Gigi Causio Voinea

Análsis de datos de proximidad

MDS métrico y no métrico

2013

Estadística Aplicada

1.1. Realizar en R el análisis de los datos kinshipdelta del paquete smacof.

1.2. Realizar en R el análisis de los datos trading del paquete smacof.

1.1. Realizar en R el análisis de los datos kinshipdelta del paquete smacof.

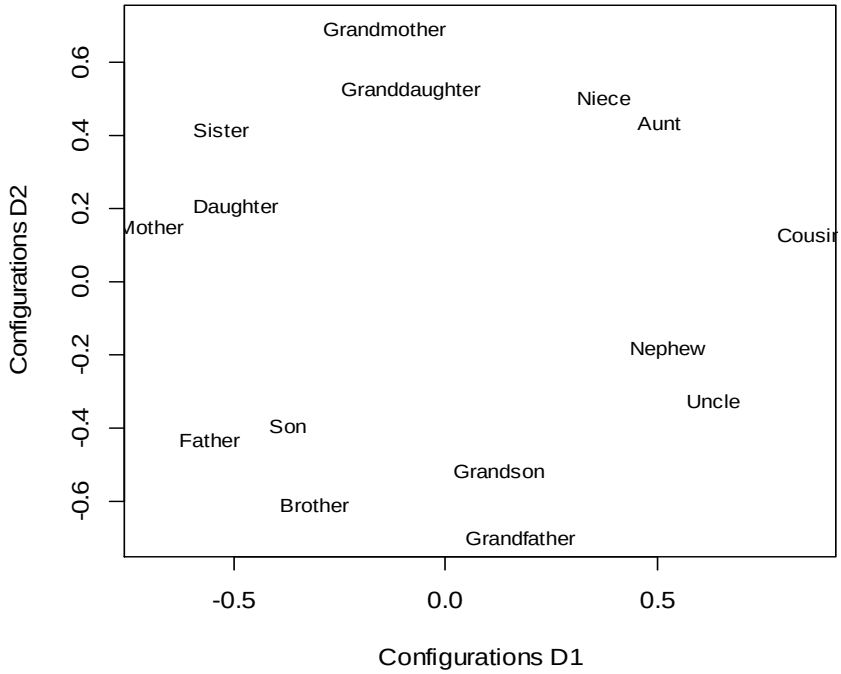
Los datos corresponden a una matriz de disimilaridades, simétrica entre *15* variable: Aunt, Brother, Cousin , Daughter, Father, Granddaughter, Grandfather, Grandmather, Grandson, Mother, Nepfew, Niece, Sister, Son y Uncle. Para este ejercicio se impone la solución simple de smacof, en dos dimensiones, métrica, de la siguiente forma:

```
>library (smacof)
> data (kinshipdelta)
> datos <- kinshipdelta
> simetricos <- smacofSym (datos, ndim=2, metric=T)
> summary (simetricos)
Configurations:
              D1        D2
Aunt       0.5052    0.4421
Brother   -0.3106   -0.6027
……
Stress per point:
             SPP     SPP(%)
Daughter  0.0367     3.6863
Son       0.0387     3.8858
……

> simetricos
Call: smacofSym (delta = datos, ndim = 2, metric = T, ties = "primary")
Model: Symmetric SMACOF
Number of objects: 15
Metric stress: 0.06988988
Number of iterations: 204
> plot (simetricos, main="Solución simple con smacof", cex.main=1)
```

En la *Figura 5.1* se observa la atracción entre nuestras variables, se observa que en la parte baja están las de género masculino y en la parte alta están las de género femenino, se observa los pequeños subgrupos: mother, daughter y sister; father, son y brother; grandson y granfather; grandmother y granddaughter; niece y aunt; nepfew y uncle; cousin. Ya se sabe que más cerca están las variables unas de otras, más atracción hay entre las mismas.

Figura 5.1: Solución simple con smacof



Tras realizar el análisis en dos dimensiones se observa que se han necesitado un número bastante grande de iteraciones, 204, para minimizar el Stress, hasta llegar a *0,0698*, sabiéndose que las funciones perdidas, como el *Stress de Kruskal* (1964), son índices que calcula el mal ajuste entre las proximidades y las distancias correspondientes. Toma valores entre *0-1*, donde *0* significa un ajuste perfecto y *1* un muy mal ajuste. El propio Kruskal sugirió, en el caso del *MDS ordinal*, los siguientes valores: 0,20 (un ajuste malo), 0,10 (un ajuste aceptable), 0,05 (un ajuste bueno), 0,025 (un muy buen ajuste) y 0,00(un ajuste perfecto). Por lo cual, nuestro ajuste es aceptable. Está similar a un coeficiente de correlación, con la excepción de que mide el mal ajuste y no el buen ajuste. Como bien se sabe, la correlación puede ser alta o baja por distintos razones. Por ejemplo puede ser alta por la culpa de los "outliers" o puede ser baja si la regresión no es lineal.

Para darnos cuenta mejor del ajuste se puede analizar el *Shepard diagram* y el *Stress per point*.

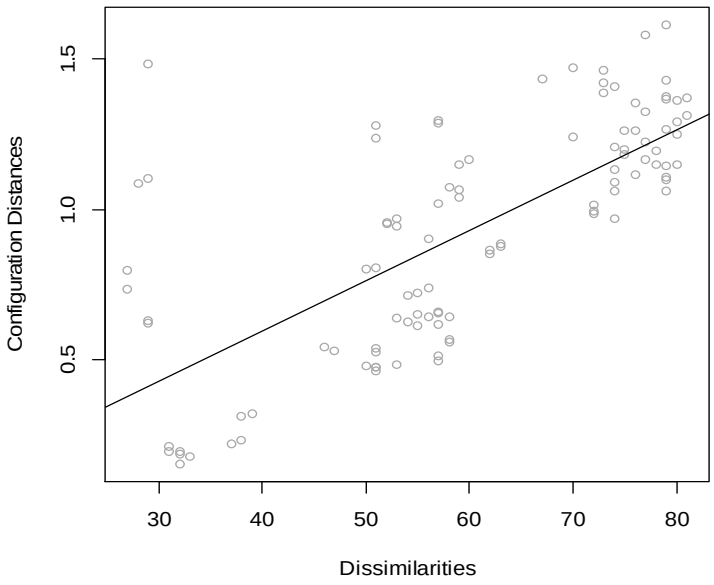
Diagrama Shepard (véase *Figura 5.2*) ha sido inventada por Shepard en 1957, y representa las similaridades frente a las disparidades o distancias. El diagrama es especialmente útil en *MDS ordinal*, pero la vamos a presentar también aquí, para la práctica y para completar nuestro análisis, donde se observa que el valor de 0,0698 no es consecuencia de una relación nelineal entre las similaridades y las distancias estimadas.

En la diagrama se puede observar también que el mayor error de ajuste están entre las variables: grandmother con granfather, grandson con granddaughter, niece con nepfew, auncle con aunt, son con daughter y mother con father.

Presentamos dos modalidades de representar el diagrama de Shepard:

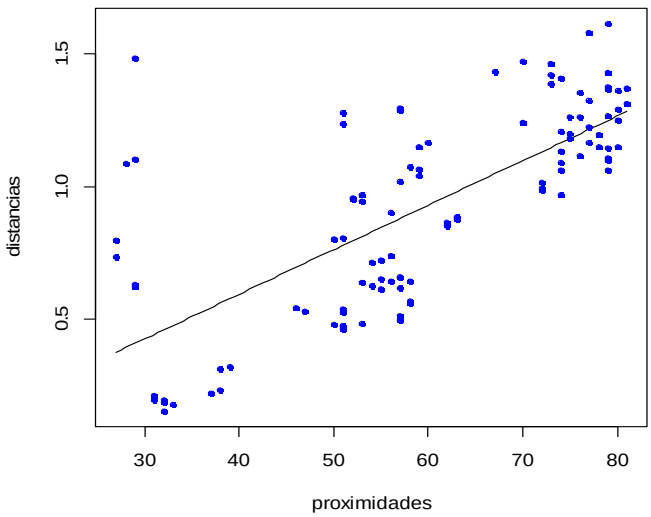
```
> dev.new()
> plot ( simetricos, plot.type="Shepard", main="Figura 5.2: Diagrama Shepard", cex.main=1)
```

Figura 5.2: Diagrama Shepard




```
> datos <- as.dist (datos)
> regresion<-lm (simetricos$confdiss~datos)
> curve (regresion$coef[1] + regresion$coef[2]*x,xlab ="proximidades", ylab ="distancias",
main="Figura 5.2: Diagrama Shepard", xlim = range (datos), ylim=range (simetricos$confdiss),
cex.main=1)
>  points (datos, simetricos$confdiss, pch=20, col="blue", bg="black")
```

Figura 5.2: Diagrama Shepard

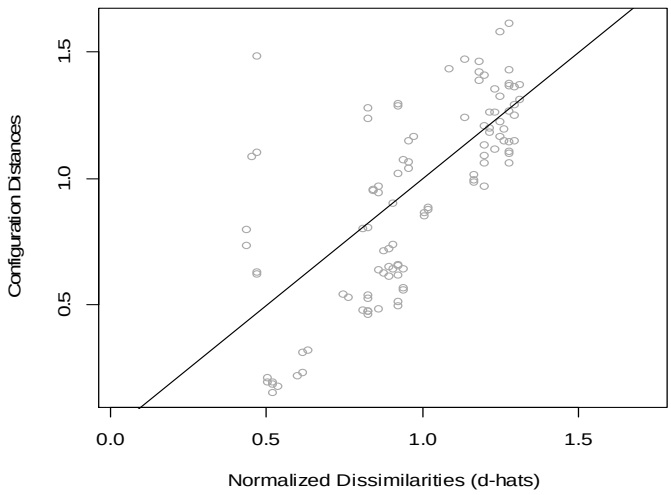


```
> plot (simetricos, plot.type="resplot", asp=1, main="Figura 5.3: Gráfico de residuos. MDS métrico", cex.main=1)
```

Stress per point es otra medida de análisis sobre los errores de ajuste que representa la media de los errores cuadráticos entre una variable y todas otras. En la tabla de Stress per point se observa el mayor stress lo tiene las variables: Grandmother igual a 0.1157, Grandfather igual a 0.1122, Grandson igual a 0.0803, Granddaughter igual a 0.0794, lo que también se ha visto reflejado en el diagrama de Shepard.

En último lugar, el gráfico de residuos nos muestra una dirección ascendente, nos muestra cual son los valores que mejor se ajusta (en nuestro caso los valores altos) al modelo, cual son los valores que peor se ajusta (los más pequeños) y nos muestra los "outliers".

Figura 5.3: Gráfico de residuos. MDS métrico



1.2. Realizar en R el análisis de los datos trading del paquete smacof. Representa una matriz de disimilaridades, entre 20 paises, según el negocio.

```
>library (smacof)
> no_metrica <- smacofSym (trading, ndim=3, metric=F,ties="secondary")
> no_metrica
Call: smacofSym (delta = trading, ndim = 3, metric = F, ties = "secondary")
Model: Symmetric SMACOF
Number of objects: 20
Nonmetric stress: 0.01129101
Number of iterations: 137
> summary (no_metrica)
Configurations:
          D1        D2        D3
Arge  -0.4454    0.1850    0.3048
Aust  -0.4703    0.1840   -0.0965
.....
```

Stress per point:

	SPP	SPP(%)
Czec	0.0029	1.3420
E.Ge	0.0049	2.2825

.....

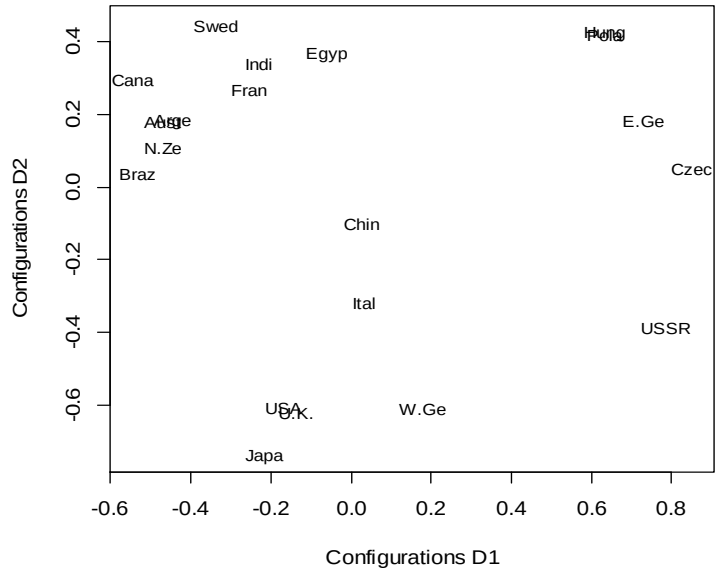
```
> plot (no_metrica, main="Figura 5.3: Solución no métrica en tres dimensiones", cex.main=1)
```

A diferencia del *modelo métrico*, el modelo de escalamiento *no métrico* no presupone una relación de tipo lineal entre proximidades y distancias, sino simplemente una relación de tipo monótona creciente, que puede corresponder a una línea recta, curva o quebrada. Es decir la relación no se establece en forma de sumas o productos, sino simplemente en términos de orden de proximidades. El primer paso es convertir las proximidades en disparidades, luego se calcula las distancias a partir de las disparidades utilizando una función de bondad de ajuste, en nuestro caso el Stress.

En la siguiente figura se observan los países con sus respectivas distancias. Este gráfico nos dice muchísimas cosas, por un lado se puede ver que en el medio está China que hace negocios en casi igual medida con todas las países (con Italia más), luego se observa los subgrupos: Brazil, New Zeeland, Argentina, Australia y quiza Canada; Frace, India, Egypt y quiza Sweden; Hungary con Poland; Czechoslovakia y East Germany; Usa con Uk y con Japan; Y algo más retirada el URSS.

Resumiendo los resultados del análisis vemos que hay 20 variables, se han tomano tres dimensiones para lo cual se han necesitado *137* iteraciones para llegar a un valor del Stress igual a *0.01129*, que por cierto es *un ajuste casi perfecto*. Podemos mirar la tabla de Stress per point y notar que lo que menos se ajusta al modelo es la variable Brazil con *0.0247* y China con *0,0228* y lo que mejor se ajusta es Czechoslovakia con *0,0029*.

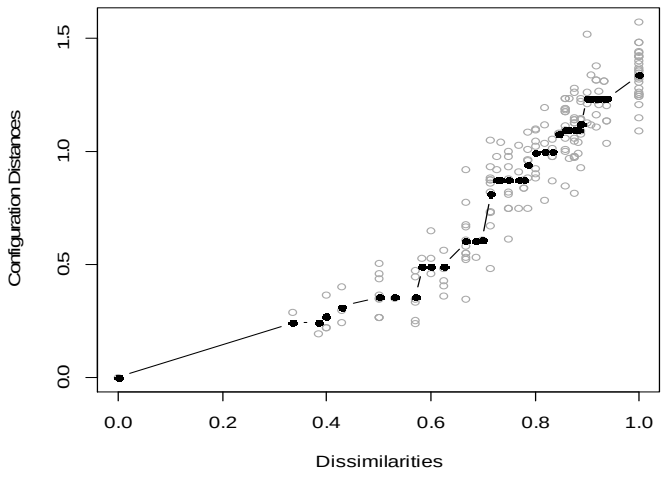
Figura 5.4: Solución no métrica en tres dimensiones




```
>  plot (no_metrica, plot.type="Shepard",main="Figura 5.5: Diagrama Shepard", cex.main=1)
```

En la *Figura 5.5* está representada el *diagrama de Shepard* que complementa la mediada del stress. Es un diagrama de dispersión que representa a las proximidades frente a las disparidades, que nos permite detectar los "outliers" y la forma de nuestra función. Sabiéndose que *la configuración estará bien ajustada cuando el diagrama refleja una función creciente, como es nuestro caso*. Un buen ajuste se representa con una diagonal en el gráfico.

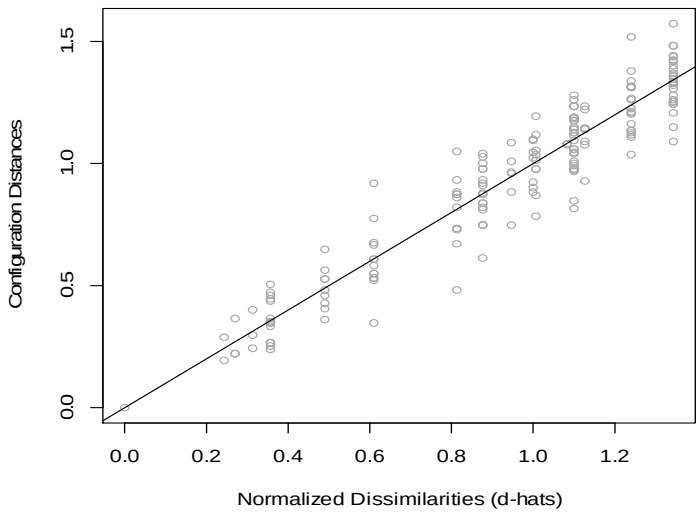
Figura 5.5: Diagrama Shepard




```
> plot (no_metrica, plot.type="resplot",main="Figura 5.5: Gráfico de residuos. MDS no métrico.", cex.main=1)
```

En la *Figura 5.5* tenemos el gráfico de los residuos que representa un ajuste lineal entre las distancias y las disparidades, donde no remarcamos anomalias.

Figura 5.5: Gráfico de residuos. MDS no métrico.



2. Realizar el análisis de diferencias individuales sobre los datos "perception".

Representan los datos de 42 sujetos, divididos en dos grupos a que se les presenta 120 estímulos. El primer grupo realiza la percepción en una escala de 0-9 en términos de anchura y alteza y el otro grupo en términos de dimensión y forma.

Se puede usar los modelos: "identity", "diagonal" y "idioscal".

```
> res.id <- smacofIndDiff (perception, constraint = "identity")
Call: smacofIndDiff(delta = perception, constraint = "identity")
Model: Three-way SMACOF
Number of objects: 16
Metric stress: 0.08491675
Number of iterations: 33
> res.diag <- smacofIndDiff (perception, constraint = "diagonal")
Call: smacofIndDiff(delta = perception, constraint = "diagonal")
Model: Three-way SMACOF
Number of objects: 16
Metric stress: 0.05531428
Number of iterations: 114
> res.idio <- smacofIndDiff (perception, constraint = "idioscal")
Call: smacofIndDiff(delta = perception, constraint = "idioscal")
Model: Three-way SMACOF
Number of objects: 16
Metric stress: 0.05531424
Number of iterations: 32
```

Si hacemos una primera comparación entre los tres métodos se observa que el stress más bajo se obtiene con el método "idioscal" y con el "diagonal", pero la diferencia es que el primero obtiene este valor del stress (0,055) con mucho menos iteraciones. Un coeficiente stress de 0,055 significa un buen ajuste. A continuación vemos las configuraciones para cada modelo por separado.

Remarcamos que la configuración del modelo "diagonal" es distinta de las otras dos configuraciones.

Figura 5.6: Configuración con el modelo <identity>

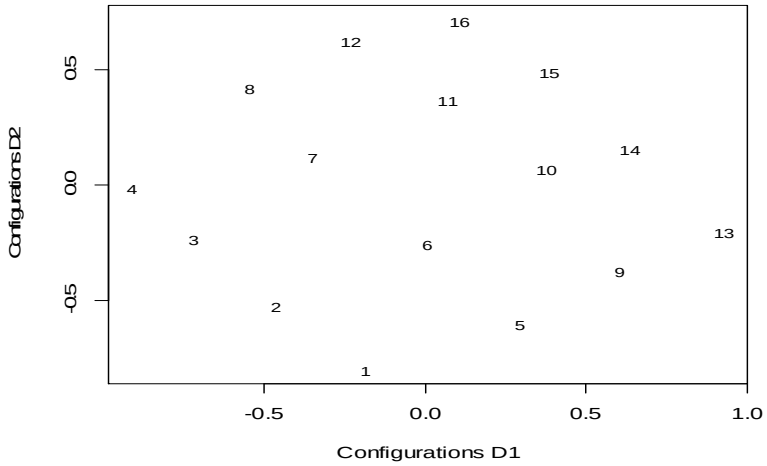


Figura 5.7: Configuración con el modelo <diagonal>

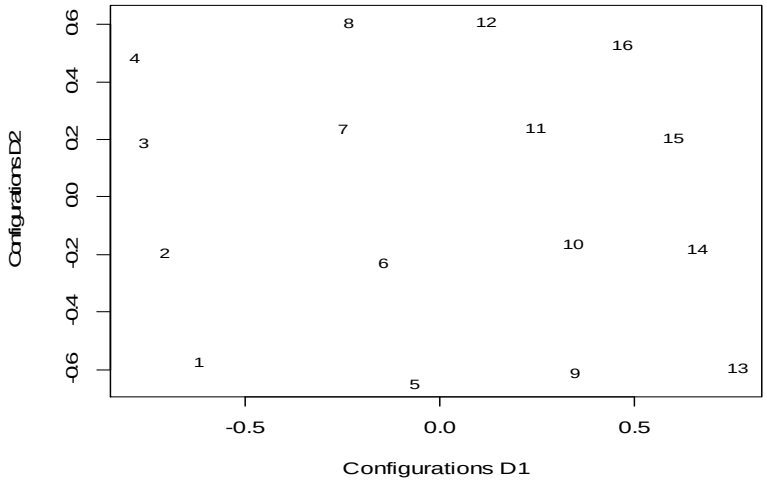
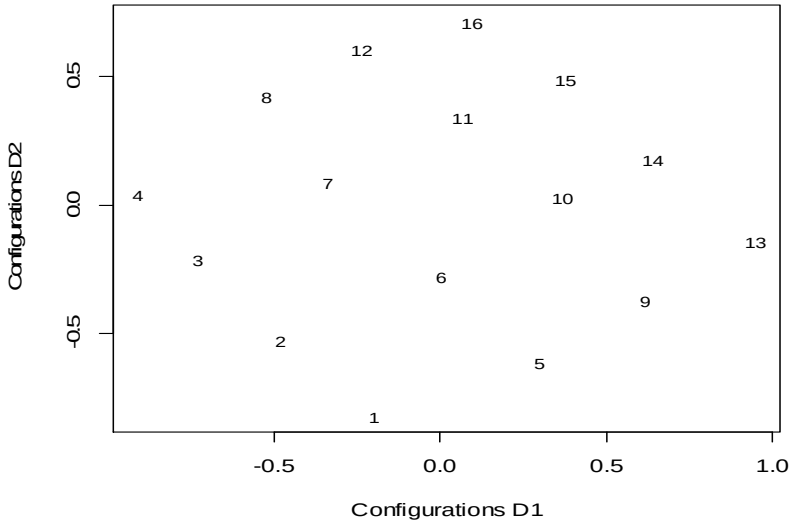


Figura 5.8: Configuración con el modelo <idioscal>



```
> plot (res.id, plot.type="Shepard", main="Figura 5.9: Diagrama Shepard. Modelo <identity>", cex.main=1)
> plot(res.diag, plot.type="Shepard", main="Figura 5.10: Diagrama Shepard. Modelo <diagonal>", cex.main=1)
> plot(res.idio, plot.type="Shepard", main="Figura 5.11: Diagrama Shepard. Modelo <idioscal>", cex.main=1)
```

Figura 5.9: Diagrama Shepard. Modelo <identity>

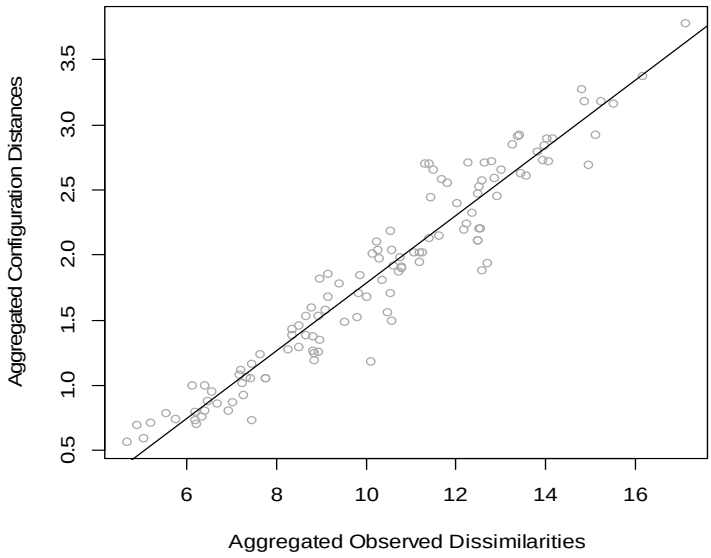


Figura 5.10: Diagrama Shepard. Modelo <diagonal>

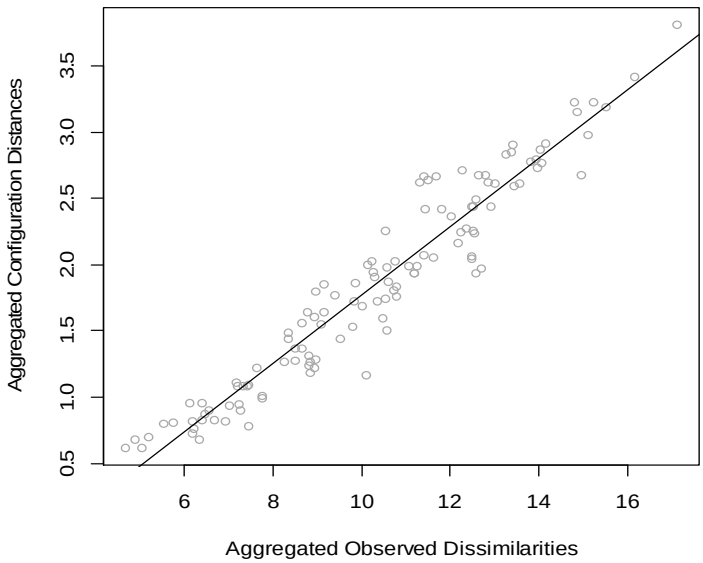
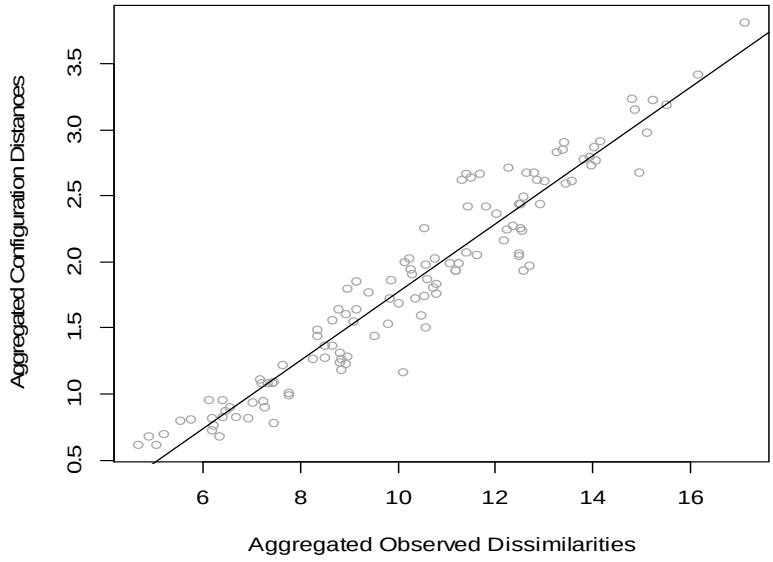


Figura 5.11: Diagrama Shepard. Modelo <idioscal>

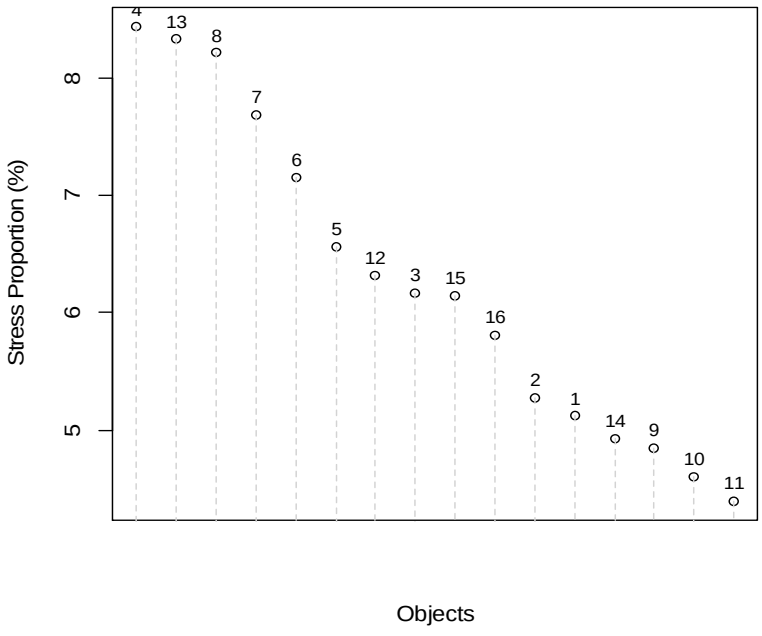


Por el otro lado, analizando las diagramas de Shepard para cada uno de los modelos se observa que hay una relación lineal creciente, que no existe diferencias entre los tres modelos y que el valor del stress obtenido justo.

Si queremos ver un gráfico de los valores del stress per point, procedamos como lo siguiente:

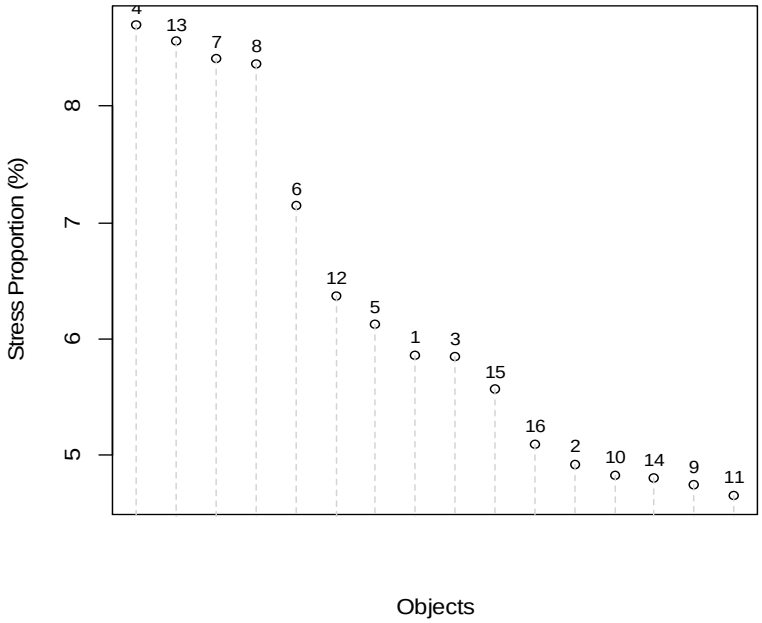
```
>plot (res.id, plot.type="stressplot", main="Figura 5.9: Gráfico stress para el modelo identity)" ,
cex.main=1)
```

Figura 5.12: Gráfico stress per point. Modelo <identity>)



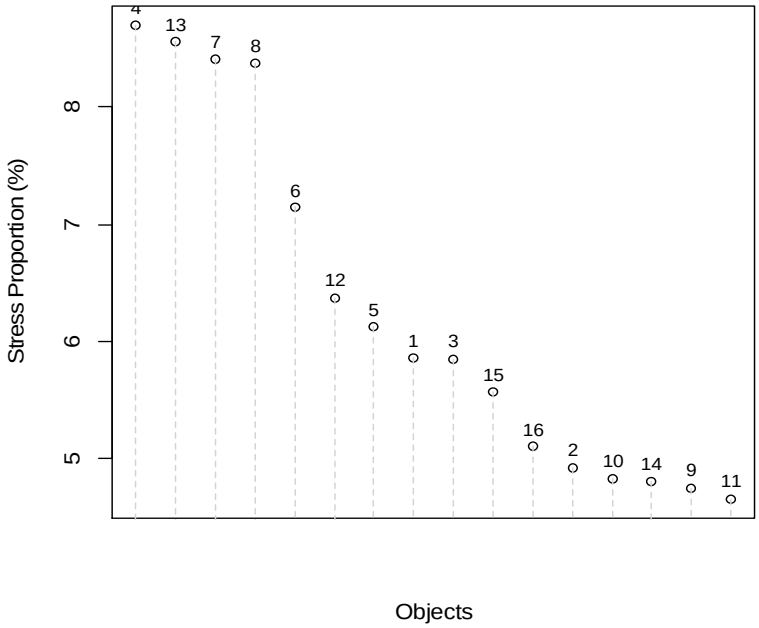
```
> plot(res.diag, plot.type="stressplot", main="Figura 5.13: Gráfico stress per point. Modelo
<diagonal>", cex.main=1)
```

Figura 5.13: Gráfico stress per point. Modelo <diagonal>




```
> plot(res.idio, plot.type="stressplot", main="Figura 5.14: Gráfico stress per point. Modelo
<idioscal>", cex.main=1)
```

Figura 5.14: Gráfico stress per point. Modelo <idioscal>



3.1. Usando los datos de la Tabla 4.1 de colors de Helm, (1959):

- *Leer los datos con SPSS.*
- *Realizar el análisis de los datos usando ALSCAL SPSS para el modelo identidad.*
- *Realizar el análisis de los datos con ALSCAL de SPSS para el modelo de diferencias individuales.*
- *Compara los resultados con los obtenidos mediante PROXSCAL*.

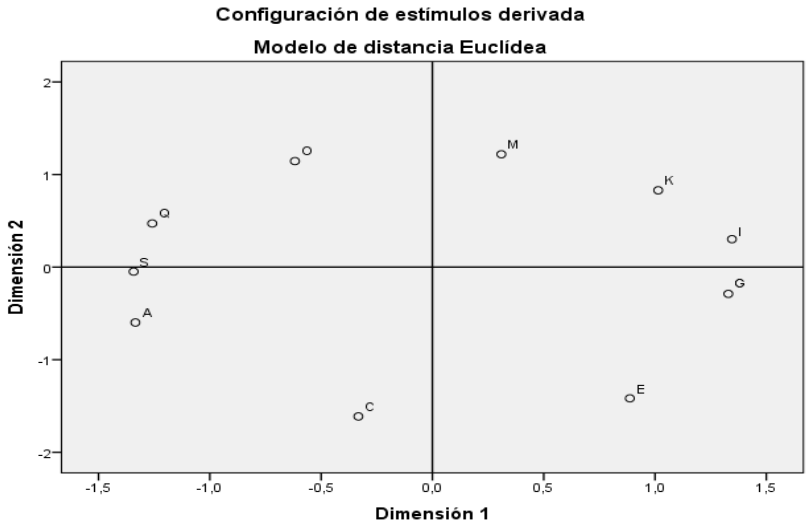
3.2. *Realizar el análisis de los datos usando ALSCAL SPSS para el modelo identidad.*

```
GET
  FILE='C:\Users\daniel\Desktop\Datos de proximidad\NOU.sav'.
DATASET NAME Conjunto_de_datos1 WINDOW=FRONT.
ALSCAL
  VARIABLES=A C E G I K M O Q S
  /SHAPE=SYMMETRIC
  /LEVEL=ORDINAL
  /CONDITION=MATRIX
  /MODEL=EUCLID
  /CRITERIA=CONVERGE(0.001) STRESSMIN(0.005) ITER(30) CUTOFF(0) DI-
MENS(2,2)
  /PLOT=DEFAULT ALL
  /PRINT=DATA HEADER.
```

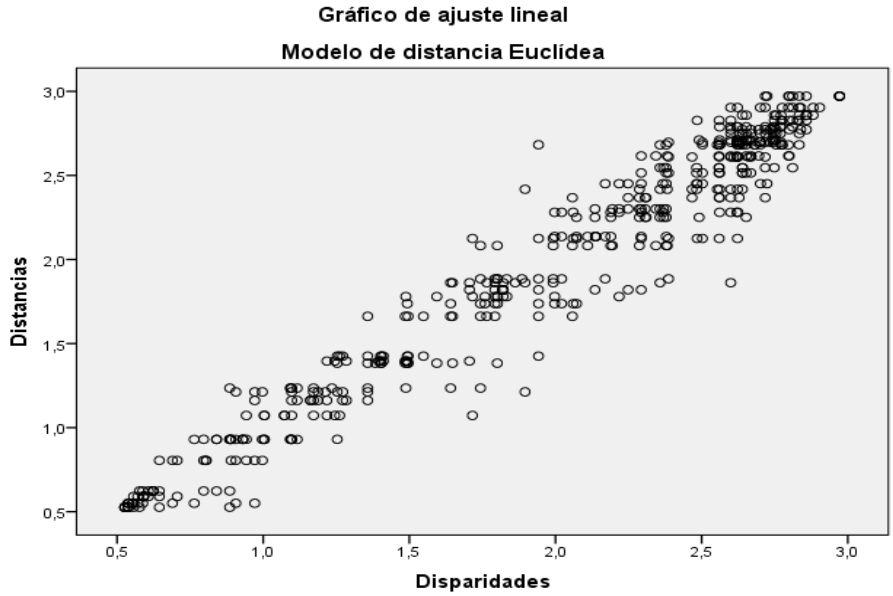
Como se puede ver SPSS nos devuelve una información general sobre nuestros datos. Se trata de las variables A,C,E,G,I,K,M,O,Q y S, que tienen datos ordinales en matrices simétricas. Para esto utilizamos el *modelo identidad* (*modelo euclídeo*) con un número de iteraciones máximas de 30 para la minimización del stress hasta *0,005* en dos dimensiones.

Después de dos iteraciones ya no se puede mejorar el **stress** que queda con un valor de *0,08158* que se considera un valor aceptable. Como se puede ver, el SPSS nos devuelve otro índice, a pesar del Stress, de ajuste del modelo a nuestros datos, el RSQ. Este índice es una correlación entre las disparidades derivadas a partir de los datos originales, y las disparidades derivadas por el modelo de escalamiento, de modo que puede ser interpretado como la proporción de varianza en las disparidades que es explicada por las distancias. Los límites del *RSQ* son entre cero y el uno, donde cero significa un ajuste malo y uno un ajuste muy bueno, por lo cual nuestro coeficiente conseguido de *0,94810* significa un muy buen ajuste.

En el gráfico “Configuración de estímulos derivada. Modelo de distancia euclídea” se puede ver la posición de cada variable en relación con las demás. Se observa los subgrupos: A,S y Q, por un lado y K,I y G, por el otro lado, con comportamiento similar dentro del grupo, y se observa que la M, C, E y O, tienen comportamiento distinto a las demás variable.



Con la ayuda del *Stress* y del *RSQ* no tenemos claro si el ajuste es bueno o no, también podemos analizar el gráfico de los residuos, entre las disparidades y las distancias, donde observamos un *ajuste lineal creciente, que significa un buen ajuste* y que el valor bajo del stress no es consecuencia de una no linealidad. Sí se puede observar que los valores altos están mejor ajustados que los bajos, este efecto se debe a que el índice Stress busca el mejor ajuste entre disparidades y distancias al cuadrado, con lo que tiende a ajustar mejor las distancias mayores que las menores, por lo que el grado de ajuste para estas últimas tiende siempre a ser más bajo que para las primeras.

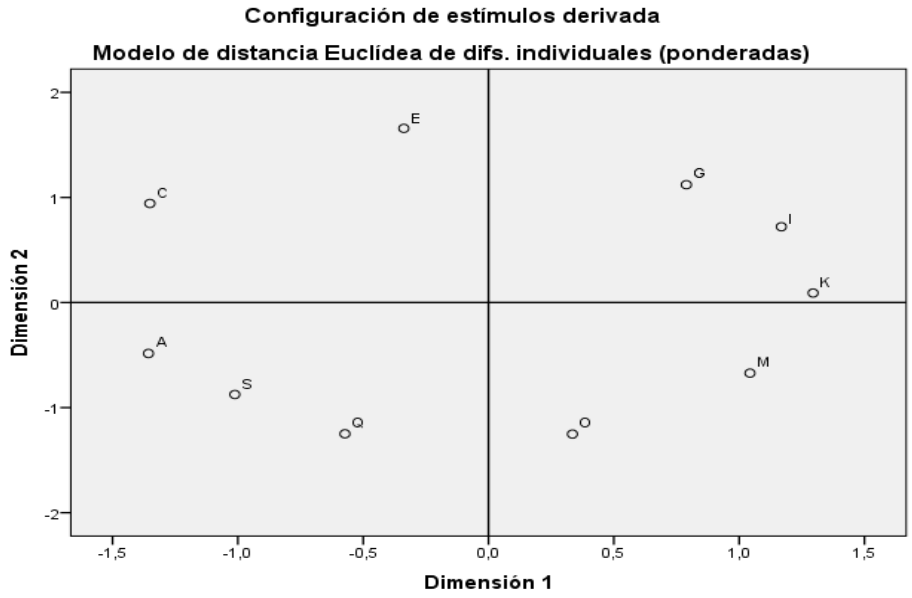


3.3 *Realizar el análisis de los datos con ALSCAL de SPSS para el modelo de diferencias individuales.*

```
ALSCAL
  VARIABLES=A C E G I K M O Q S
  /SHAPE=SYMMETRIC
  /LEVEL=ORDINAL
  /CONDITION=MATRIX
  /MODEL=INDSCAL
  /CRITERIA=CONVERGE(0.001) STRESSMIN(0.005) ITER(30) CUTOFF(0) DI-
MENS(2,2)
  /PLOT=DEFAULT
  /PRINT=DATA.
```

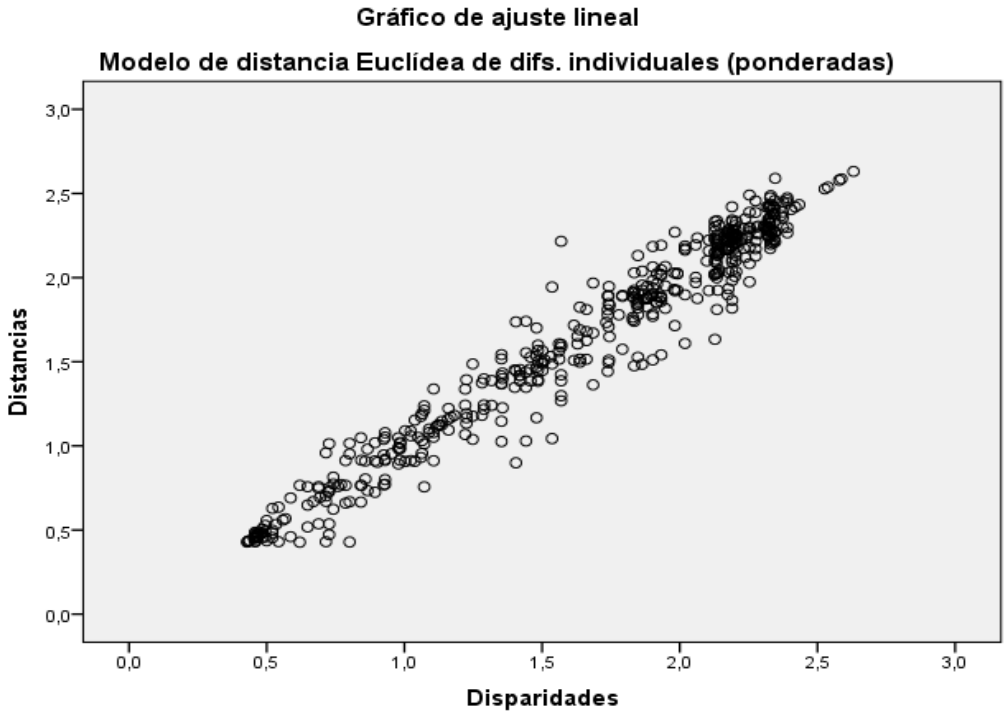
Entonces tenemos los mismos datos y las mismas condiciones, lo único que cambia es el modelo que pasamos del modelo identidad (euclideo) al *modelo de diferencias individuales* (*INDSCAL*).

Después de tres iteraciones se ha llegado a un nivel mínimo del stress de ***0,07584*** que también es *aceptable*, y es un pelín menor que en el caso del modelo identidad. El coeficiente RSQ toma el valor de ***0,95556***, que también es un pelín mejor que en el caso anterior. Veamos a continuación el gráfico de la configuración calculada y el de los residuos.



El gráfico de configuración con el modelo de diferencias individuales es algo diferente que el otro conseguido con el modelo identidad. Ahora tenemos los sub grupos G, I, K y M por un lado, el subgrupo A,S y Q por el otro lado, la variable O que está en el medio de estos dos subgrupos y las variables C y E con un comportamiento distinto.

Y, si analizamos el gráfico de los residuos observamos un mucho mejor ajuste a comparación con el otro modelo, **hay mucho menos dispersión y más linealidad**.



CONCLUSIÓN

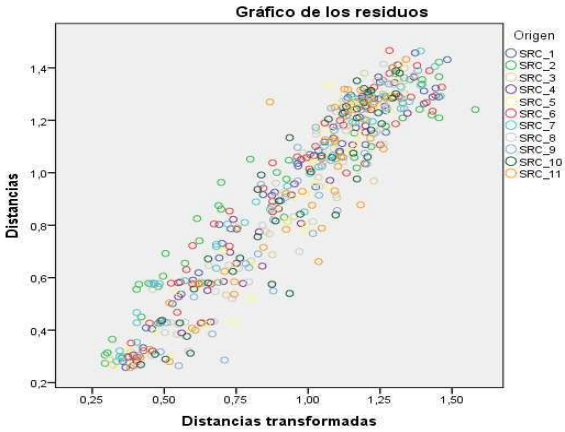
Con estos datos el modelo de diferencias individuales realiza una mejor configuración, esto nos lo dice el Stress, el RSQ y en principal el gráfico de los residuos.

3.3 Compara los resultados con los obtenidos mediante PROXSCAL

Recordamos los resultados obtenidos con el PROXSCAL:

Modelo de diferencias individuales: S stress=0,03582

Haciendo una comparación entre los dos modelos identidad y diferencias individuales, observamos que en ambos algoritmos, PROXSCAL y ALSCAL, el modelo de las diferencias individuales nos produce menos Stress y un mejor ajuste. Con el PROXSCAL el stress está más pequeño, pero con ALSCAL el grafico de los residuos tiene menos dispersión y más linealidad. Y, si miramos la configuración de los puntos en común también es muy parecida entre los dos algoritmos, con el modelo de diferencias individuales. Por lo tanto, nos quedamos con el modelo de diferencias individuales tanto con PROXSCAL como con ALSCAL.



a) Modelo de identidad: S stress=0,03974