Tarea 2

Christian Badillo Luis Nuñez Luz Maria Santana Sealtiel Pichardo

Tabla de contenidos

1 Parte 1

1. Observa que algunos datos están en segundos y otros en minutos. Comenta los problemas que esto puede generar en el análisis de componentes principales.

Utilizar variables con diferentes unidades de medición podría traer problemas en el análisis de componentes principales, ya que el análisis podría ser sencible a alguna de ellas. Por lo que es conveniente trabajar con los datos centrados y estandarizados, es decir, con la matriz R.

2. Calcula la varianza de cada una de las variables y haz el cociente de la máxima entre la mínima. Comenta (¿qué variables se deben transformar para evitar el cociente tan grande? ¿Cómo transformarlas?).

Vamos a cargar la base de datos.

```
athletic <- read.csv("athletic.csv")
# Primero vamos a poner todo en minutos
athletic[,1:3] <- athletic[,1:3] / 60
```

Cálculo de varianzas.

```
varianzas <- apply (athletic, 2, var)
varianzas
```

```
      X100m        X200m        X400m        X800m       X1500m       X5000m
3.430625e-05 1.288943e-03 5.896945e-04 4.055758e-03 2.430774e-02 6.418581e-01
     X10000m     Marathon
3.246071e+00 8.513404e+01
```

```
var_max <- max(varianzas)
var_max
```

```
[1] 85.13404
```

```
var_min <- min(varianzas)
var_min
```

```
[1] 3.430625e-05
```

```
cociente_varianzas <- var_max /  var_min
cociente_varianzas
```

```
[1] 2481590
```

El cociente calculado es de 2481590, es un número demasiado grande que indica que la diferencia entre varianzas es muy grande. Esto ocurre por la variable Marathon, ya que es la que posee más varianza (85.134042), incluso entre las variables que también están en minutos. Esto es porque su rango de valores inicia desde el 128 hasta el 164.7. Por ello, esa variable es la que se podría normalizar (restar a cada valor su media y divir entre su varianza) para poder utilizar la matriz de varianzas y covarianzas. En R se tiene la función sclae() para realizar ese proceso

```r
athletic_normalizada <- scale(athletic)
```

3. Calcula la matriz de correlaciones (no imprimir). Comenta las relaciones entre las variables.

```r
matriz_correlaciones <- cor(athletic)
```

De manera general podemos decir que las variables se correlacionan de manera positiva ya que no se tienen valores menores a cero. Aquellas variables con mayor correlación (cercanas a 1) son 100m con 400m, 400m con 800m, 400m con 1500m, 800m con 1500m, 500m con 1500m (0.928114), 1500m con 10000m (0.9337307), 5000m con 10000m (0.9738873) siendo estas 3 últimas las más grandes.

En general, la mayoría de las variables se encuentran asociadas de manera positiva y con una relación fuerte. La más débil es la de 0.26 entre la prueba de 200 y 400m Y en general es la prueba de 200m la que tiene menor correlación con las demás

4. Calcule los componentes principales. (¿es recomendable usar la matriz de correlaciones?, explica). Describe brevemente los resultados.

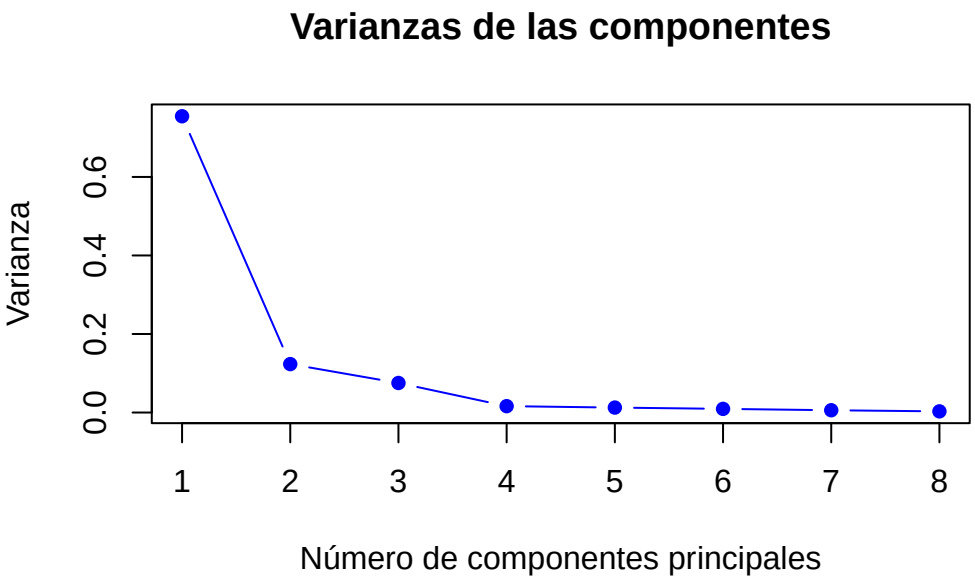
```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     2.4570 0.9931 0.77589 0.36044 0.31726 0.27263 0.21559
Proportion of Variance 0.7546 0.1233 0.07525 0.01624 0.01258 0.00929 0.00581
Cumulative Proportion  0.7546 0.8779 0.95313 0.96937 0.98195 0.99125 0.99706
                           PC8
Standard deviation     0.15349
Proportion of Variance 0.00294
Cumulative Proportion  1.00000
```

De la matriz S, haciendo el cociente de la varianza máxima con la mínima se obtiene 20990.91 que es un valor muy grande y hace que usar la matriz S no sea viable. Por lo cual conviene utilizar la matriz de correlaciones.

5. Haz una gráfica de la varianza de las componentes (screeplot).Comenta.

6. Explica tu criterio para selección de número de componentes. ¿qué proporción de la varianza total se explica con el número de variables que seleccionaste?

Vamos a extraer la varianza de cada componente

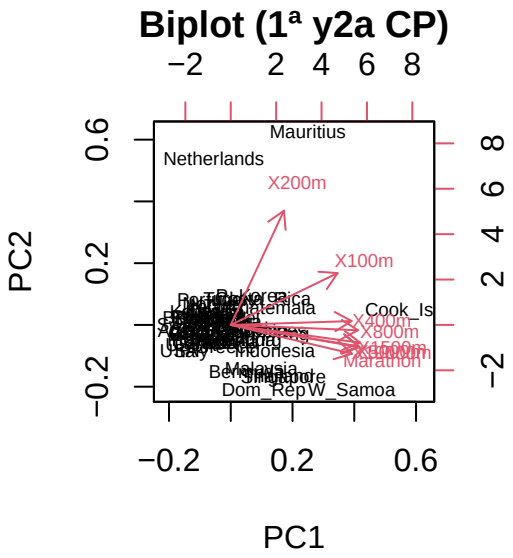


Convendría quedarse con entre 2 y 3 componentes. Si nos quedaramos con dos ya tendríamos el 87% de la varianza explicada y se facilitaría la interpretación. Si nos quedaramos con tres ya tendríamos el 95% de la varianza explicada, aunque la interpretación podría ser menos sencilla.

Además, conviene más interpretar a los que se encuentran lejos del eje horizontal, ya que esa primera componente por sí sola se lleva el 75% de la varianza.

7. Haz el biplot (1ª y2a CP) comenta (identifica grupos de países, valores discrepantes y comportamiento de las variables originales).

```
biplot(pca,main = "Biplot (1ª y2a CP)",
       cex = 0.6)
```



Dadas las cargas, la segunda componente tiene mayor carga en la variable 200m, por lo que si encontramos en el biplot variables por encima de la 2da componente, estas serían valores altos

respecto a la prueba de 200 m

Como la primera componente tiene mayor carga en las variables de pruebas de 100 y 400 y más metros entonces las variables que se encuentren más a la derecha de la primera componente se asociarán a los países que se desempeñan mejor en las pruebas de 100, 400 o más metros.

Convendría examinar a Cook_ls, W_samoa que tienen los puntajes más altos de las pruebas que implican más de. 400 m. Netherlands y Mauritius tendría los puntuajes más altos en pruebas que implican menos de 400m

Y que probablemente República dominicana sea la que menos destace de los países. Revisando los valores originales:

```
               X100m     X200m     X400m X800m X1500m X5000m X10000m Marathon
Cook_Is     0.2030000 0.3866667 0.8823333  2.02   4.24  16.70   35.38   164.70
W_Samoa     0.1803333 0.3643333 0.8166667  2.02   4.24  16.28   34.71   161.83
Netherlands 0.1753333 0.4991667 0.7516667  1.74   3.62  13.36   27.61   129.02
Mauritius   0.1865000 0.5575000 0.7950000  1.88   3.83  15.06   31.77   152.23
Dom_Rep     0.1690000 0.3441667 0.7800000  1.82   3.82  14.91   31.45   154.12
```

8. Calcula la correlación de la 1ª componente con cada una de las variables originales. Comenta.

```
    X100m     X200m     X400m     X800m    X1500m    X5000m   X10000m  Marathon
0.3229085 0.1607366 0.3654821 0.3846850 0.3907921 0.3869970 0.3892441 0.3665215
```

La primera componente correlaciona de manera similar con las variables de pruebas de 100, 400, 800, 1,500, 5,000, 10,000 m y el marathon. Solo correlaciona de manera más baja con la variable de la prueba de 200m

9. Compara las cargas (loadings) de la 1ª y la 2ª CP. Haz los barplot correspondientes, compáralos y comenta.

```
cargas_PC1
```

```
    X100m     X200m     X400m     X800m    X1500m    X5000m   X10000m  Marathon
0.3229085 0.1607366 0.3654821 0.3846850 0.3907921 0.3869970 0.3892441 0.3665215
```

```
cargas_PC2
```

```
      X100m       X200m       X400m       X800m      X1500m      X5000m
 0.38741198  0.85443890  0.02894096 -0.03985022 -0.13305276 -0.16033018
    X10000m    Marathon
-0.16627762 -0.21532173
```

La primera componente tiene mayor carga en las variables de pruebas de 100m y 400m y más metros la segunda componente tiene mayor carga en las pruebas de 200 m. Por lo que, al examinar el biplot, los valores que se muevan más a la derecha podrían interpretarse como los que se desempeñan de mejor manera en la mayoría de las pruebas. Mientras que si hay valores que se mueven más hacia arriba, se interpretaría como aquellos que se desempeñan mejor en las pruebas de pocos metros pero que se desempeñan peor en las otras pruebas (debido a las cargas negativas).

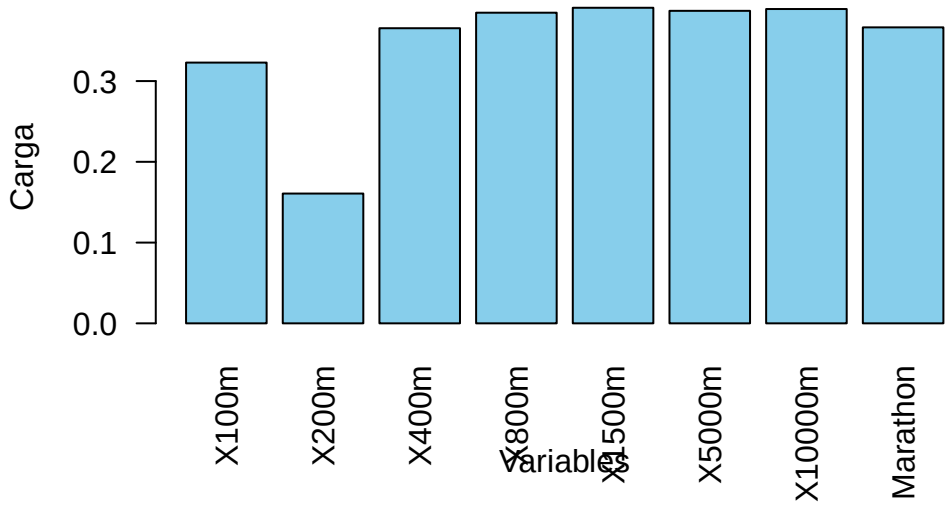


Figura 1: Cargas de la 1ª Componente Principal (CP1)

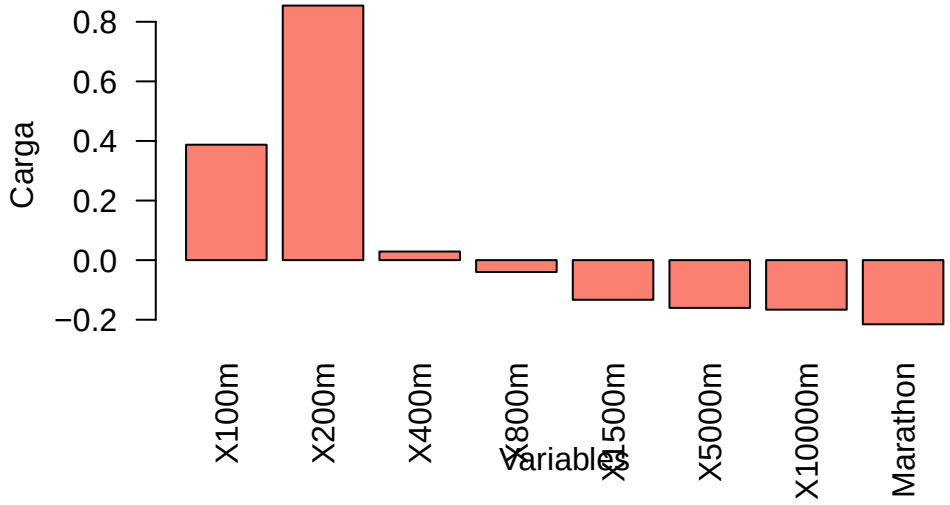


Figura 2: Cargas de la 2ª Componente Principal (CP2)

10. Verifica que CP1 es ortogonal a la CP2.

Para comprobar que la primera componente es ortogonal a la segunda, basta con hacer una correlación entre los scores de cada componente Como la correlación es muy cercana a 0, podemos decir que las dos componente son ortogonales

```
cor(pca$x[,1], pca$x[,2])
```

```
[1] 1.370769e-16
```

2 Parte 3

1. Con los datos **Distancias20ciudades.xlsx** haz un escalamiento métrico. Presenta las coordenadas en dos dimensiones, su gráfica y la medida de bondad de ajuste.
2. Con los datos haz un escalamiento no-métrico en dos dimensiones. Presenta las coordenadas en dos dimensiones, su gráfica y la medida de bondad de ajuste STRESS. También haz las gráficas dij vs ˆ dij (como las vistas en clase función Sheppard de R) y coméntala .
3. Compara lo obtenido contra un mapa y comenta.

2.1 K = 2

Se leen los datos

```
              Aca Aguascalientes Campeche Cancún CdCuauhtemoc CdJuarez
Aca             0            898     1500   1990         1210     2230
Aguascalientes  898            0     1668   2160         1740     1360
Campeche       1500         1668        0    490          726     2991
Cancun         1990         2160      490      0         1215      393
CdCuauhtemoc   1210         1740      726   1215            0     3069
CdJuarez       2230         1360     2991    393         3069        0
               CdObregon CdVictoria Colima Cuernavaca Culiacán Chetumal
Aca                 2002       1092    688        300     1650     1700
Aguascalientes      1380        515    450        597      969     2869
Campeche            3053       1603   1850       1210     2417      430
Cancun              3039       2090   2346       1702     2915      383
CdCuauhtemoc        2769       1705   1896       1251     2497      632
CdJuarez            1029       1443   1754       1931     1468     3197
               Chilpancingo Chihuahua Durango Ensenada Guadalajara Guanajuato
Aca                     115      1860    1310     3285         936        750
Aguascalientes          485       978     430     2608         252        168
Campeche               1391      2630    2070     4060        1697       1522
Cancun                 1890      3121    2567     4549        2200       2013
CdCuauhtemoc           1325      2965    2153     2129        1777       1591
CdJuarez               2119       371    1043     1365        1557       1536
               Hermosillo LaPaz Leon
Aca                  2340  4695  779
Aguascalientes       1655  4016  127
Campeche             3105  5455 1550
Cancun               3605  5955 2047
CdCuauhtemoc         3185  5535 1623
CdJuarez              772  2772 1477
```

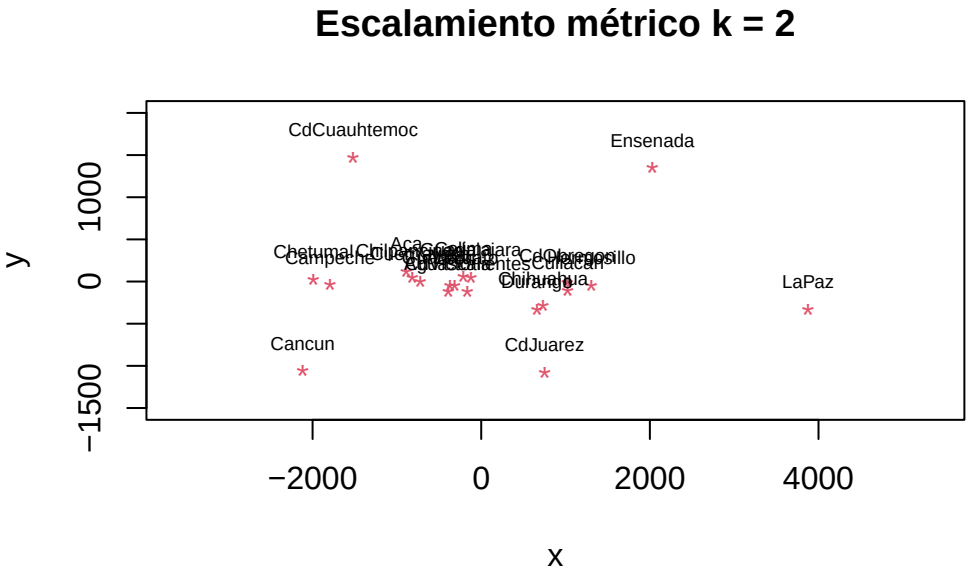
2.1.1 Escalamiento métrico

Se aplica el escalamiento multidimensional de tipo métrico para k = 2 dimensiones y se presentan algunas de las coordenadas

```
                      [,1]         [,2]
Aca               -887.4295   140.51361
Aguascalientes    -162.6740   -89.86821
Campeche         -1791.7962   -17.74549
Cancun           -2117.2191 -1037.64834
CdCuauhtemoc     -1518.7966  1494.33998
CdJuarez           751.4815 -1056.03523
```

2.1.1.1 Gráfica

Se presenta la gráfica de los puntos en k = 2 dimensiones y la etiqueta de las ciudades:



2.1.1.2 Medida GOF

Como se puede apreciar, el ajuste es bastante malo cuando no se toma el cuenta el valor absoluto de los eigenvalores (0.58) y aumenta apenas de manera aceptable cuando se toman en cuenta los lambda's en su valor absoluto (0.76). Esto indica que el escalamiento en k = 2 dimensiones reproduce el 58% de las distancias originales.

```
[1] 0.5899928 0.7652376
```

La aparición de eigenvalores negativos nos indica que no se pudo obtener una representación perfecta de los datos y que, además, las distancias no son euclidianas.

```
 [1]  4.090087e+07  6.632100e+06  5.677672e+06  4.065645e+06  2.150335e+06
 [6]  1.528815e+06  7.352914e+05  2.877820e+05  9.612529e+04  4.067565e+04
[11]  6.247092e-10 -8.157445e+02 -4.343101e+04 -8.095482e+04 -1.892648e+05
[16] -1.972728e+05 -5.764401e+05 -1.531651e+06 -3.378991e+06 -5.791722e+06
[21] -6.659488e+06
```

2.1.2 Escalamiento no métrico

Se aplica el escalamiento no métrico para mismas dimensiones y se presentan las primeras coordenadas:

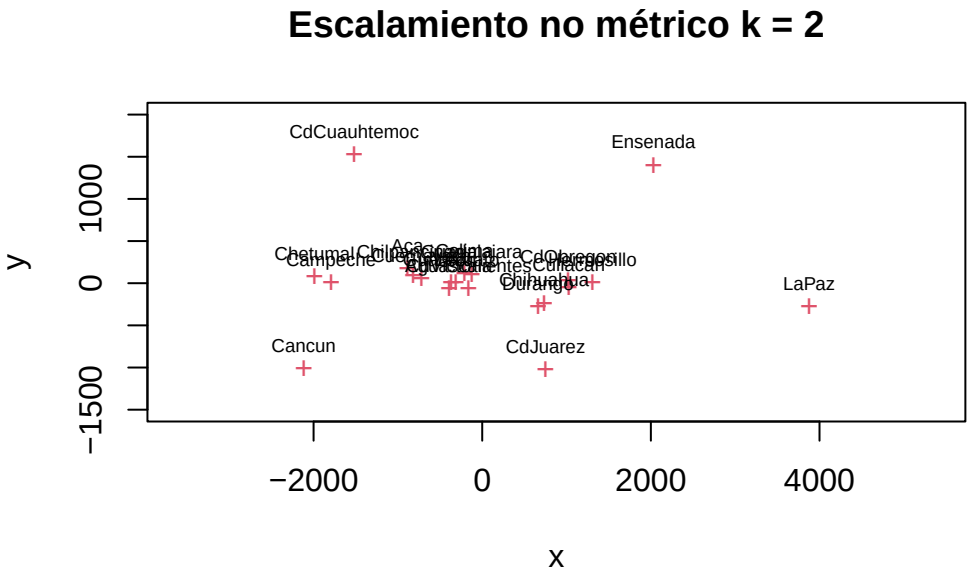
```
initial  value 18.365363
final  value 18.365330
converged
```

```
                    [,1]        [,2]
Aca             -887.4299   140.51422
Aguascalientes  -162.6733   -89.86838
Campeche       -1791.7978   -17.74544
Cancun         -2117.2186 -1037.64783
```

```
CdCuauhtemoc   -1518.7958  1494.33788
CdJuarez         751.4817 -1056.03537
```

2.1.2.1 Gráfica

Se presenta la gráfica del escalamiento no métrico con mismas dimensiones. Genera distancias parecidas a las del escalamiento métrico.



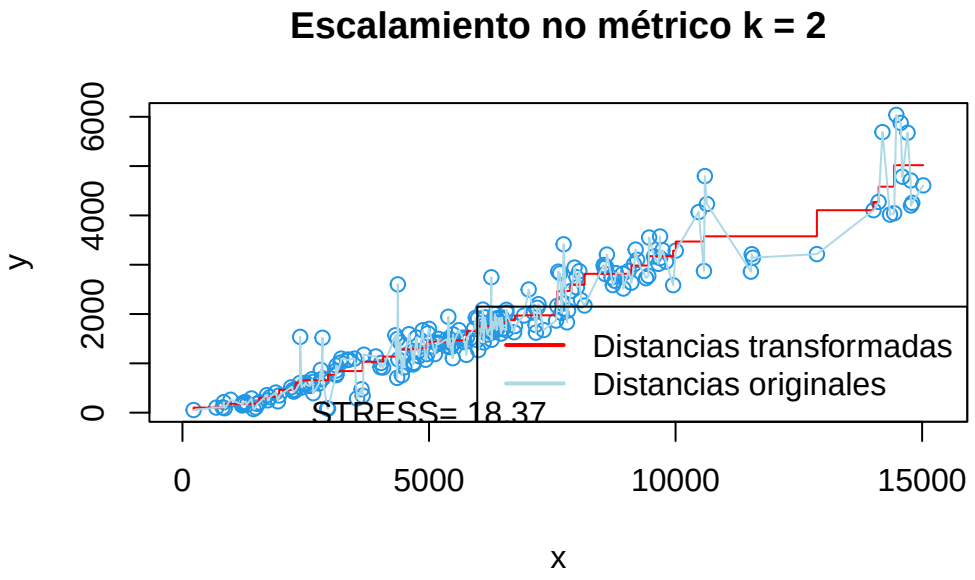
2.1.2.2 STRESS

La medida STRESS indica qué tan bien se ajusta la transformación de las distancias originales (para hacerlas euclideas) a las originales. Como el valor es mayor a 0.05, esto indica que el ajuste es malo y que las nuevas distancias euclidianas no se parecen a las originales después de la transformación.

```
[1] 18.36533
```

2.1.2.3 dij - ^dij

Se utiliza la función Shepard para generar el gráfico que permite comparar las distancias originales con las transformadas:



2.2 K = 3

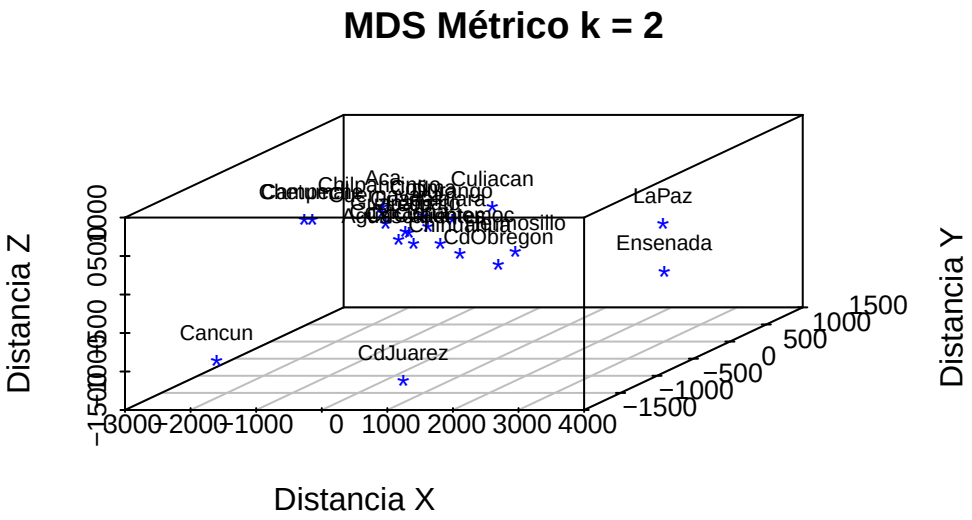
Como análisis extra, se aplica el escalamiento multidimensional de tipo métrico para k = 3 dimensiones.

Presentación de algunas de las coordenadas en 3 dimensiones

	[,1]	[,2]	[,3]
Aca	-887.4295	140.51361	442.88343
Aguascalientes	-162.6740	-89.86821	55.41163
Campeche	-1791.7962	-17.74549	334.11826
Cancun	-2117.2191	-1037.64834	-1040.83739
CdCuauhtemoc	-1518.7966	1494.33998	-636.73384
CdJuarez	751.4815	-1056.03523	-1294.14101

2.2.0.1 Gráfica

Gráfica de los puntos en k = 3 dimensiones



2.2.0.2 Medida GOF

Como se puede apreciar, el ajuste también es malo cuando no se toma el cuenta el valor absoluto de los eigenvalores (0.66) y aumenta de manera aceptable cuando se toman en cuenta los lambda's en su valor absoluto 0.85 Esto indica que el escalamiento en k = 3 reproduce el 66% de las distancias originales, de nuevo, destacando que estas no son euclideas y no parece mejorar mucho la representación de los datos.

```
[1] 0.6604656 0.8566429
```

2.2.0.3 Escalamiento no métrico

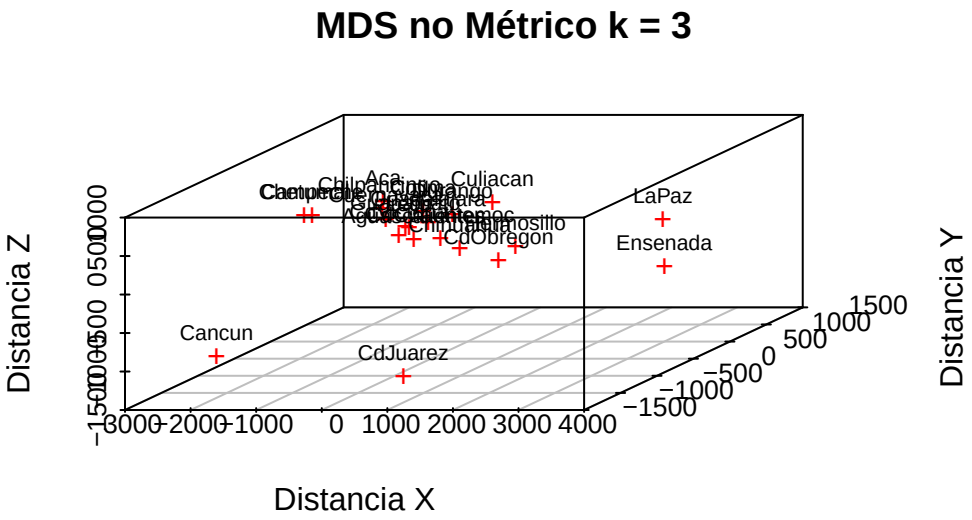
Se aplica el escalamiento no métrico para k = 3 dimensiones y se presentan sus primeras coordenadas en tres dimensiones

```
initial  value 18.719675
final  value 18.719627
converged
```

	[,1]	[,2]	[,3]
Aca	-887.4300	140.51408	442.88390
Aguascalientes	-162.6732	-89.86839	55.41128
Campeche	-1791.7985	-17.74557	334.11794
Cancun	-2117.2182	-1037.64747	-1040.83677
CdCuauhtemoc	-1518.7962	1494.33821	-636.73311
CdJuarez	751.4809	-1056.03369	-1294.13927

2.2.0.4 Gráfica

Se genera una gráfica en tres dimensiones para representar a los datos.



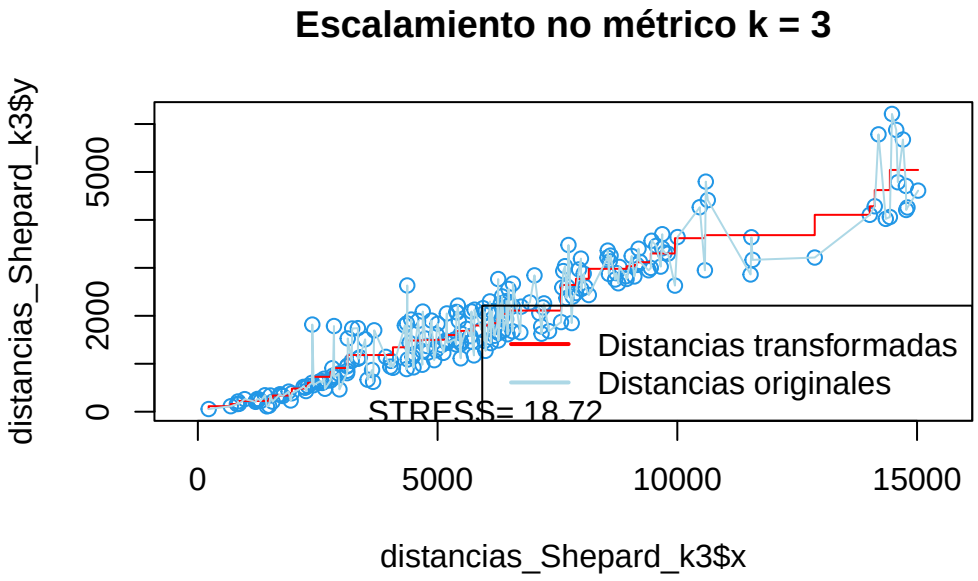
2.2.0.5 STRESS

Como el valor es mayor a 0.05 esto indica que el ajuste también es malo y que las nuevas distancias euclidianas no se parecen a las originales. Incluso aumentando el error un poco más respecto a cuando k = 2

```
[1] 18.71963
```

2.2.0.6 dij - ˆdij

A simple vista, no se nota mejoría respecto al caso donde k = 2. En conclusión, probablemente se necesiten más dimensiones para representar de manera adecuada a las distancias entre ciudades.



3 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)
```

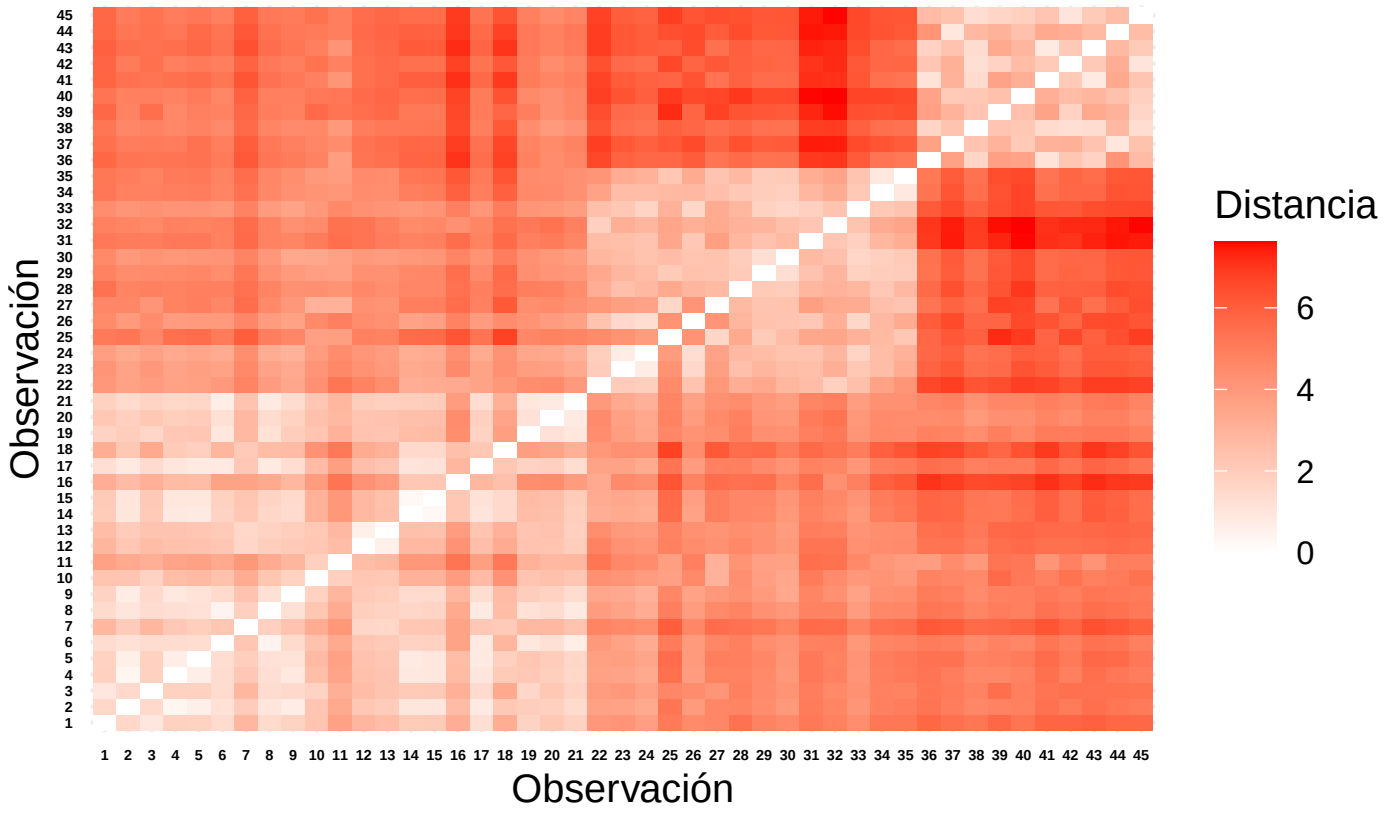


Figura 3: Mapa de Calor de la Matrix de Distancias.

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.

3.1 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() %>%
    dplyr::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`.

3.1.1 Liga Sencilla.

3.1.1.1 Dos Grupos.

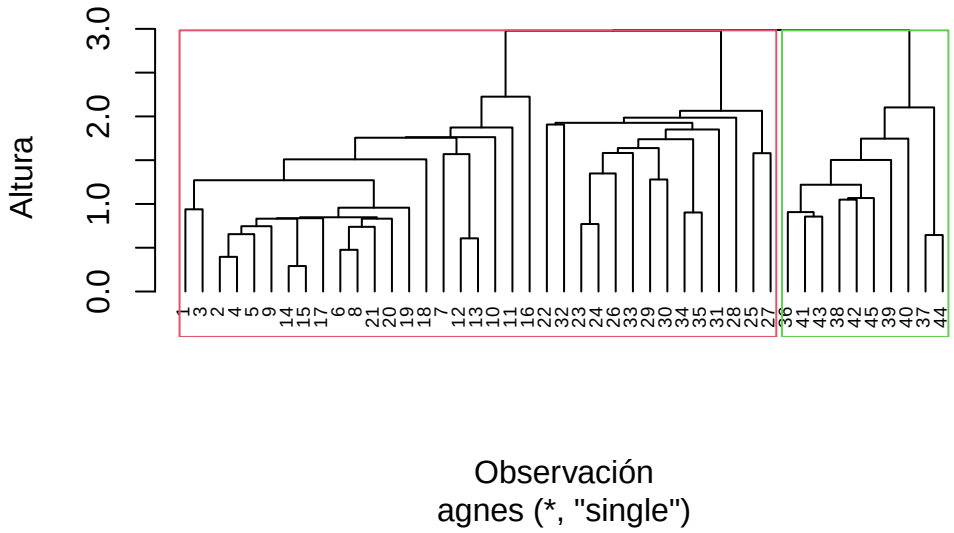


Figura 4: Dendograma: 2 grupos (liga sencilla).

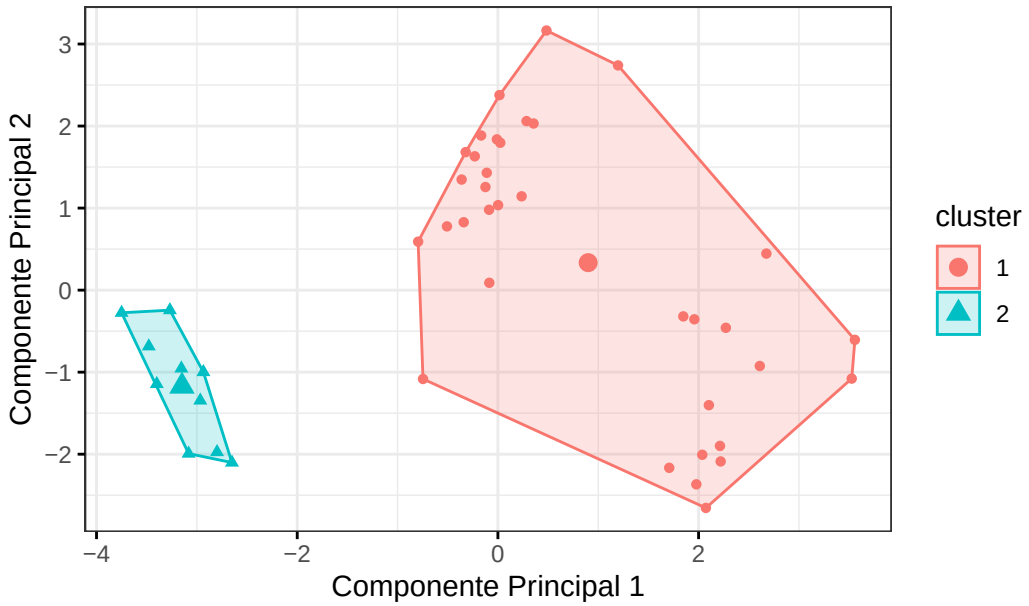


Figura 5: Conglomerados: 2 grupos (liga sencilla).

3.1.1.2 Tres Grupos.

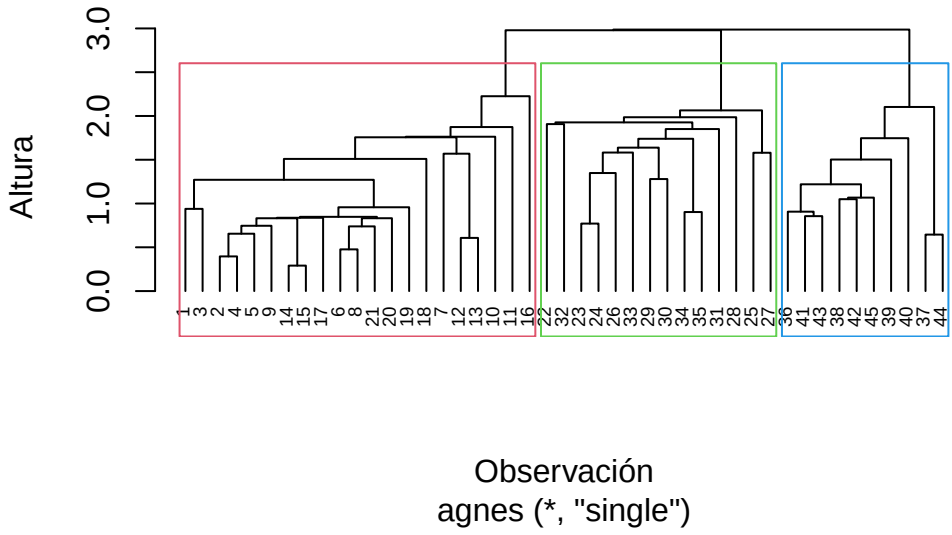


Figura 6: Dendograma: 3 grupos (liga sencilla).

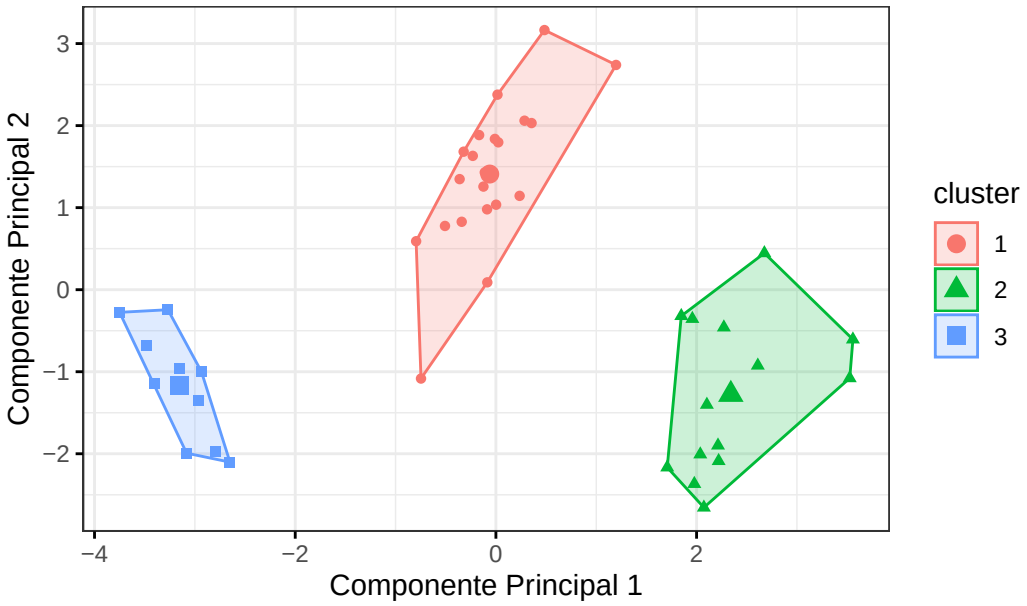


Figura 7: Conglomerados: 3 grupos (liga sencilla).

3.1.1.3 Cuatro Grupos.

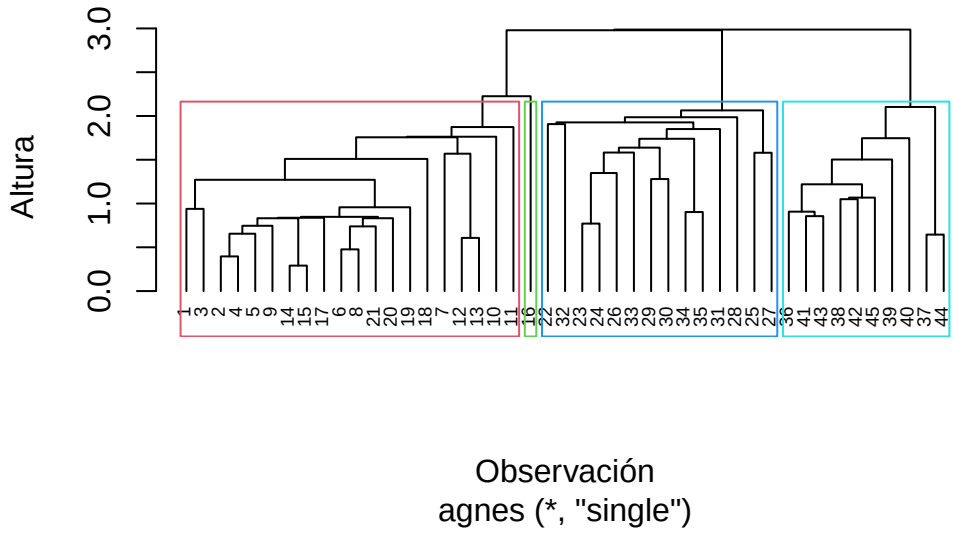


Figura 8: Dendograma: 4 grupos (liga sencilla).

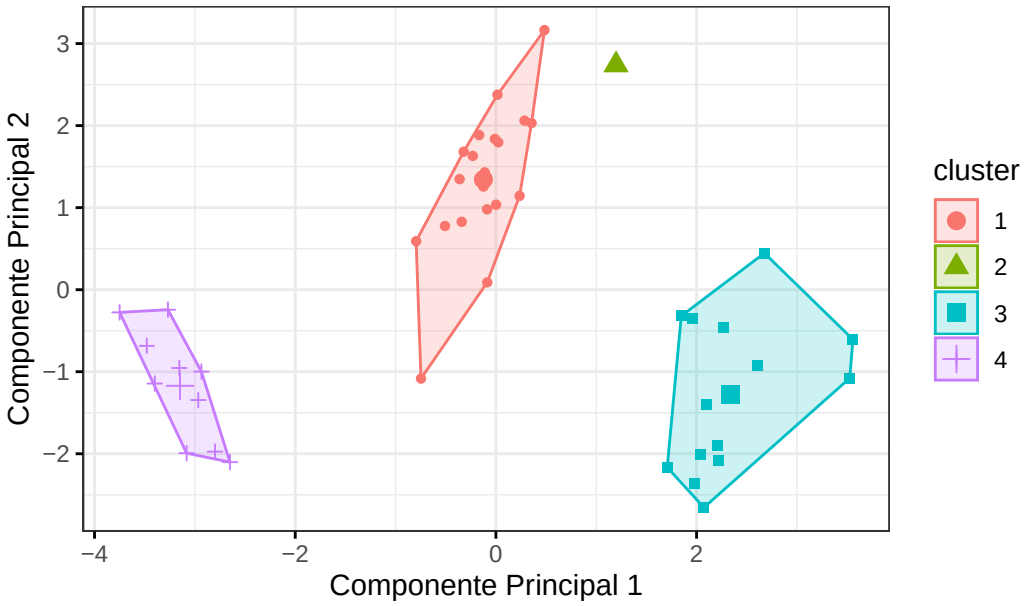


Figura 9: Conglomerados: 4 grupos (liga sencilla).

3.1.1.4 Cinco Grupos.

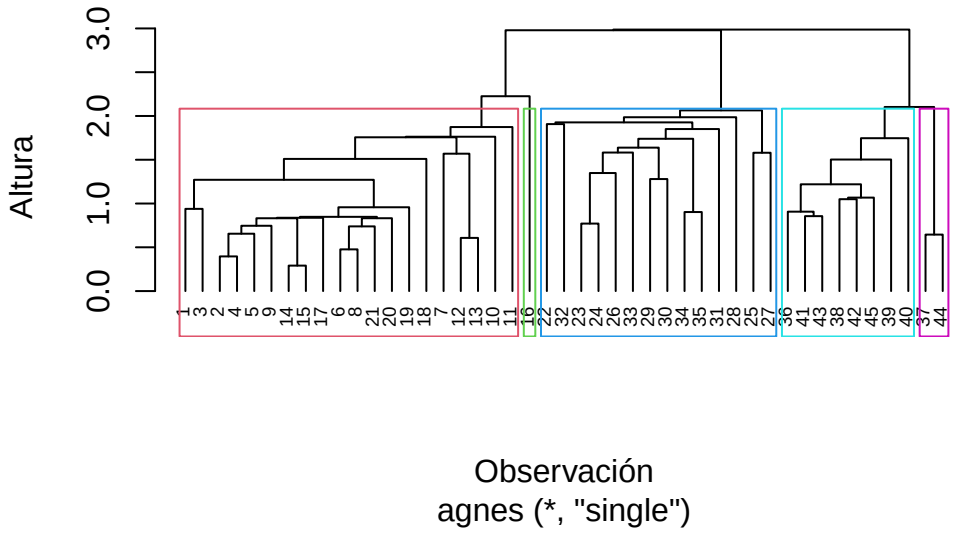


Figura 10: Dendograma: 5 grupos (liga sencilla).

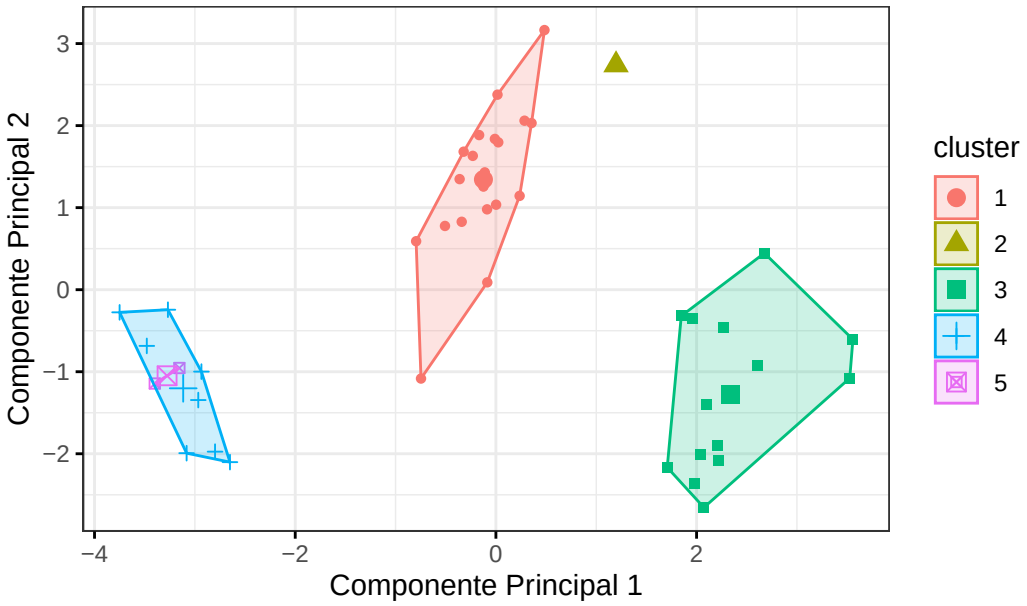


Figura 11: Conglomerados: 5 grupos (liga sencilla).

3.1.1.5 Seis Grupos.

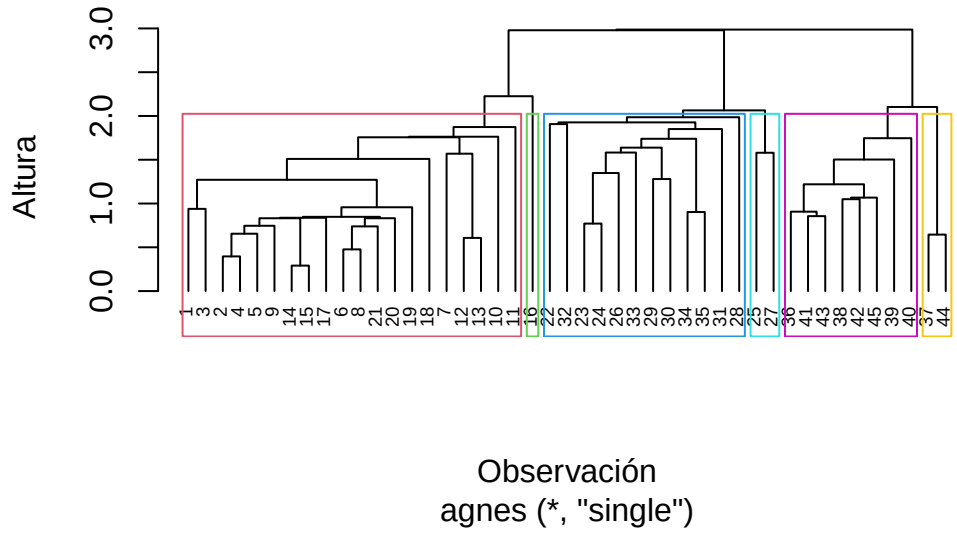


Figura 12: Dendograma: 6 grupos (liga sencilla).

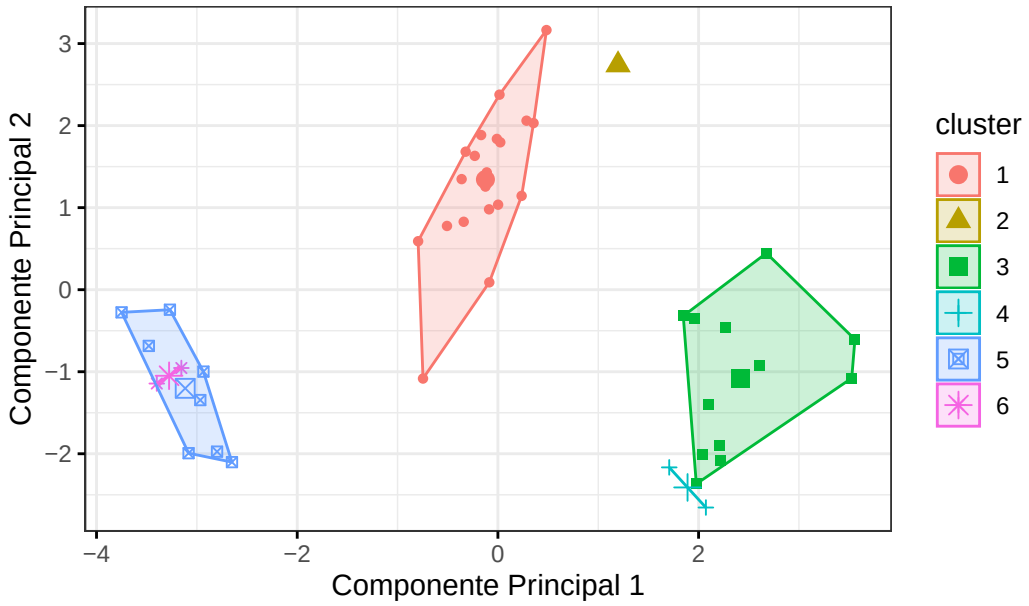


Figura 13: Conglomerados: 6 grupos (liga sencilla).

3.1.1.6 Siete Grupos.

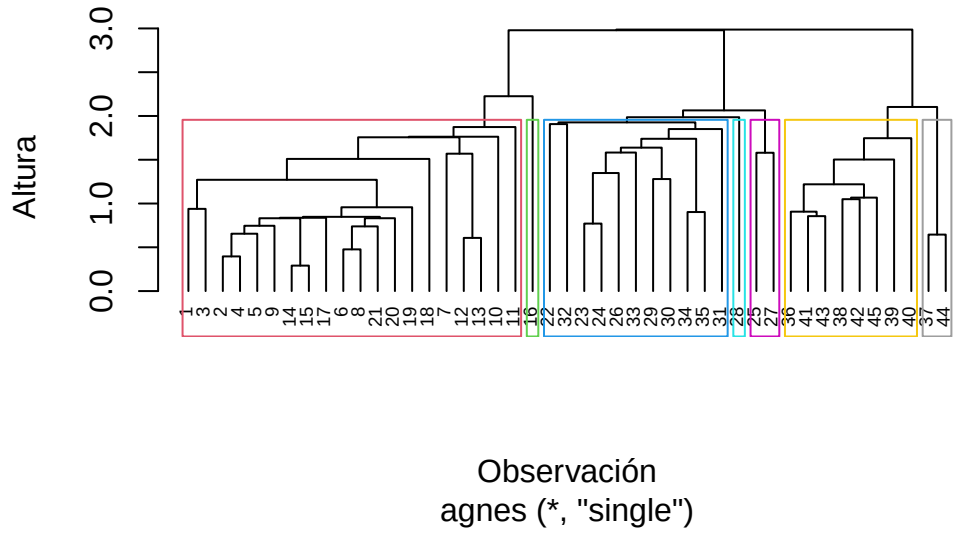


Figura 14: Dendograma: 7 grupos (liga sencilla).

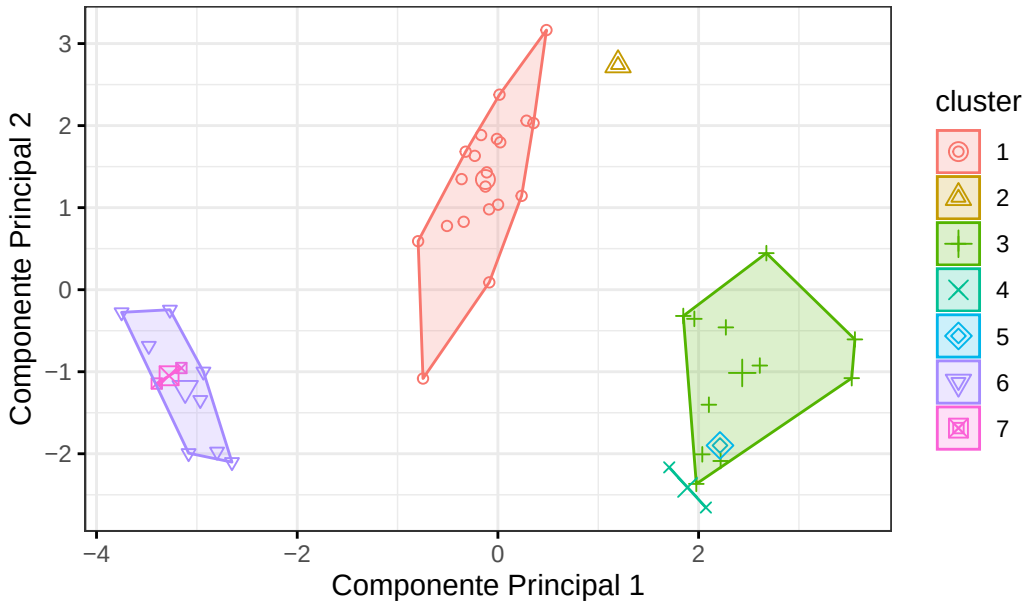


Figura 15: Conglomerados: 7 grupos (liga sencilla).

3.1.1.7 Ocho Grupos.

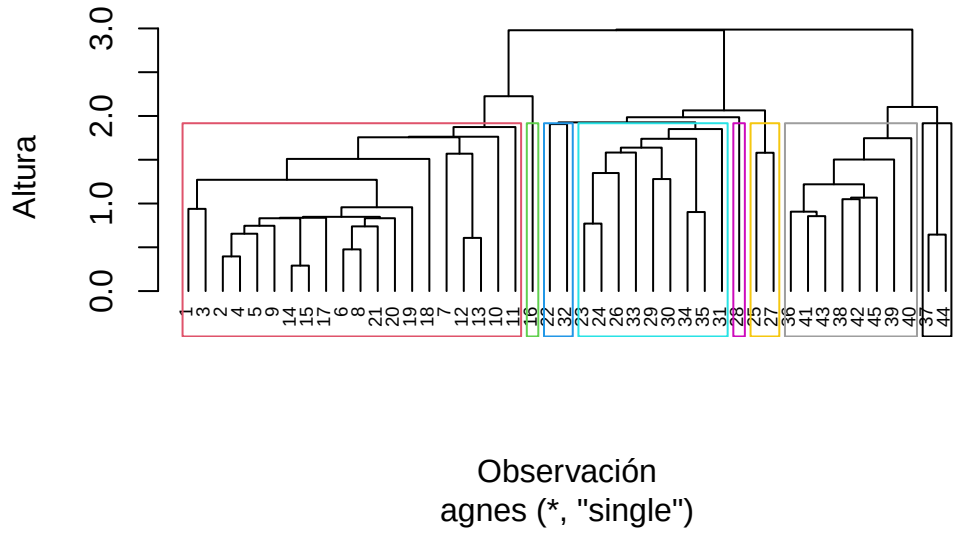


Figura 16: Dendograma: 8 grupos (liga sencilla).

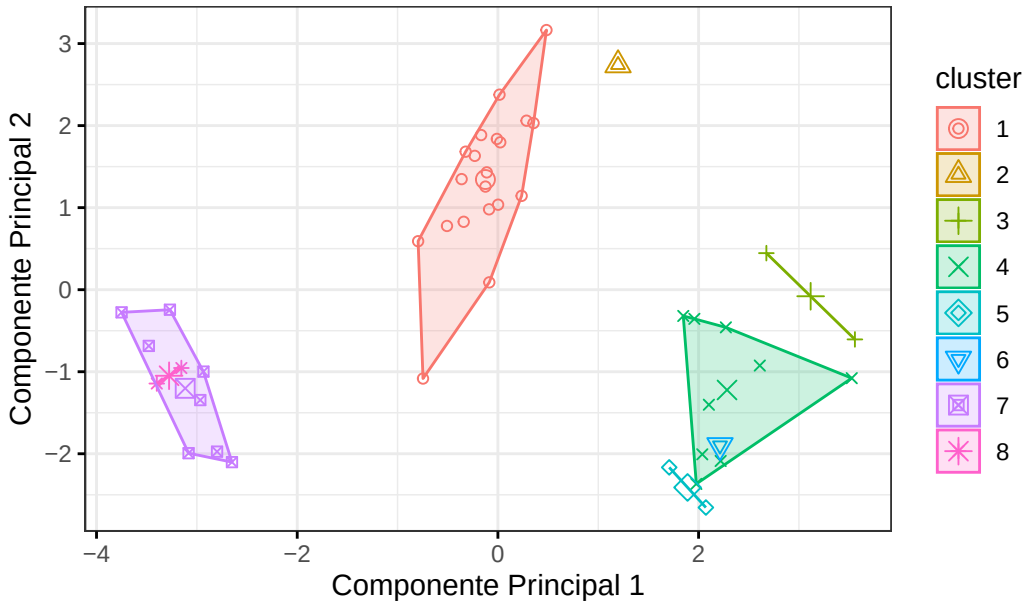


Figura 17: Conglomerados: 8 grupos (liga sencilla).

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.

3.1.2 Liga Completa.

3.1.2.1 Dos Grupos.

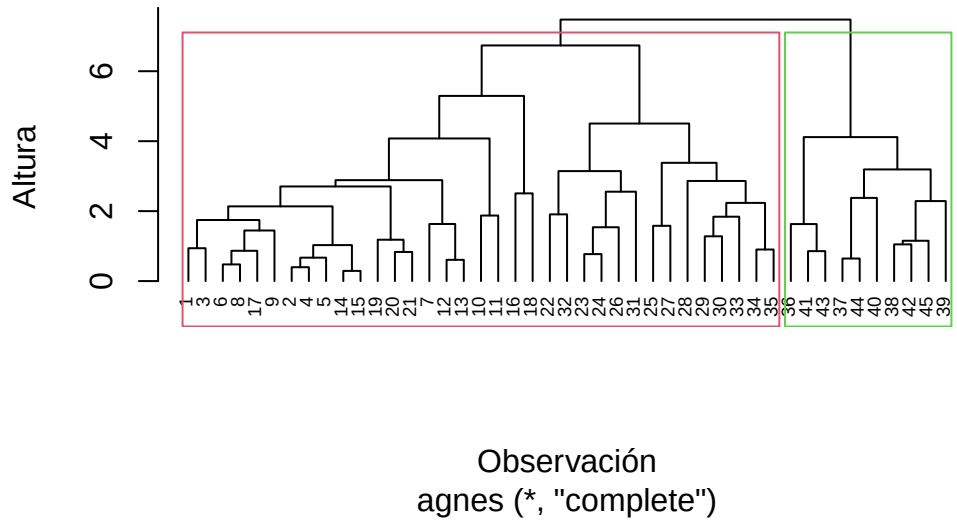


Figura 18: Dendograma: 2 grupos (liga completa).

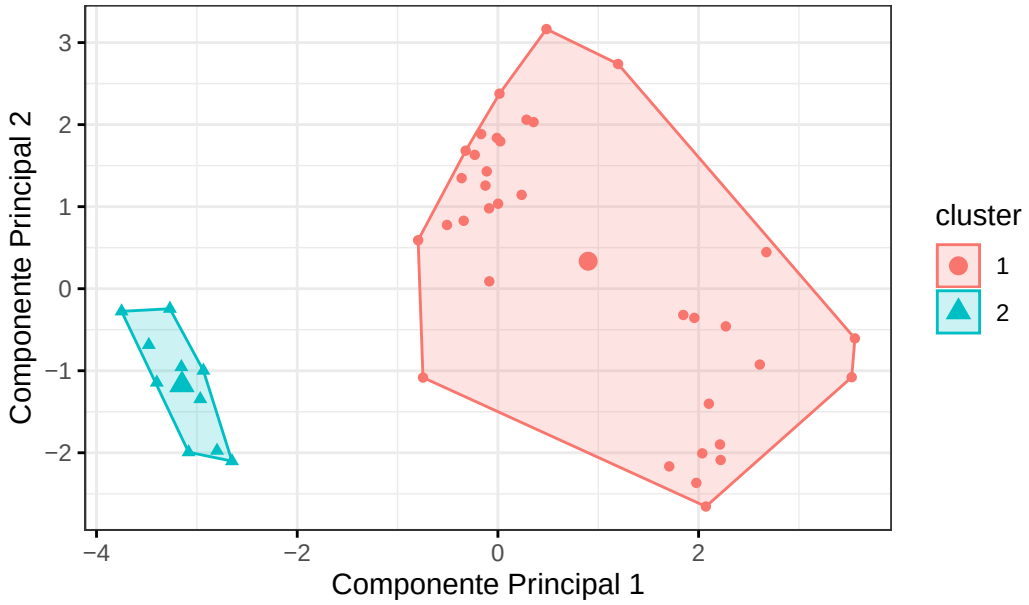


Figura 19: Conglomerados: 2 grupos (liga completa).

3.1.2.2 Tres Grupos.

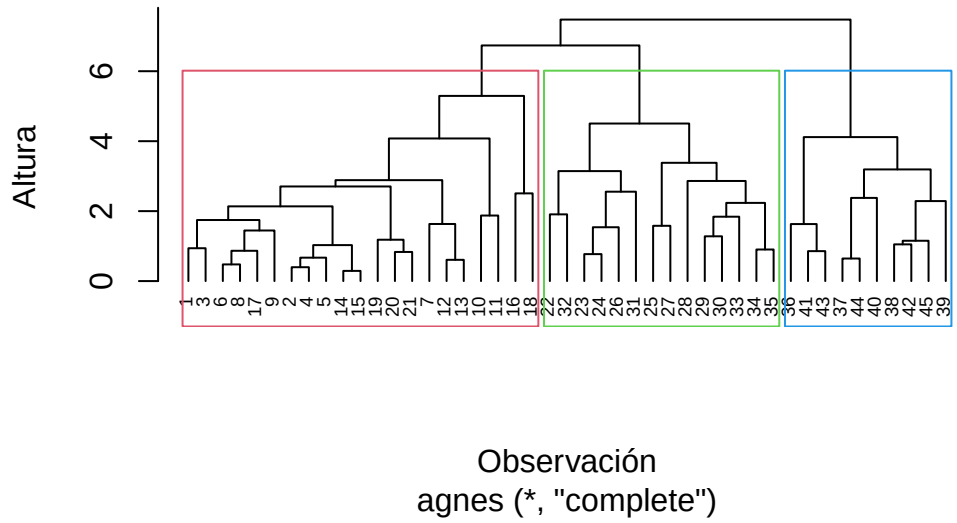


Figura 20: Dendograma: 3 grupos (liga completa).

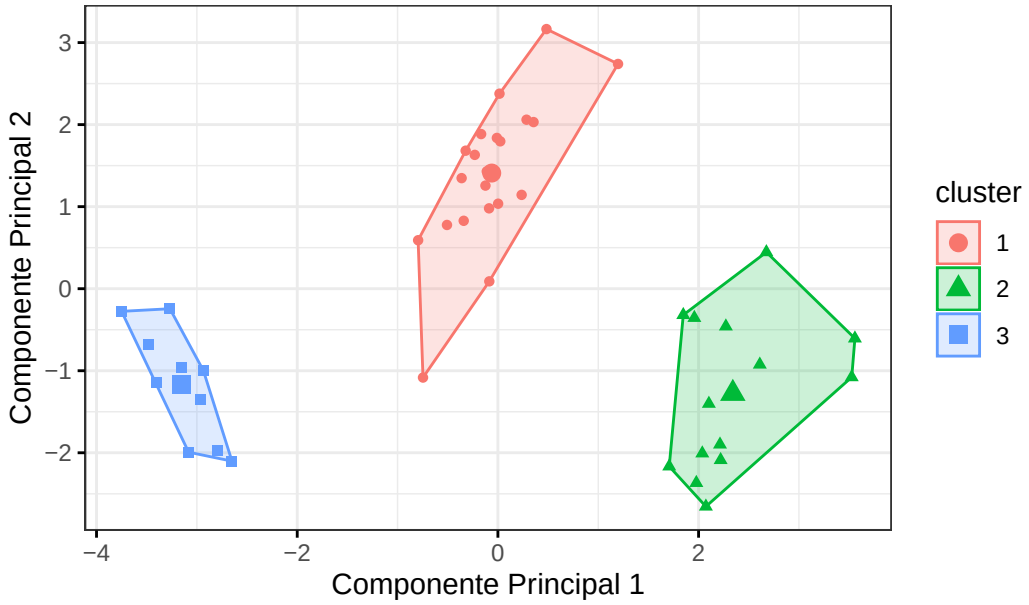


Figura 21: Conglomerados: 3 grupos (liga completa).

3.1.2.3 Cuatro Grupos.

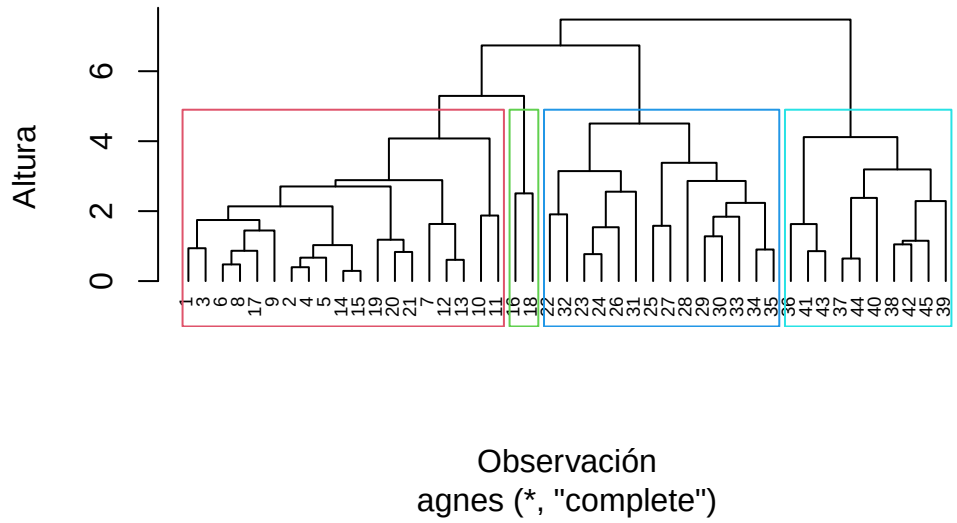


Figura 22: Dendograma: 4 grupos (liga completa).

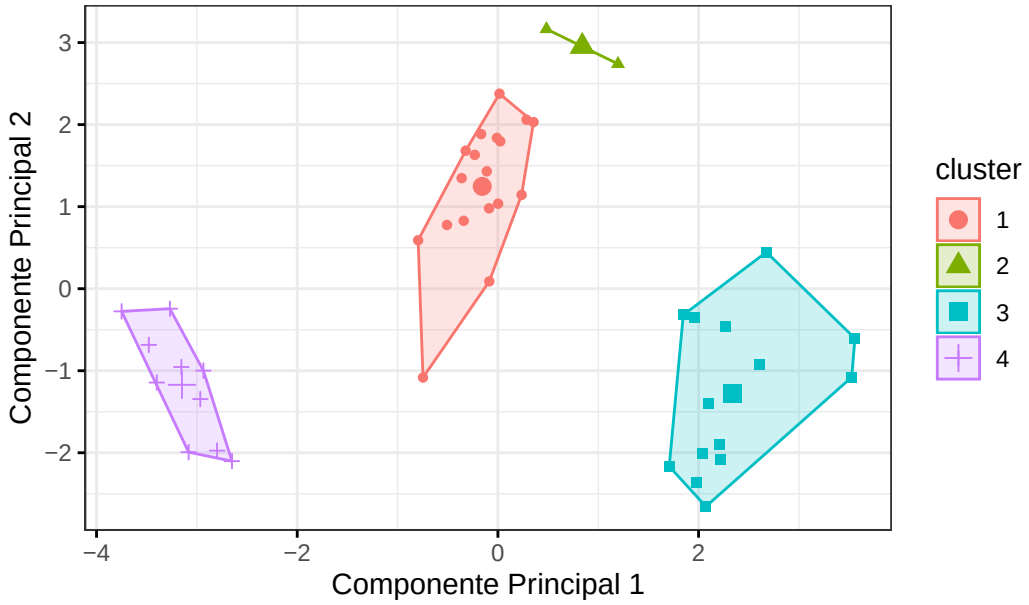


Figura 23: Conglomerados: 4 grupos (liga completa).

3.1.2.4 Cinco Grupos.

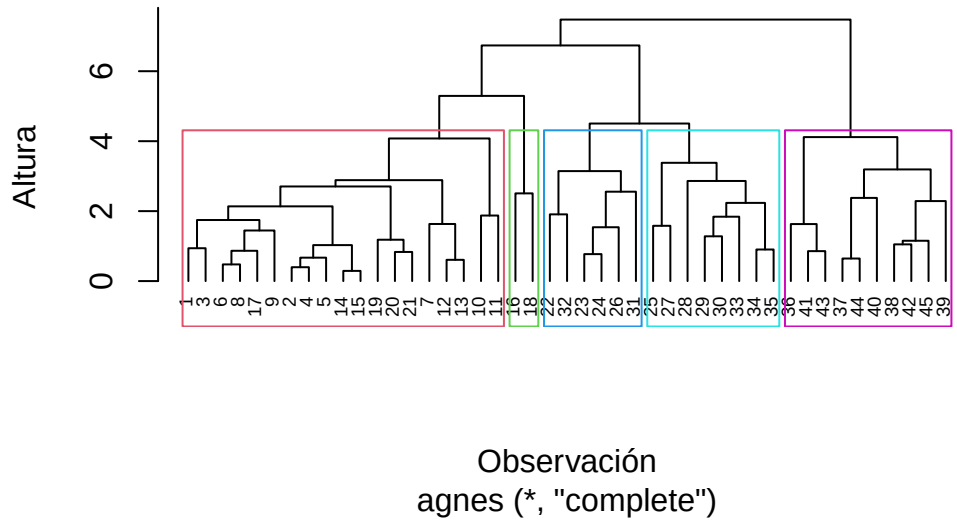


Figura 24: Dendograma: 5 grupos (liga completa).

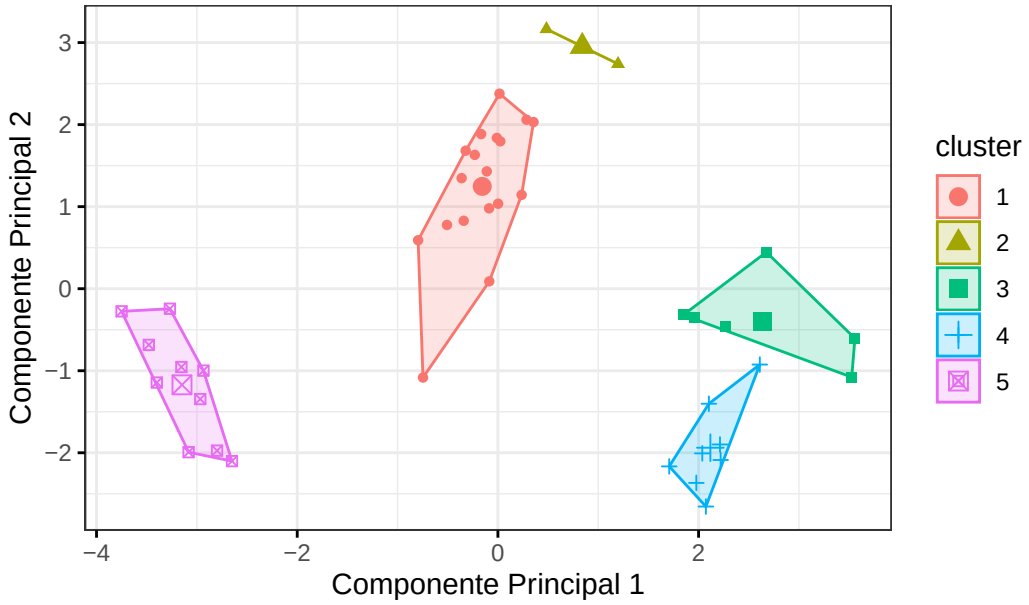


Figura 25: Conglomerados: 5 grupos (liga completa).

3.1.2.5 Seis Grupos.

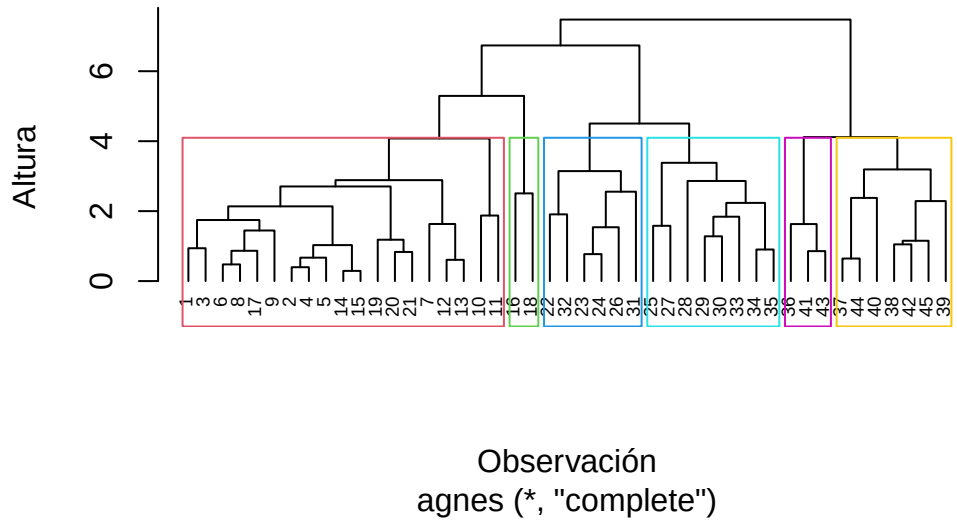


Figura 26: Dendograma: 6 grupos (liga completa).

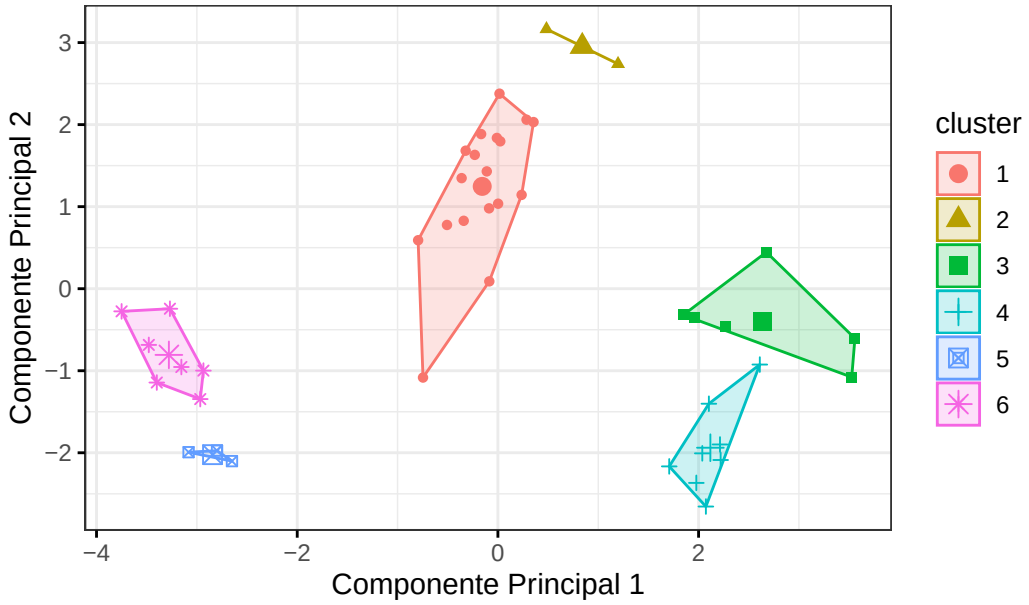


Figura 27: Conglomerados: 6 grupos (liga completa).

3.1.2.6 Siete Grupos.

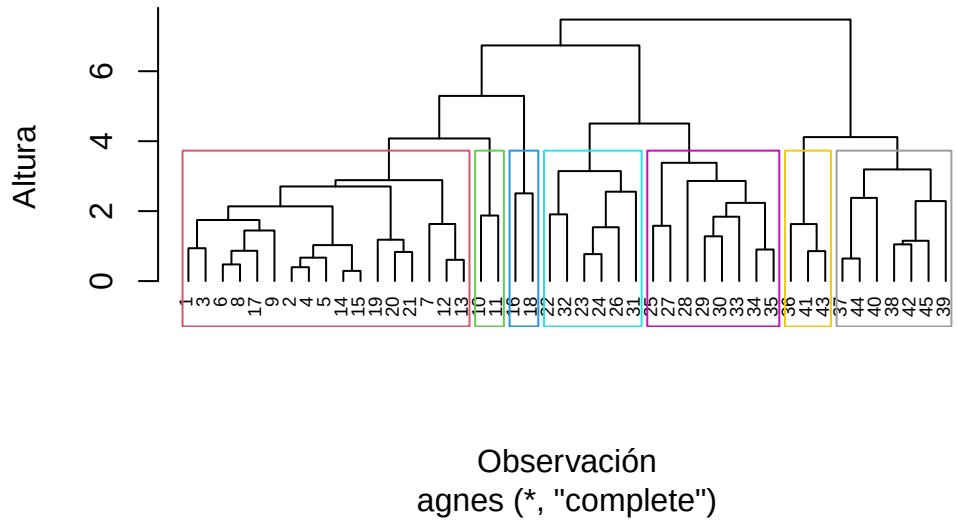


Figura 28: Dendograma: 7 grupos (liga completa).

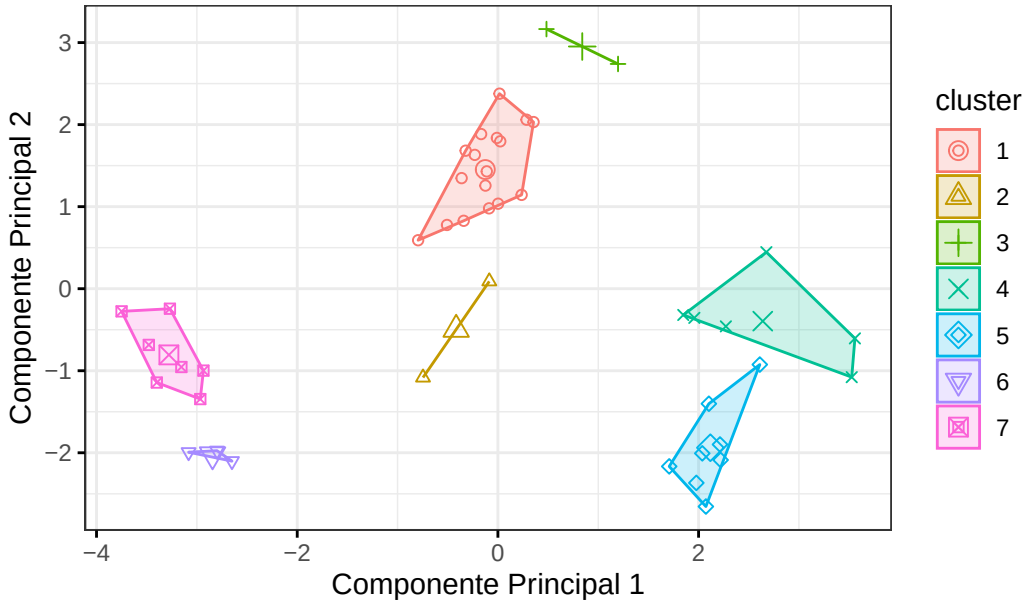


Figura 29: Conglomerados: 7 grupos (liga completa).

3.1.2.7 Ocho Grupos.

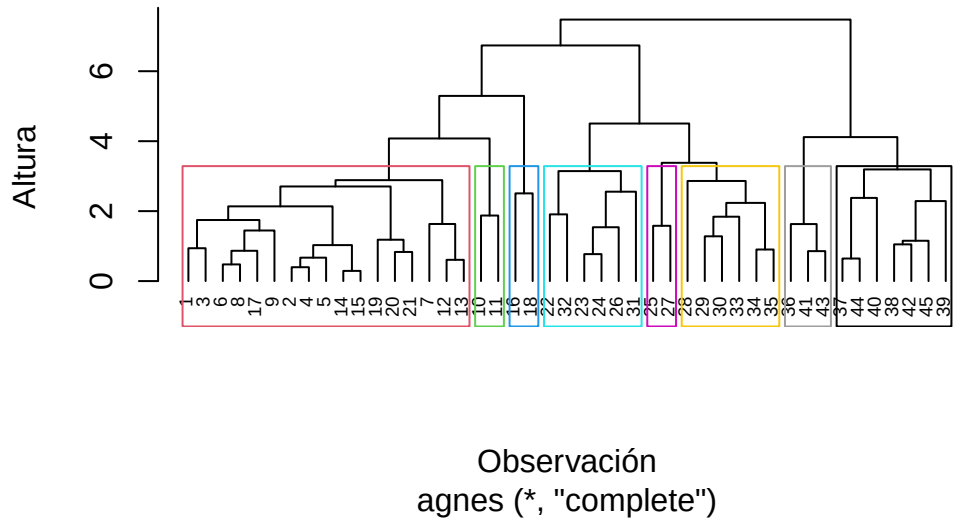


Figura 30: Dendograma: 8 grupos (liga completa).

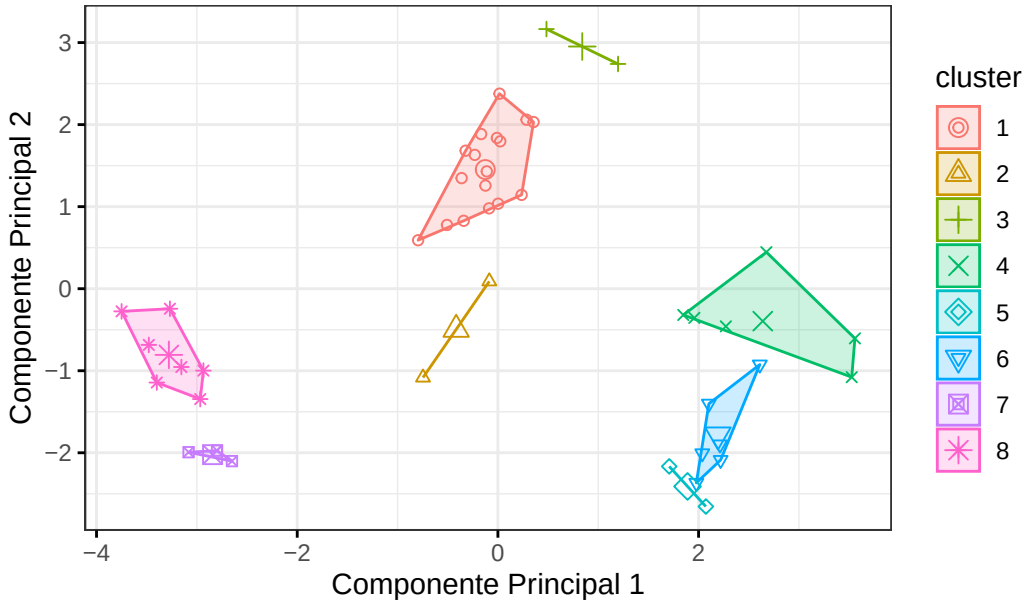


Figura 31: Conglomerados: 8 grupos (liga completa).

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.

3.1.3 Métdo de Ward.

3.1.3.1 Dos Grupos.

Figura 32: Dendograma: 2 grupos (métdo de Ward).

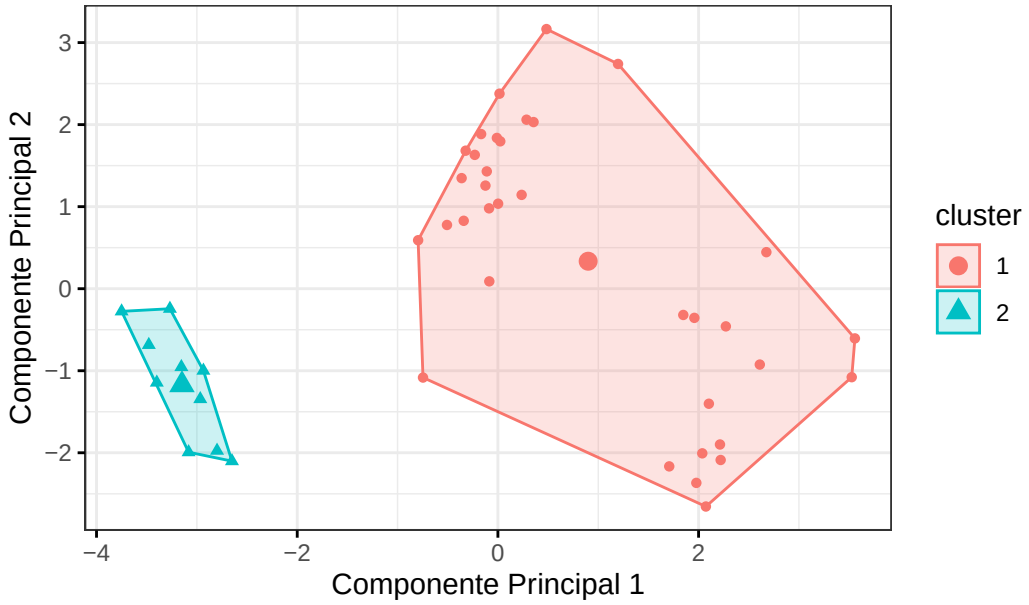


Figura 33: Conglomerados: 2 grupos (métdo de Ward).

3.1.3.2 Tres Grupos.

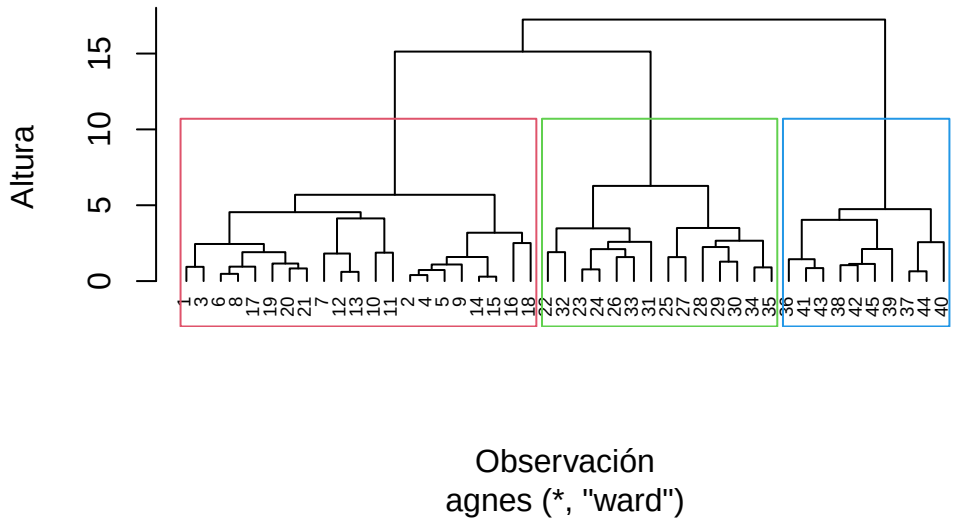


Figura 34: Dendograma: 3 grupos (métdo de Ward).

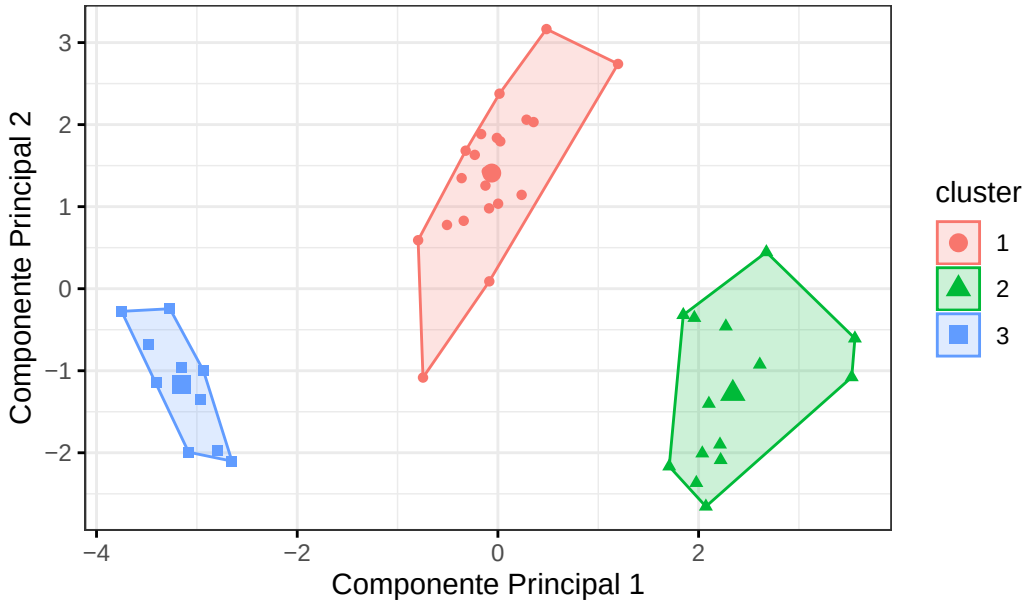


Figura 35: Conglomerados: 3 grupos (métdo de Ward).

3.1.3.3 Cuatro Grupos.

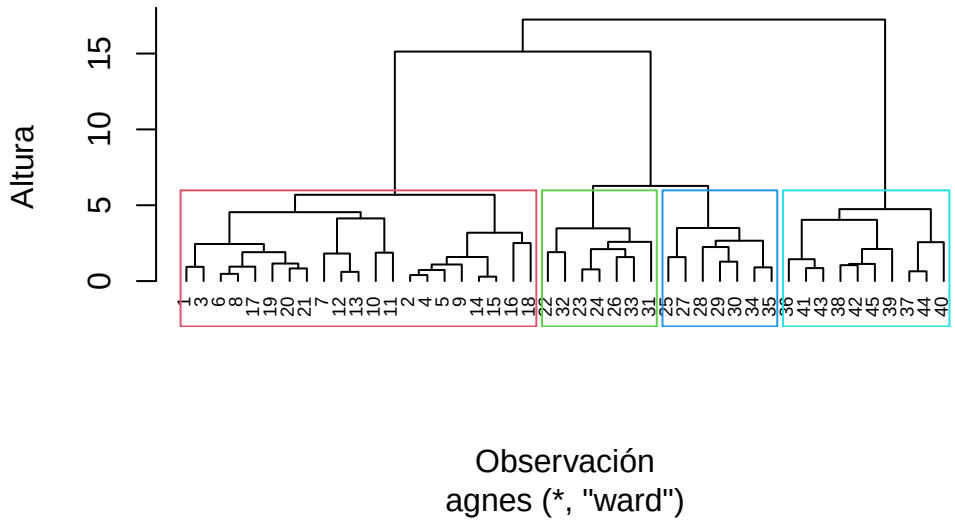


Figura 36: Dendograma: 4 grupos (métdo de Ward).

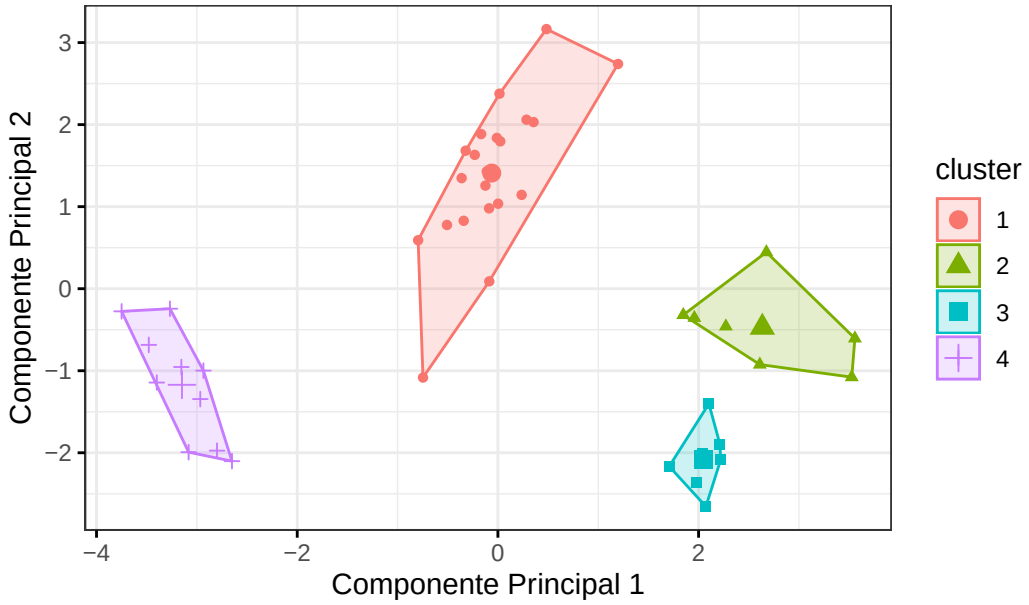


Figura 37: Conglomerados: 4 grupos (métdo de Ward).

3.1.3.4 Cinco Grupos.

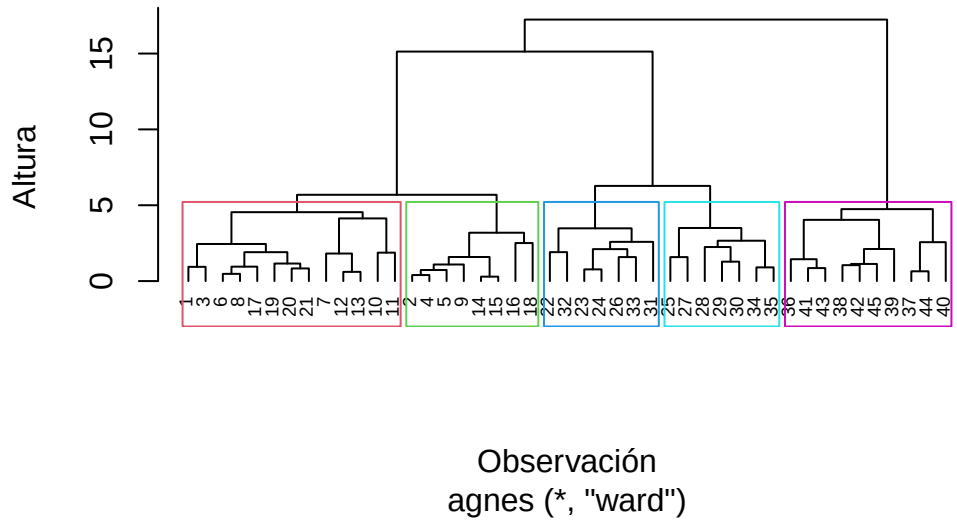


Figura 38: Dendograma: 5 grupos (métdo de Ward).

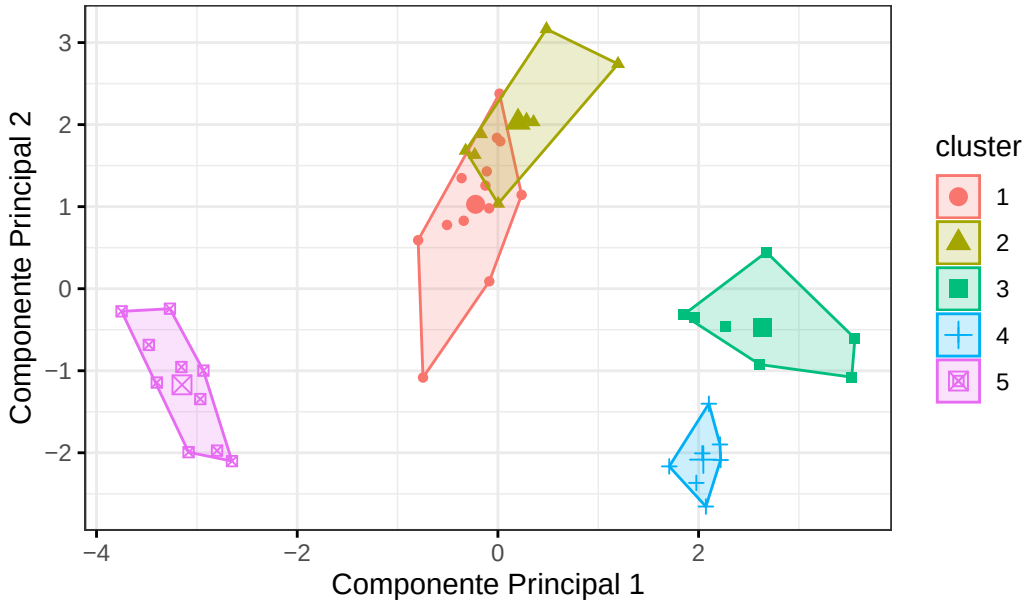


Figura 39: Conglomerados: 5 grupos (métdo de Ward).

3.1.3.5 Seis Grupos.

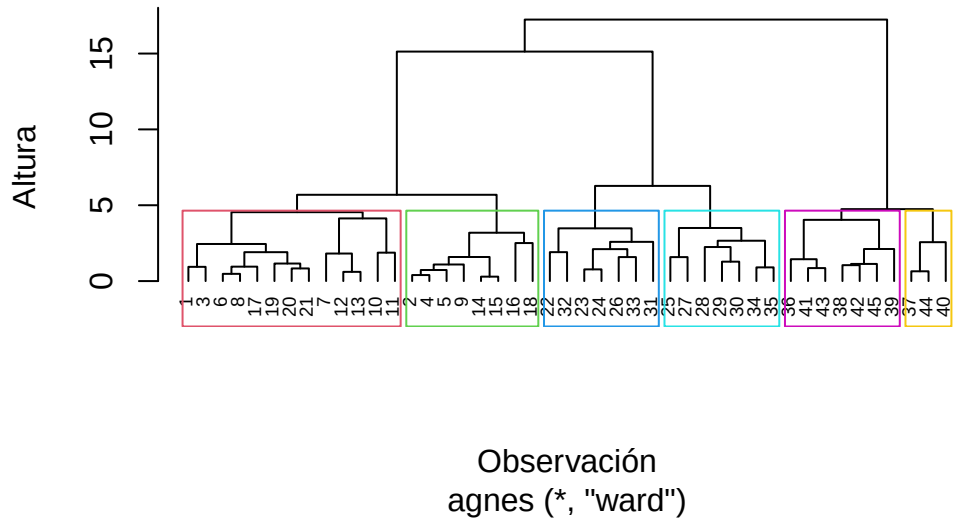


Figura 40: Dendograma: 6 grupos (métdo de Ward).

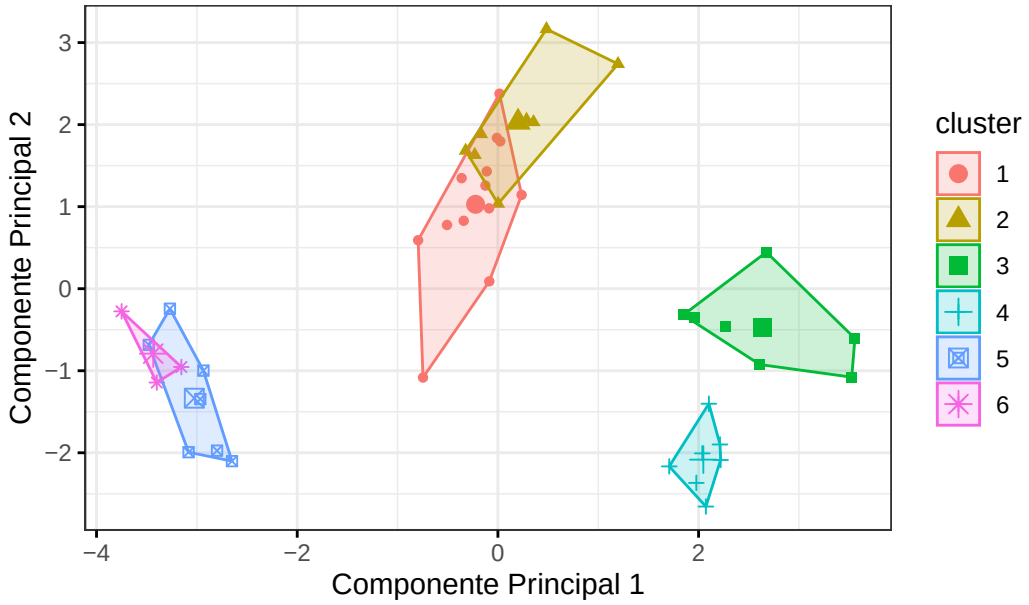


Figura 41: Conglomerados: 6 grupos (métdo de Ward).

3.1.3.6 Siete Grupos.

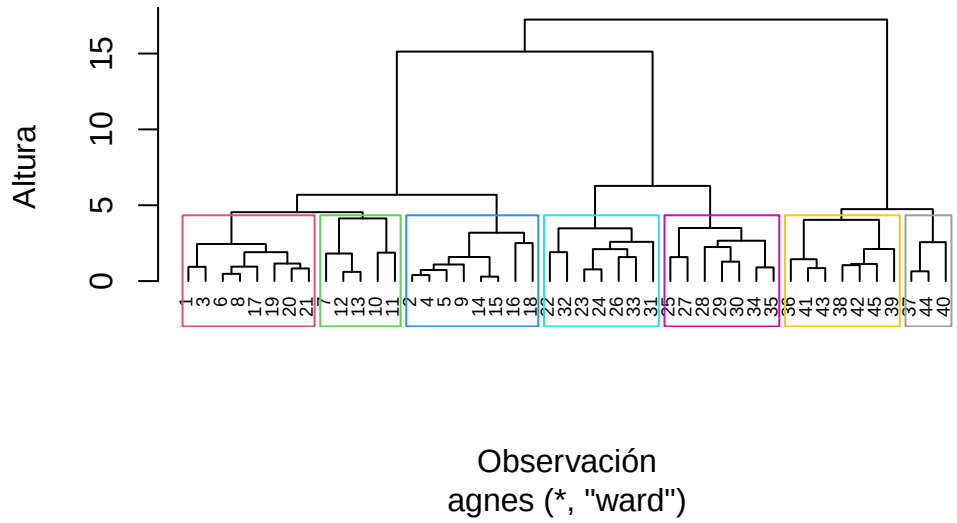


Figura 42: Dendograma: 7 grupos (métdo de Ward).

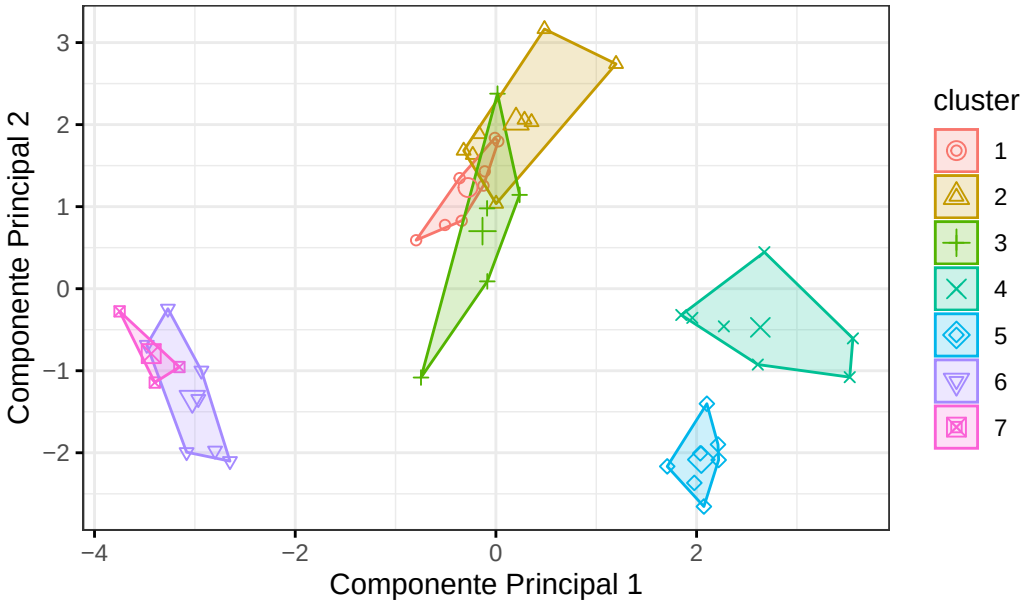


Figura 43: Conglomerados: 7 grupos (métdo de Ward).

3.1.3.7 Ocho Grupos.

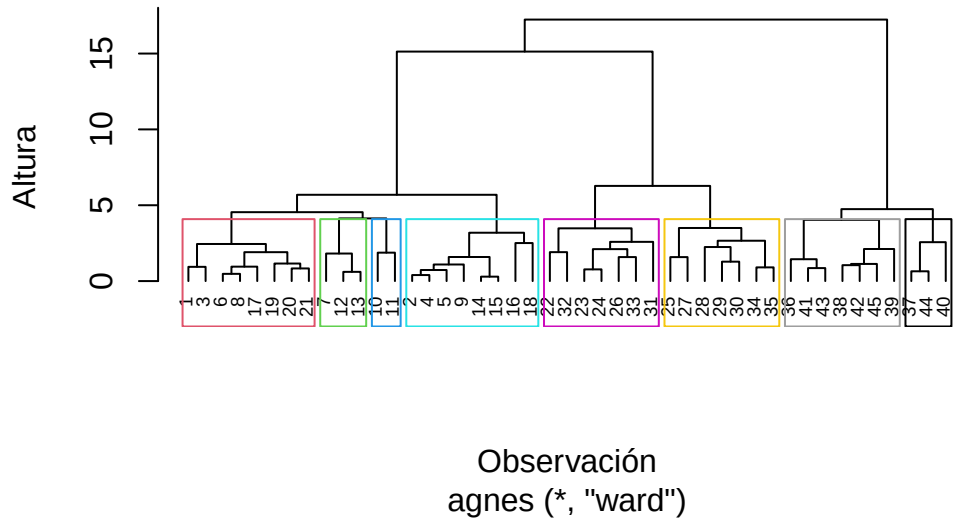


Figura 44: Dendograma: 8 grupos (métdo de Ward).

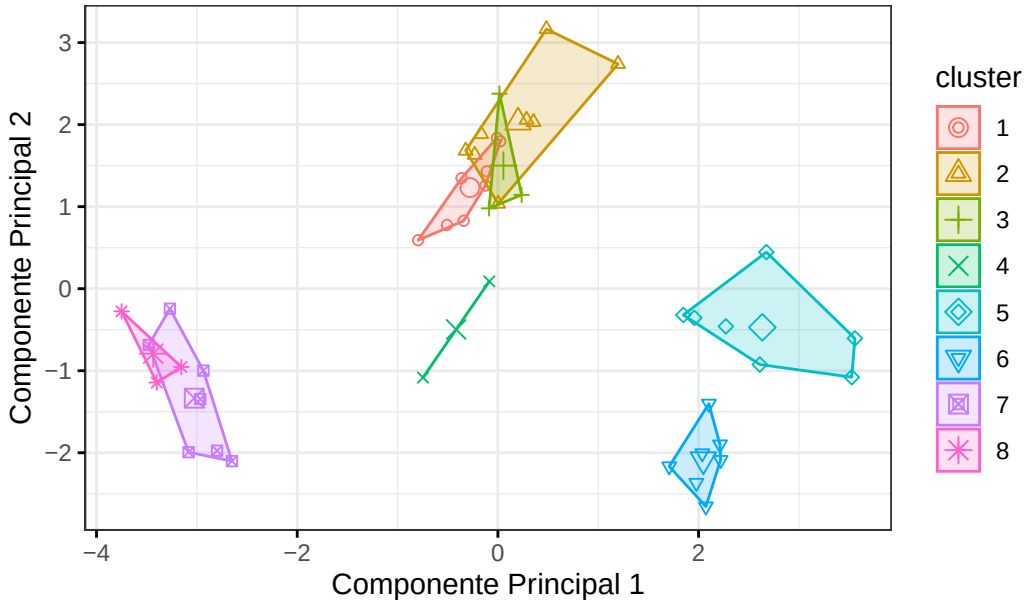


Figura 45: Conglomerados: 8 grupos (métdo de Ward).

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.

3.2 K-Means.

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

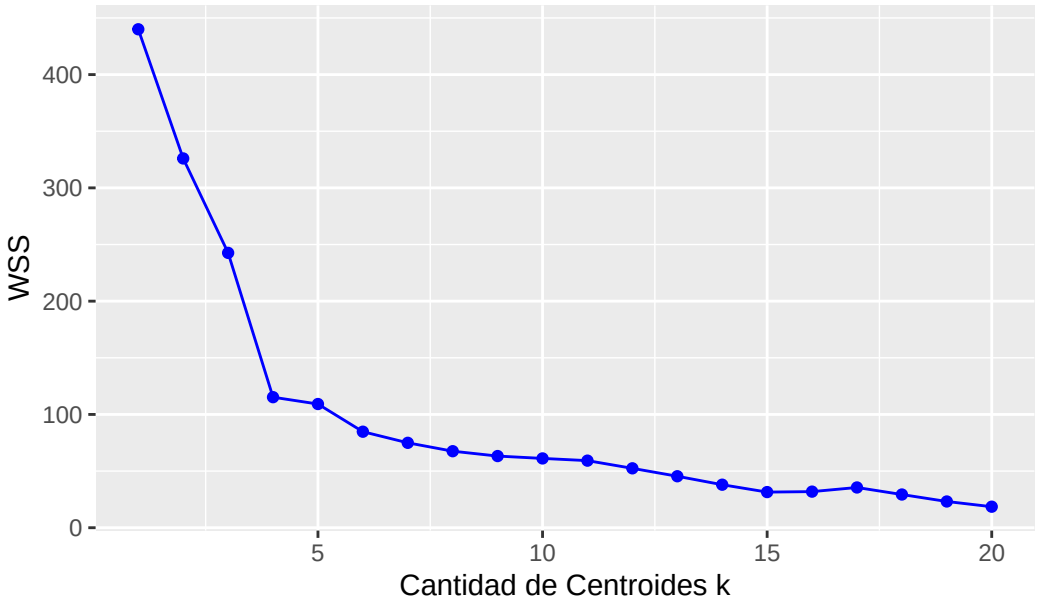


Figura 46: Gráfico de Codo para K-Means.

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.

3.2.1 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)
```

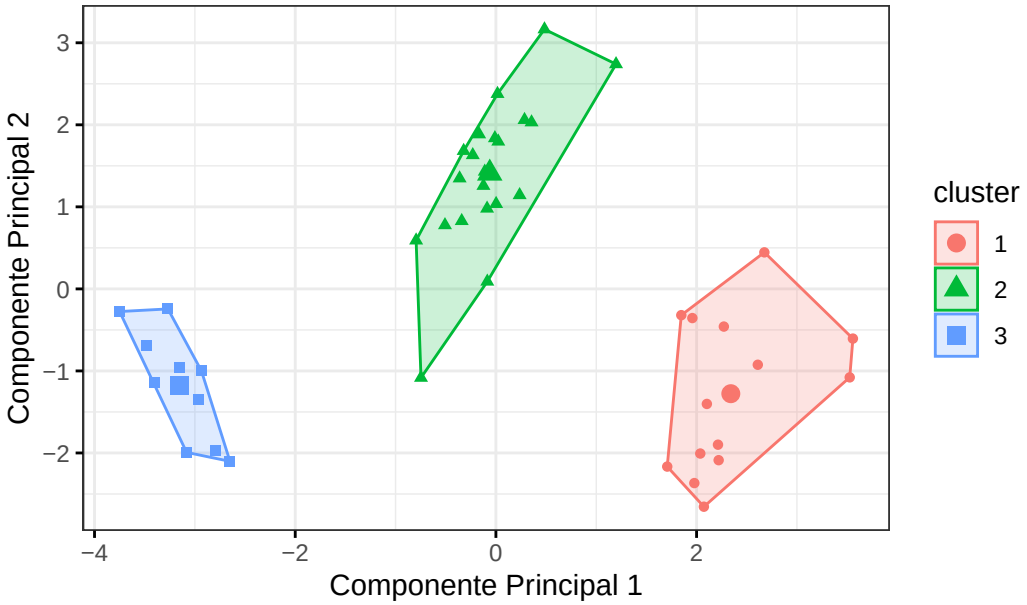


Figura 47: K-Means: 3 grupos.

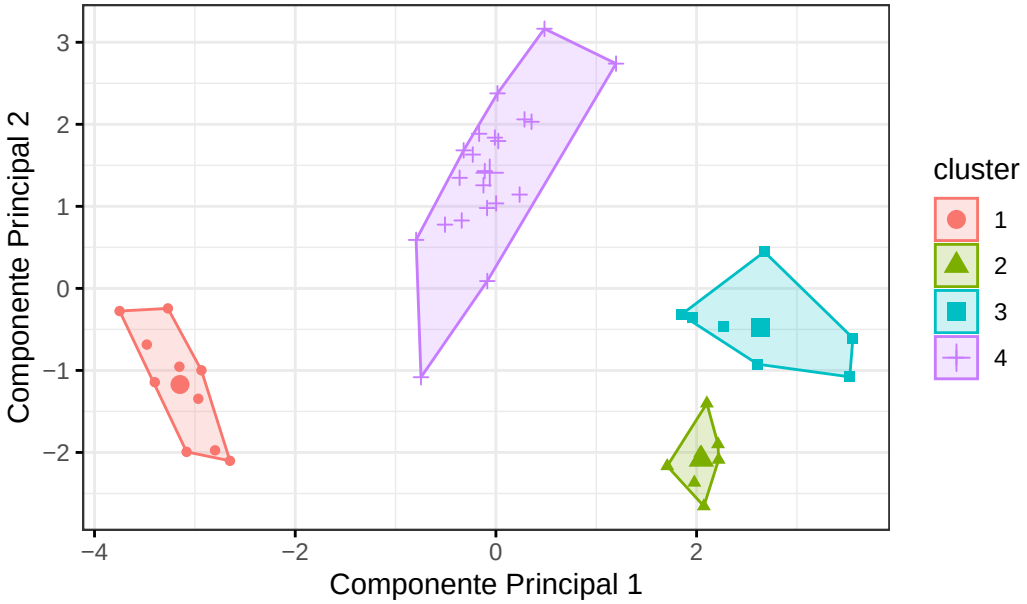


Figura 48: K-Means: 4 grupos.

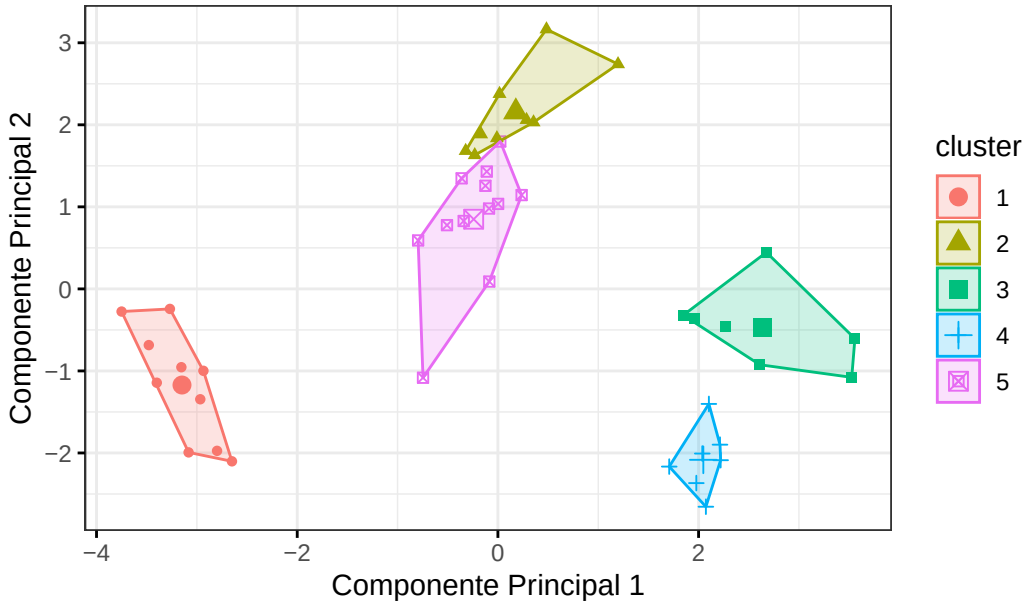


Figura 49: K-Means: 5 grupos.

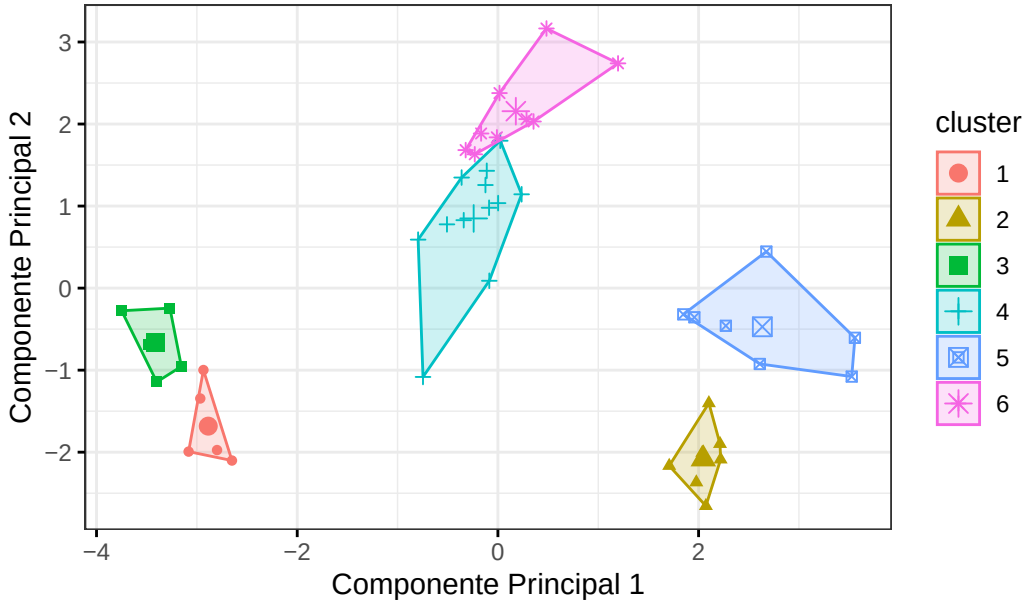


Figura 50: K-Means: 6 grupos.

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.

3.3 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

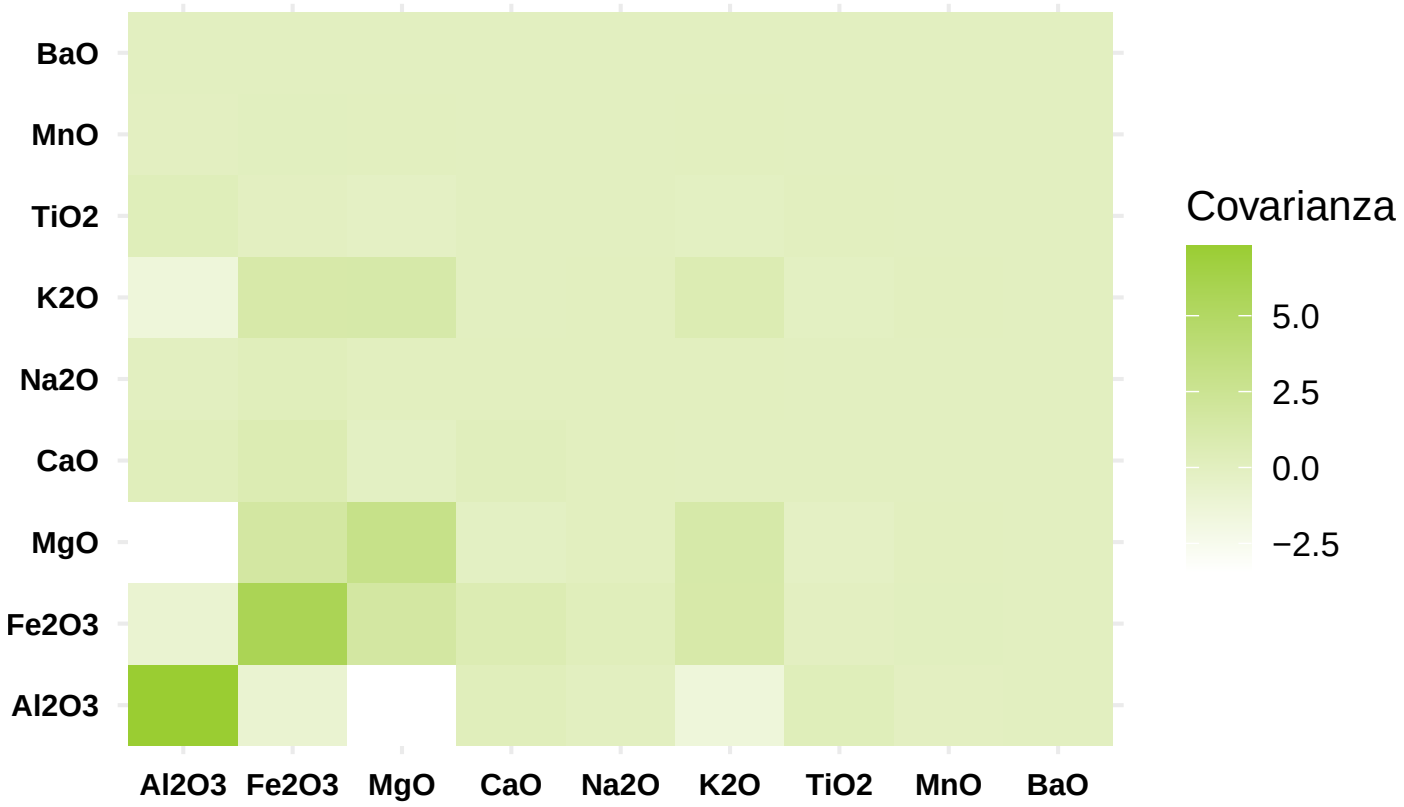


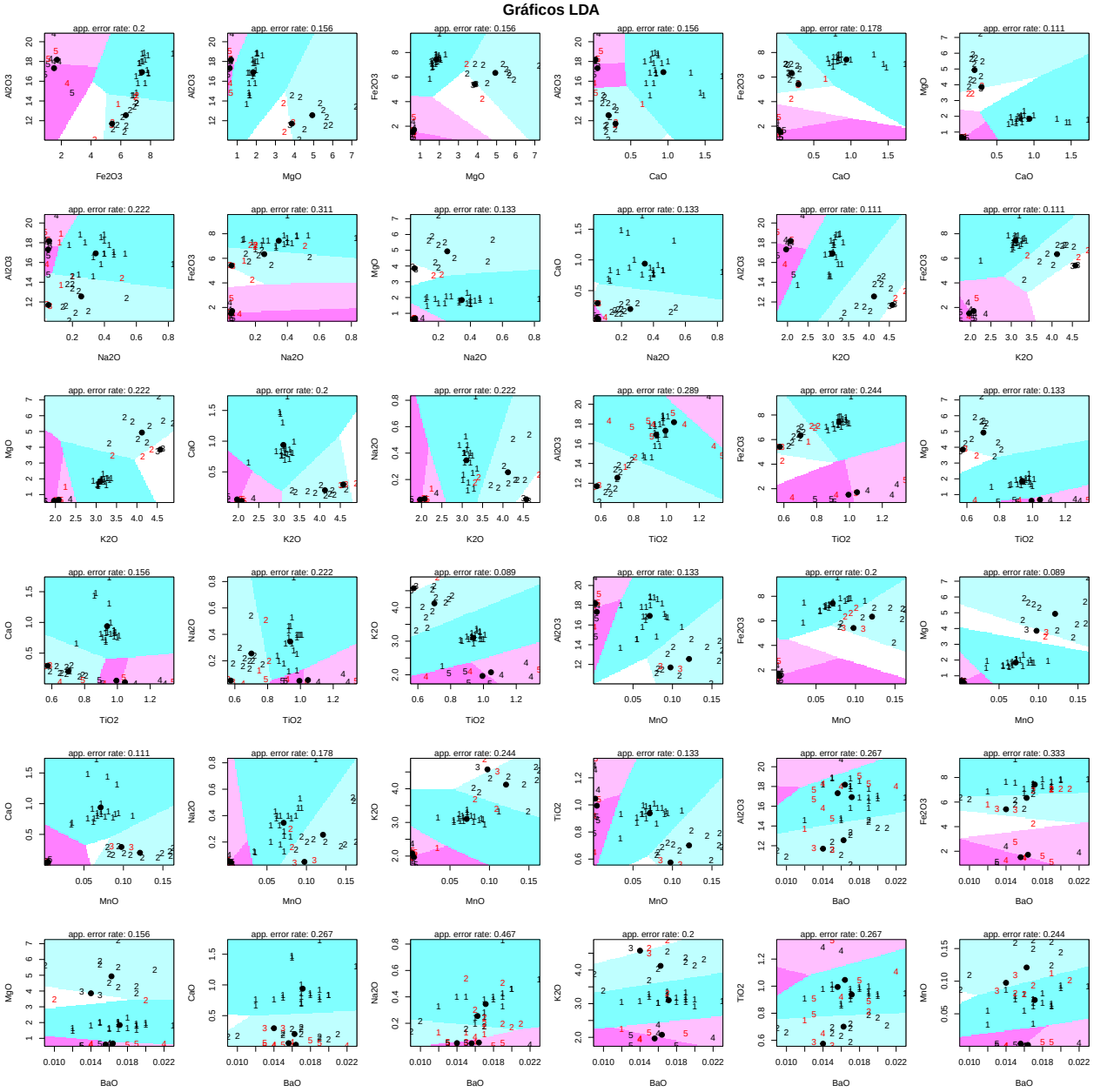
Figura 51: Mapa de Calor de la Matrix de Covarianzas.

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.

3.3.1 Discriminante Lineal.

3.3.1.1 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.

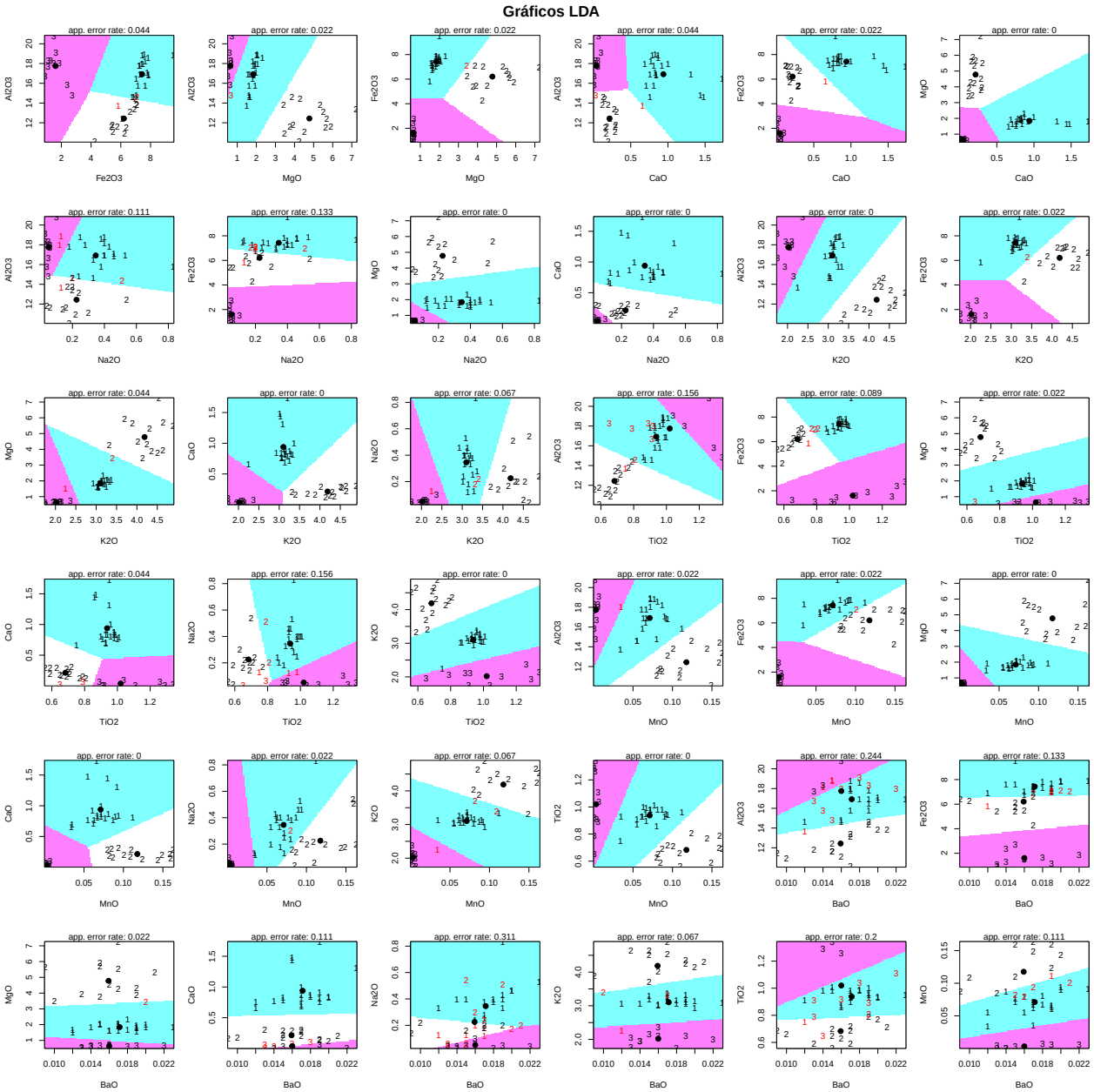
3.3.1.2 Predicción de Región.

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.

3.3.2 Discriminante Cuadrático.

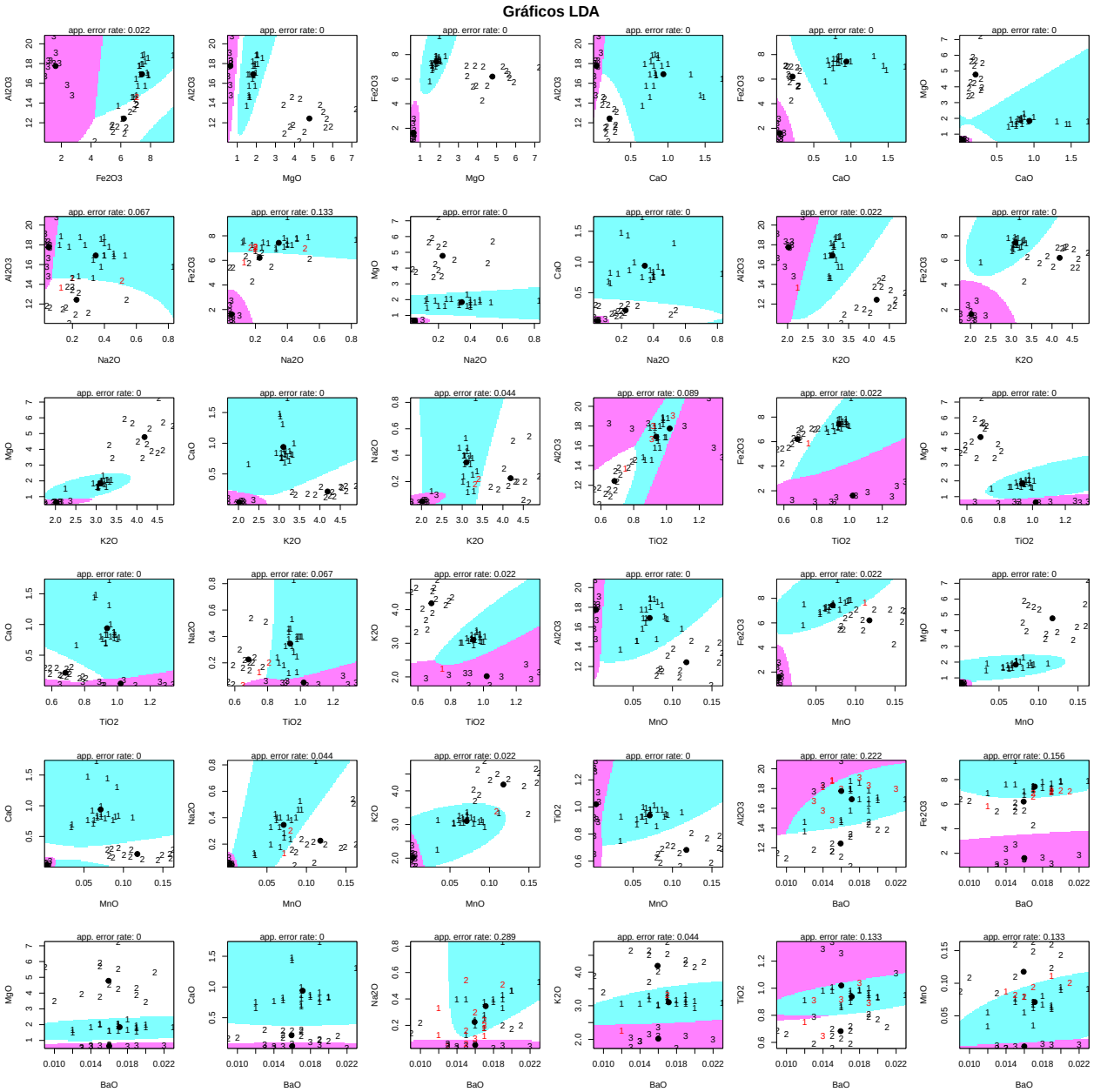
3.3.2.1 Predicción de Horno.

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.

3.3.2.2 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.

3.4 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.