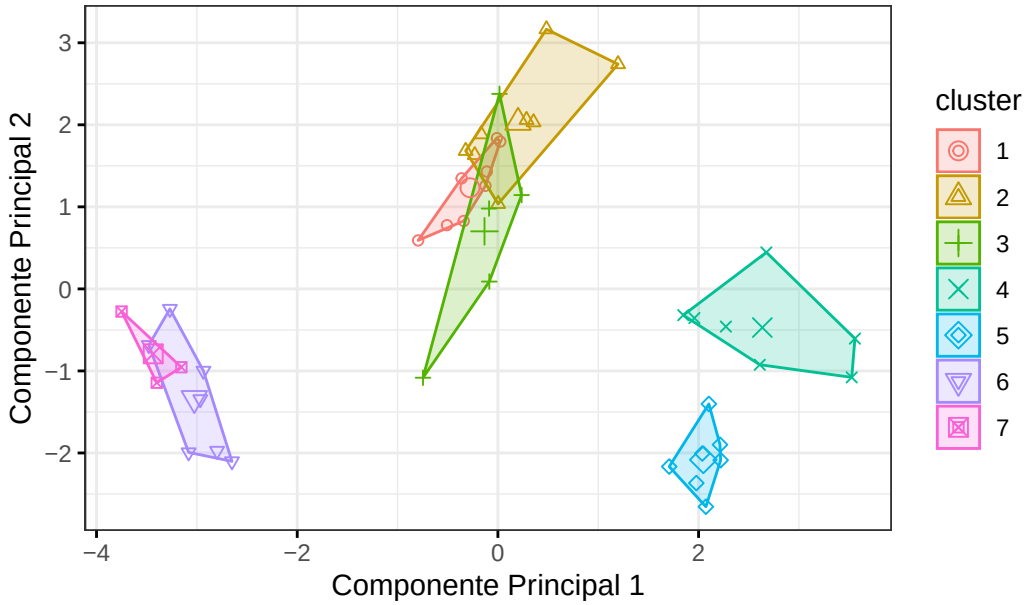


Figura 41: Conglomerados: 7 grupos (métdo de Ward).

1.1.3.7 Ocho Grupos.

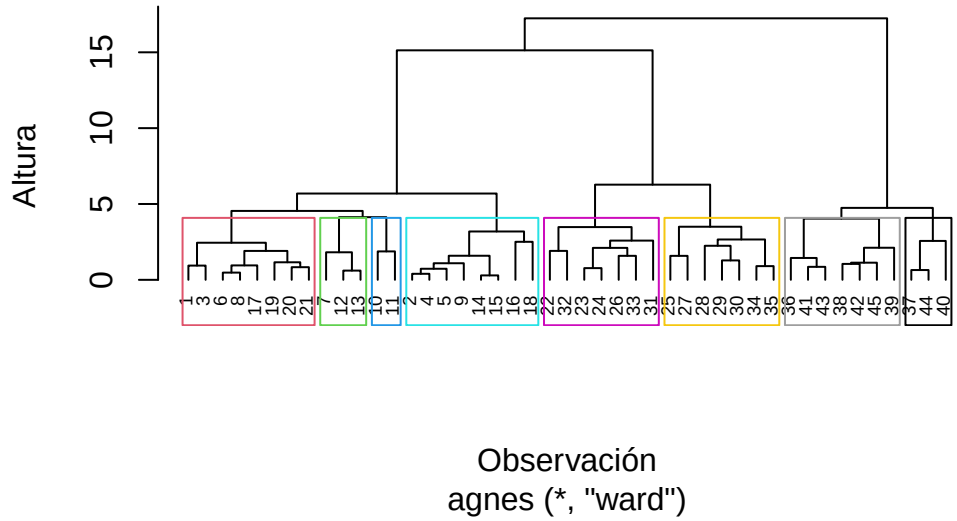


Figura 42: Dendograma: 8 grupos (métdo de Ward).

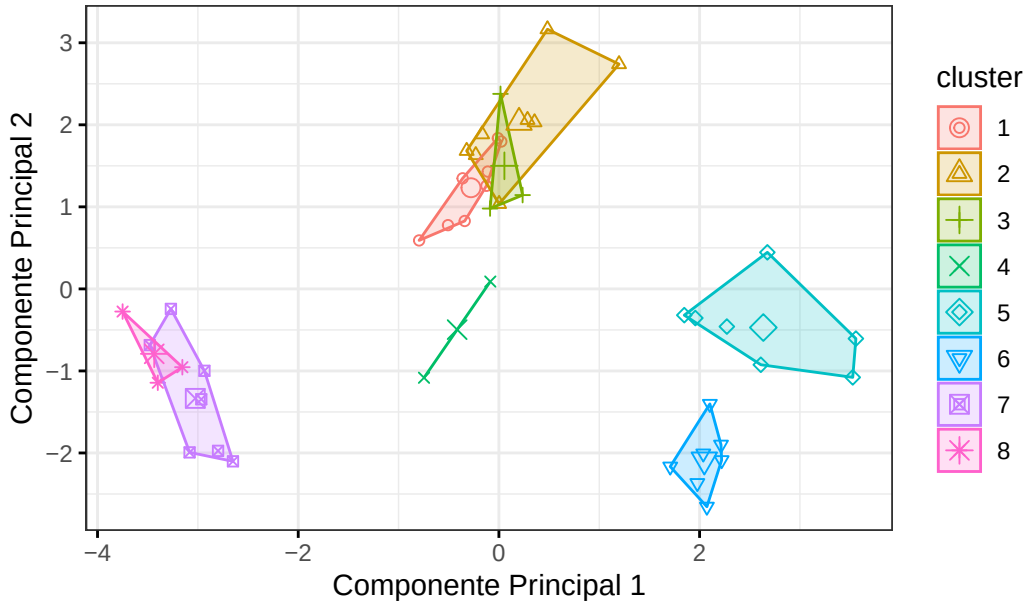


Figura 43: Conglomerados: 8 grupos (métdo de Ward).

El método de Ward parece ser el método que crea de una mejor forma los grupos para nuestros datos, la mejor k en este caso puede ser 3 o 4, ya que dividen de mejor forma las observaciones tanto en el dendograma como en la proyección.

1.2 K-Means.

Realicemos primero un análisis para ver cuál k minimiza de buena manera la función de costo WSS.

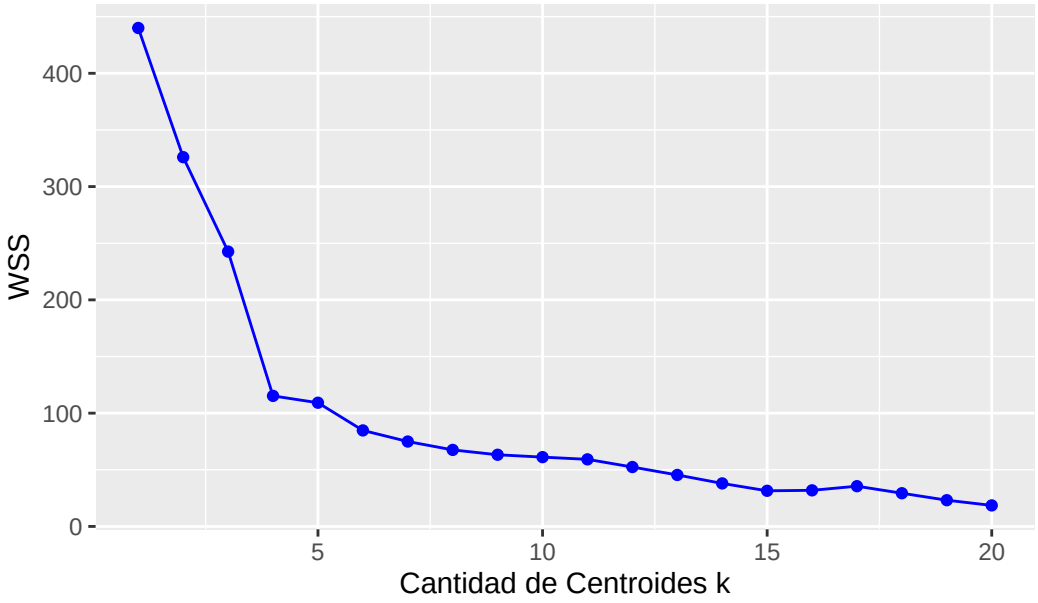


Figura 44: Gráfico de Codo para K-Means.

Podemos ver que la función tiene un salto abrupto entre $k = 3$ y $k = 4$. Para otros k mayores el cambio en la función de costo WSS es menor, por lo cuál podemos suponer que un $k = 4$ debería ser apropiado, lo que va acorde a lo observado en el método de Ward anteriormente.

1.2.1 Comparación de K-Means.

Se comparan las divisiones creadas por un K-Means de k igual a 3, 4, 5 y 6.

```
k3 <- kmeans(data.centered, 3, iter.max = 1000, nstart = 20)
k4 <- kmeans(data.centered, 4, iter.max = 1000, nstart = 20)
k5 <- kmeans(data.centered, 5, iter.max = 1000, nstart = 20)
k6 <- kmeans(data.centered, 6, iter.max = 1000, nstart = 20)
```

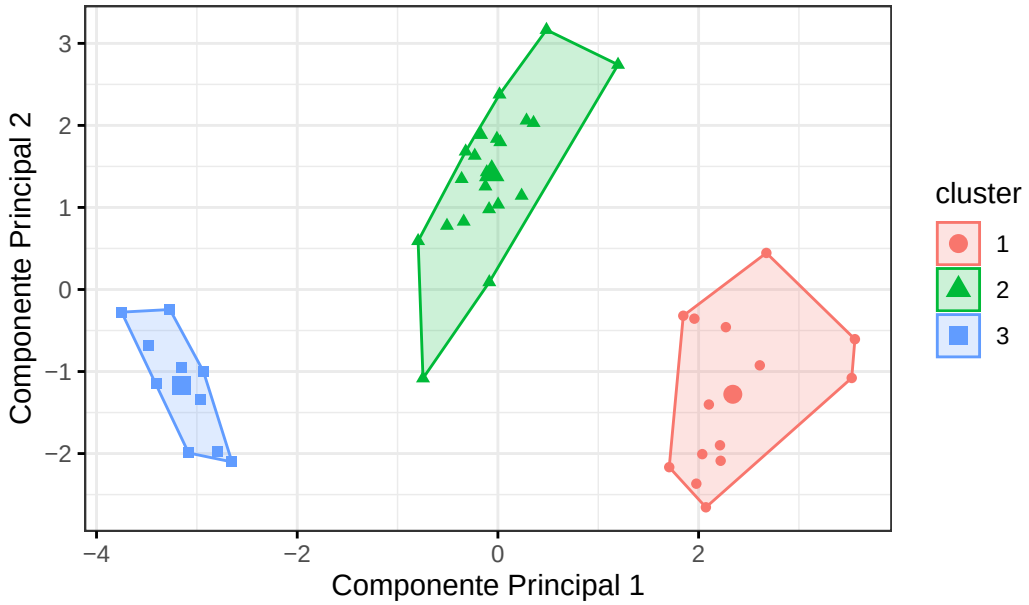


Figura 45: K-Means: 3 grupos.

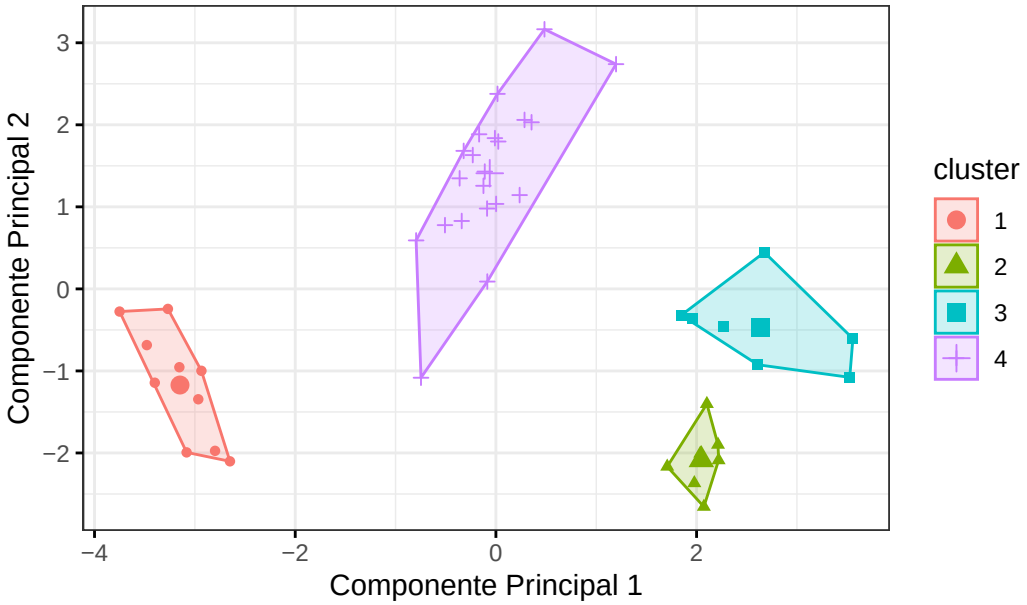


Figura 46: K-Means: 4 grupos.

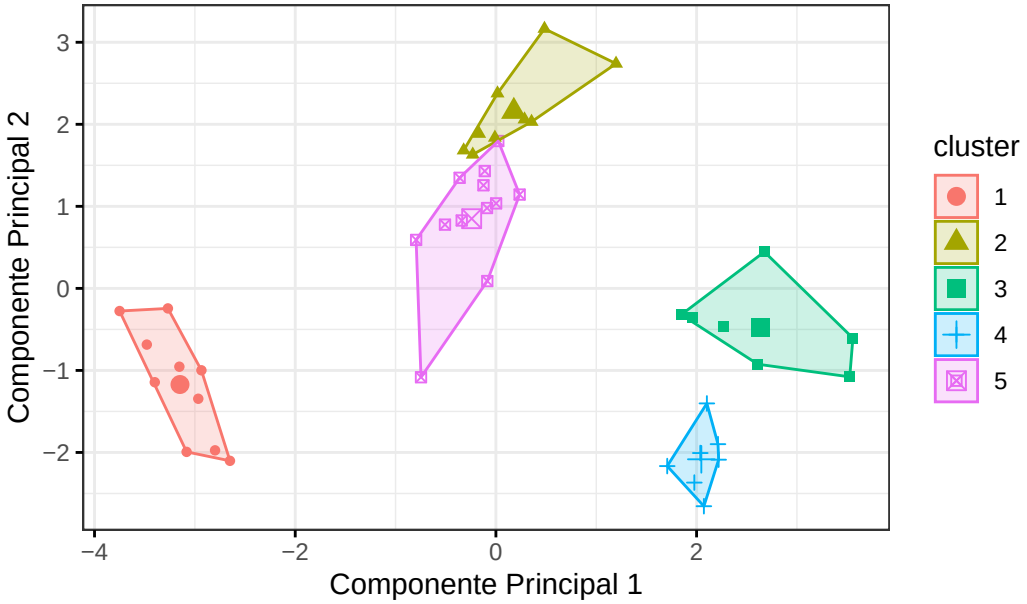


Figura 47: K-Means: 5 grupos.

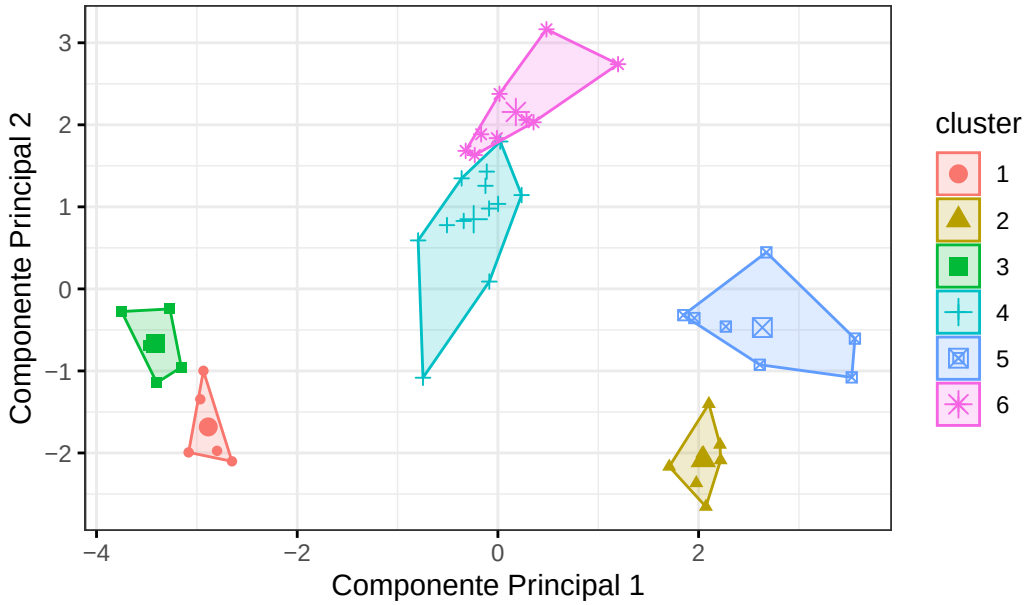


Figura 48: K-Means: 6 grupos.

Como se puede observar, el K-means con 3 y 4 grupos son buenos para la creación de grupos siendo $k = 4$ una muy buena estructura para la formación de los grupos.

1.3 Análsis de Discriminante.

Dado que la principal diferencia entre el discriminante lineal y el cuadrático es la suposición de varianzas iguales, es indispensable el verificar como es la varianza entre las variables.

Tabla 2: Matrix de Covarianza de los Datos.

	Al2O3	Fe2O3	MgO	CaO	Na2O	K2O	TiO2	MnO	BaO
Al2O3	7.3062828	-0.9078520	-3.4490768	0.2845131	0.0078601	-1.4089318	0.3417348	-0.0717160	0.0025340
Fe2O3	-0.9078520	5.7879286	1.6480894	0.7242160	0.2889112	1.2632318	-0.0630879	0.0754472	0.0015156
MgO	-3.4490768	1.6480894	3.0349498	-0.1470693	0.0464387	1.2985023	-0.2151439	0.0638965	-0.0003453
CaO	0.2845131	0.7242160	-0.1470693	0.2063689	0.0417895	0.0217977	0.0134667	0.0030066	0.0003377
Na2O	0.0078601	0.2889112	0.0464387	0.0417895	0.0317710	0.0491909	0.0013598	0.0044514	0.0001944
K2O	-1.4089318	1.2632318	1.2985023	0.0217977	0.0491909	0.7271409	-0.0926318	0.0339407	0.0001775
TiO2	0.3417348	-0.0630879	-0.2151439	0.0134667	0.0013598	-0.0926318	0.0323318	-0.0045737	0.0001270
MnO	-0.0717160	0.0754472	0.0638965	0.0030066	0.0044514	0.0339407	-0.0045737	0.0021903	0.0000251
BaO	0.0025340	0.0015156	-0.0003453	0.0003377	0.0001944	0.0001775	0.0001270	0.0000251	0.0000089

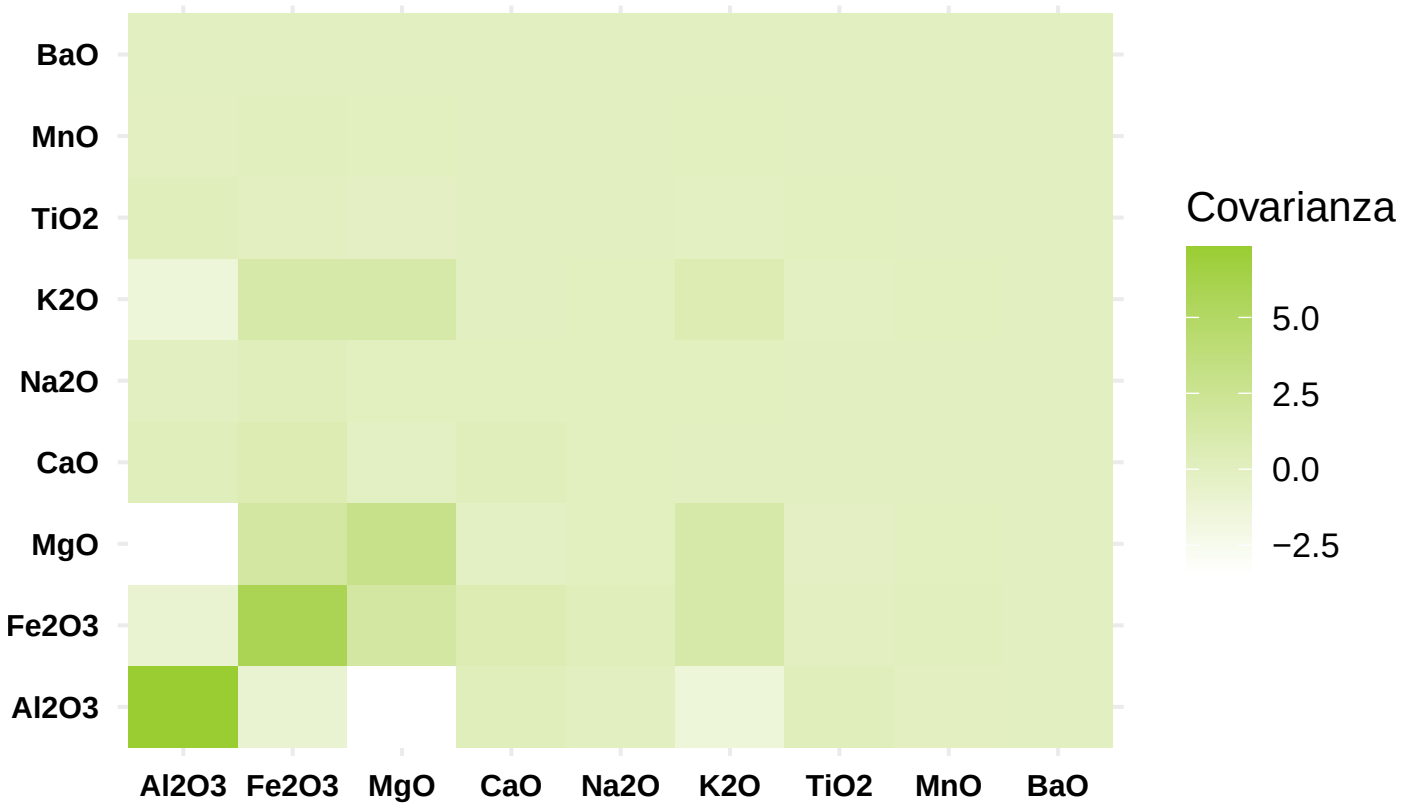


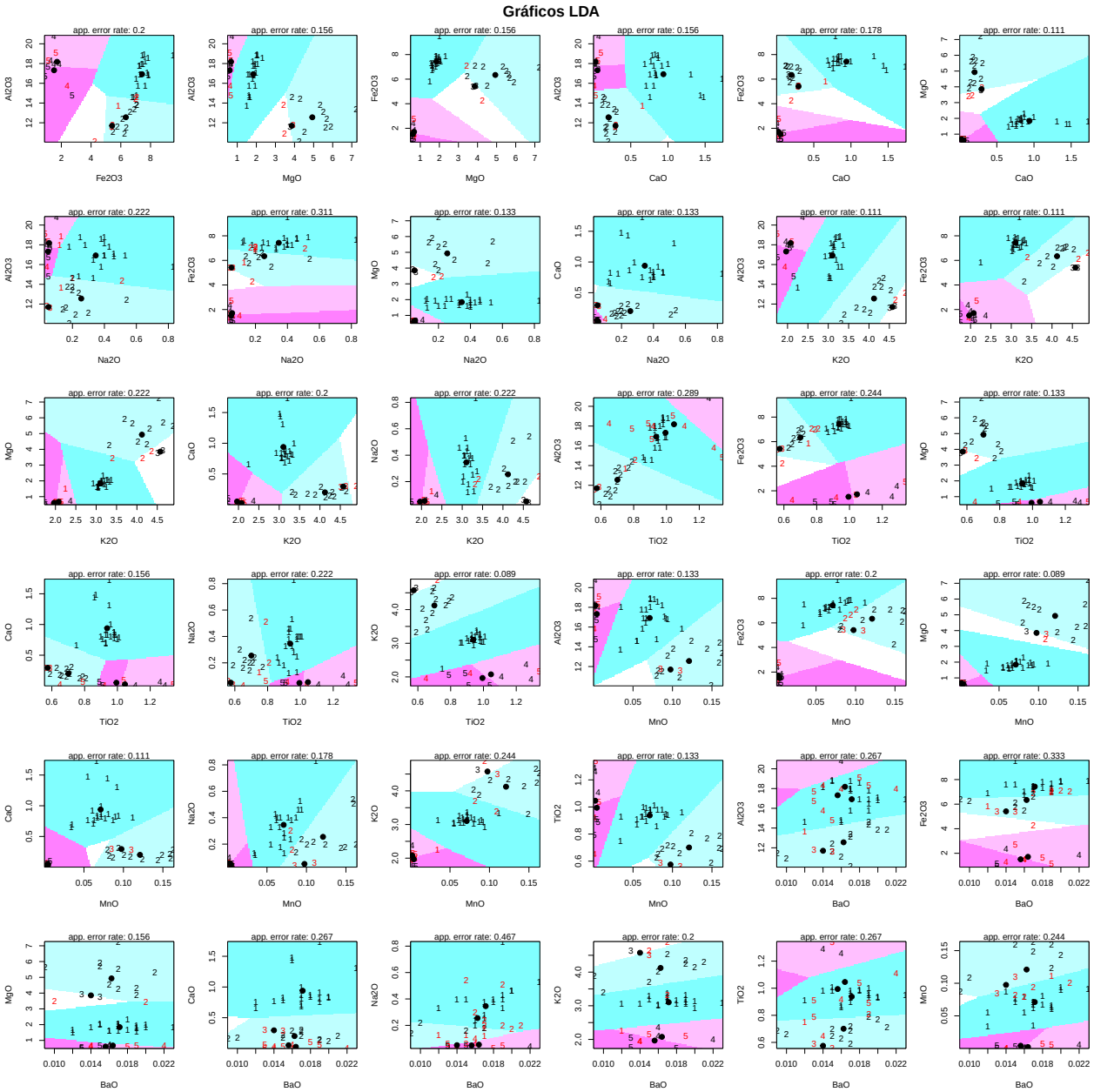
Figura 49: Mapa de Calor de la Matrix de Covarianzas.

Lo que se puede observar es que la varianza es muy pequeña dada la escala de medición de los datos, sin embargo, no se nota una gran diferencia entre las variables a excepción de la concentración del Óxido de Hierro III (Fe_2O_3) y del Óxido de Aluminio (Al_2O_3). Por lo cual se espera que el discriminante lineal y el cuadrático no difieran demasiado entre si.

1.3.1 Discriminante Lineal.

1.3.1.1 Predicción de Horno.

```
partimat(as.factor(kiln) ~ ., data=data,method="lda",
         main = "Gráficos LDA",
         plot.matrix = F)
```



Existen 36 posibles combinaciones para visualizar las 4 reglas de decisión creadas por el discriminante lineal para clasificar el tipo de horno, see observa una tasa elevada de error en varias combinaciones.

	Real	1	2	3	4	5
Predicción						
1		21	0	0	0	0
2		0	12	0	0	0
3		0	0	2	0	0
4		0	0	0	2	1
5		0	0	0	3	4

La proporción de elementos bien clasificados es 0.847619, lo cual indica que el modelo se desempeña bien para clasificar el horno usado para crear las vasijas.

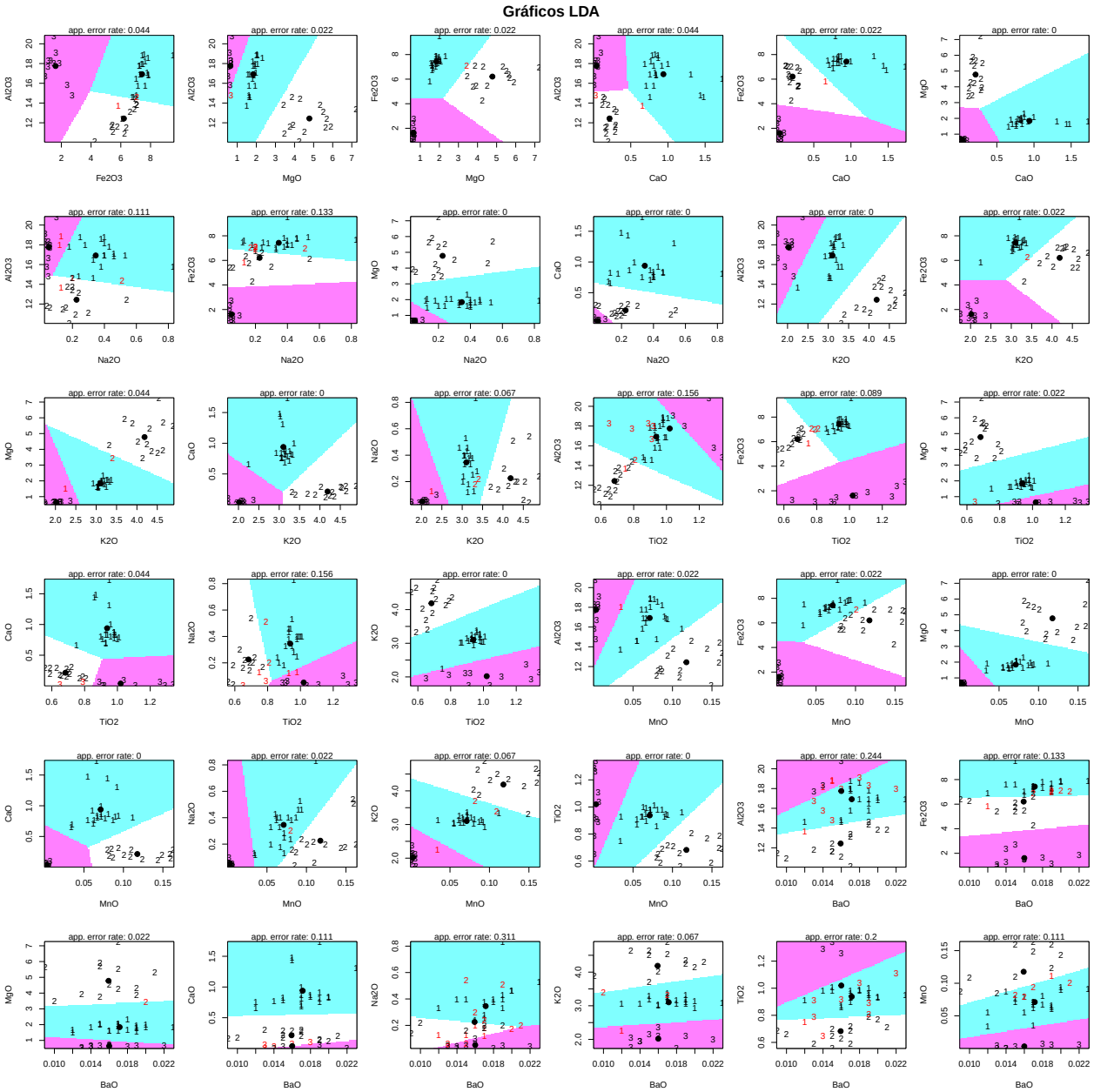
1.3.1.2 Predicción de Región.

Primero se crean las regiones con base en el tipo de horno.

```
region.data <- data %>%
    mutate(region =
                case_when(
                    kiln == 1 ~ "1",
                    kiln == 2 | kiln == 3 ~ "2",
                    kiln == 4 | kiln == 5 ~ "3")) %>%
    dplyr::select(!c(kiln))
```

Ajustamos el modelo de discriminante lineal.

```
partimat(as.factor(region) ~ ., data=region.data, method="lda",
         main = "Gráficos LDA",
         plot.matrix = F)
```



La mayoría de los gráficos indican que la tasa de error es bastante baja para la predicción de la región.

	Real	1	2	3
Predicción				
1		21	0	0
2		0	14	0
3		0	0	10

La proporción de elementos bien clasificados es 1, es decir, el modelo es capaz de clasificar correctamente todas las vasijas en su región correspondiente.

1.3.2 Discriminante Cuadrático.

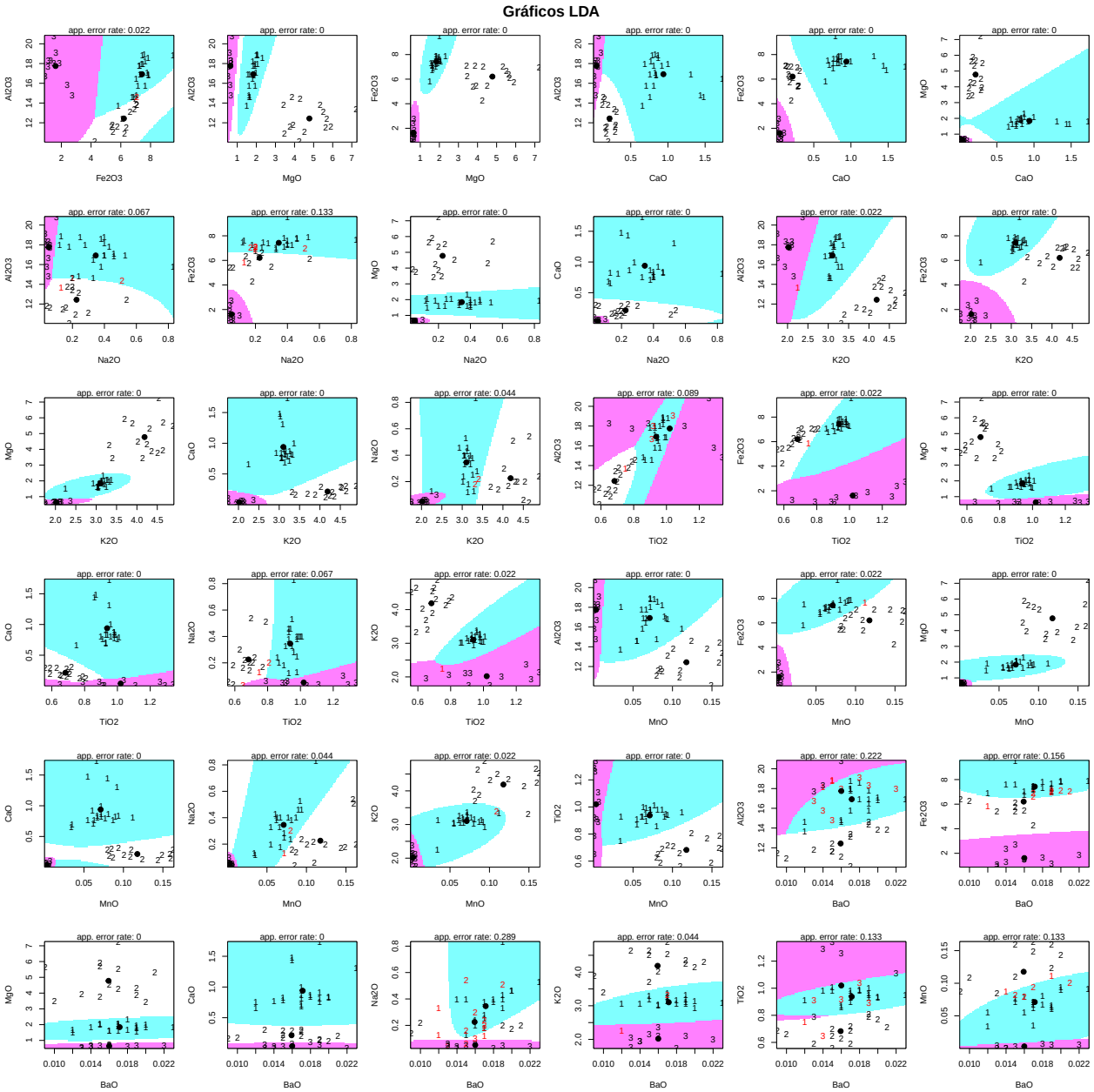
1.3.2.1 Predicción de Horno.

Dado que para la clase 3 de hornos (variable `kiln` en los datos), solo tiene 2 observaciones, la función `MASS::qda` no se puede correr debido a la pequeña cantidad de datos, por lo cual se omitirá su estimación.

1.3.2.2 Predicción de Región.

Ajustamos el modelo de discriminante cuadrático.

```
partimat(as.factor(region) ~ ., data=region.data, method="qda",
         main = "Gráficos LDA",
         plot.matrix = F)
```



La mayoría de los gráficos indican que la tasa de error es bastante baja para la predicción de la región.

	Real	1	2	3
Predicción				
1		21	0	0
2		0	14	0
3		0	0	10

La proporción de elementos bien clasificados es 1, que es idéntico al discriminante lineal.

1.4 Conclusión.

A través de los distintos análisis de conglomerados se pudo detectar que las vasijas se pueden agrupar en al menos 3 grupos distintivos lo cual no concuerda con el número de hornos pero si con el número de regiones de las cuales proviene, de hecho se logró un modelo perfecto en el discriminante lineal y cuadrático cuando se predice la región, por tanto se concluye que la composición química de las vasijas difiere por la región y no por el tipo de horno.