EL ANÁLISIS DE ESCALAMIENTO MULTIDIMENSIONAL: UNA ALTERNATIVA Y UN COMPLEMENTO A OTRAS TÉCNICAS MULTIVARIANTES.

Dra. Flor María Guerrero Casas
José Manuel Ramírez Hurtado

Departamento de Economía y Empresa
Universidad Pablo de Olavide
Ctra. de Utrera, km. 1 - 41013 SEVILLA (ESPAÑA)
Tfn. 95 434 9279-9171 / Fax: 95 434 9339
fguecas@dee.upo.es jmramhur@dee.upo.es

Resumen: En los últimos años la proliferación de datos y el fácil acceso a los mismos ha hecho que, en la mayoría de las investigaciones, se analicen grandes conjuntos de datos, utilizando para ello las técnicas multivariantes. En este sentido, hay que indicar que las técnicas multivariantes cobran cada vez mayor importancia en las investigaciones.

Dentro de las técnicas multivariantes podemos citar al Escalamiento Multidimensional *(Multidimensional Scaling, MDS)*. El MDS es una técnica multivariante de interdependencia que trata de representar en un espacio geométrico de pocas dimensiones las proximidades existentes entre un conjunto de objetos o de estímulos. Esta técnica, aunque tiene sus raíces a principios del siglo XX, hoy día sigue siendo infrautilizada en muchas áreas.

En este trabajo se pretende dar una visión general del funcionamiento del MDS, comparándolo con otras técnicas multivariantes más tradicionales como son el Análisis Factorial y el Análisis Cluster, de modo que pueda servir como alternativa y como complemento a las mismas en cualquier investigación que utilice dichas técnicas. También se incluye un análisis comparativo de los resultados de estas técnicas, mediante una aplicación a la infraestructura del sector turístico en Andalucía.

Palabras clave: Análisis multivariante, escalamiento, distancia, estímulo, análisis factorial, análisis cluster, turismo.

1. INTRODUCCIÓN.

El escalamiento multidimensional, más conocido como *MultiDimensional Scaling (MDS)*, tiene sus orígenes a principios de siglo XX en el campo de la Psicología. Surge cuando se pretendía estudiar la relación que existía entre la intensidad física de ciertos estímulos con su intensidad subjetiva.

El MDS es una técnica de representación espacial que trata de visualizar sobre un mapa un conjunto de estímulos (firmas, productos, candidatos políticos, ideas u otros artículos) cuya posición relativa se desea analizar. El propósito del MDS es transformar los juicios de similitud o preferencia llevados a cabo por una serie de individuos sobre un conjunto de objetos o estímulos en distancias susceptibles de ser representadas en un espacio multidimensional. El MDS está basado en la comparación de objetos o de estímulos, de forma que si un individuo juzga a los objetos A y B como los más similares entonces las técnicas de MDS colocarán a los objetos A y B en el gráfico de forma que la distancia entre ellos sea más pequeña que la distancia entre cualquier otro par de objetos.

En la actualidad, el MDS puede ser apto para gran cantidad de tipos diferentes de datos de entrada (tablas de contingencia, matrices de proximidad, datos de perfil, correlaciones, etc.).

El MDS puede ayudar a determinar:

- qué dimensiones utilizan los encuestados a la hora de evaluar a los objetos.
- cuántas dimensiones utilizan.
- la importancia relativa de cada dimensión.
- cómo se relacionan perceptualmente los objetos.

Existen otras técnicas multivariantes, como son el análisis factorial y el análisis cluster, que persiguen objetivos muy similares al MDS pero que difieren en una serie de aspectos. Sin embargo, la utilización de alguna de estas técnicas no supone que no se pueda utilizar el escalamiento multidimensional, sino que esta última técnica puede servir como alternativa o bien como complemento a las otras técnicas multivariantes.

En definitiva, el MDS es una técnica multivariante que crea un gráfico aproximado a partir de las similitudes o preferencias de un conjunto de objetos.

2. EL MODELO GENERAL DE ESCALAMIENTO MULTIDIMENSIONAL.

De modo general, podemos decir que el MDS toma como entrada una matriz de proximidades, $\Delta \in M_{nxn}$, donde n es el número de estímulos. Cada elemento δ_{ij} de Δ representa la proximidad entre el estímulo i y el estímulo j.

$$\Delta = \begin{pmatrix} \delta_{11} & \delta_{12} & \cdots & \delta_{1n} \\ \delta_{21} & \delta_{22} & \cdots & \delta_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{n1} & \delta_{n2} & \cdots & \delta_{nn} \end{pmatrix}$$

A partir de esta matriz de proximidades el MDS nos proporciona como salida una matriz $X \in M_{nxm}$, donde n, al igual que antes, es el número de estímulos, y m es el número de dimensiones. Cada valor x_{ij} representa la coordenada del estímulo i en la dimensión j (más adelante veremos el procedimiento para obtener esta matriz).

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}$$

A partir de esta matriz X se puede calcular la distancia existente entre dos estímulos cualesquiera i y j, simplemente aplicando la fórmula general de la distancia de Minkowski:

$$d_{ij} = \left[\sum_{t=1}^{m} (x_{it} - x_{jt})^p \right]^p$$

donde p puede ser un valor entre 1 e infinito. A partir de estas distancias podemos obtener una matriz de distancias que denominamos $D \in M_{nxn}$:

$$D = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{pmatrix}$$

La solución proporcionada por el MDS debe ser de tal modo que haya la máxima correspondencia entre la matriz de proximidades inicial Δ y la matriz de distancias obtenidas D. Para que exista la máxima correspondencia MDS proporciona varias medidas, que veremos más adelante, y que nos informan sobre la bondad del modelo.

3. MODELOS DE ESCALAMIENTO MULTIDIMENSIONAL.

Existen dos modelos básicos de MDS que son: el modelo de escalamiento métrico y el modelo de escalamiento no métrico. En el primero de ellos consideramos que los datos están medidos en escala de razón o en escala de intervalo y en el segundo consideramos que los datos están medidos en escala ordinal. No se ha desarrollado todavía ningún modelo para datos en escala nominal.

- Modelo de escalamiento métrico.-

Todo modelo de escalamiento parte de la idea de que las distancias son una función de las proximidades, es decir, $d_{ij}=f(\delta_{ij})$. En el modelo de escalamiento métrico partimos del supuesto de que la relación entre las proximidades y las distancias es de tipo lineal: $d_{ij}=a+b\delta_{ij}$. El primer procedimiento de escalamiento métrico se debió a Torgerson (1952, 1958) quién se basó en un teorema de Young y Householder (1938), según el cual a partir de una matriz de distancias, $D \in M_{nxn}$, se puede obtener una matriz $B \in M_{nxn}$ de productos escalares entre vectores. El procedimiento consiste en transformar la matriz de proximidades $\Delta \in M_{nxn}$ en una matriz de distancias $D \in M_{nxn}$, de tal forma que verifique los tres axiomas de la distancia euclídea:

1. No negatividad	$d_{ij} \geq 0 = d_{ii}$
2. Simetría	$d_{ij} = d_{ji}$
3. Desigualdad triangular	$d_{ij} \leq d_{ik} + d_{kj}$

Tabla: Axiomas de la distancia euclídea.

Los dos primeros axiomas son fáciles de cumplir, pero el tercer axioma no se cumple siempre. Este problema se conoce con el nombre de "estimación de la constante aditiva". Torgerson solucionó este problema, estimando el valor mínimo de *c* que verifica la desigualdad triangular de la siguiente forma:

$$c_{\min} = \max_{(i,j,k)} \{\delta_{ij} - \delta_{ik} - \delta_{kj}\}$$

De esta forma las distancias se obtienen sumando a las proximidades la constante *c*, es decir, $d_{ij}=\delta_{ij}+c$. Por ejemplo, supongamos que tenemos la siguiente matriz de proximidades:

$$\Delta = \begin{pmatrix} 0 & 1 & 5 \\ 1 & 0 & 2 \\ 5 & 2 & 0 \end{pmatrix}$$

Esta matriz no verifica la desigualdad triangular puesto que no se cumple que $\delta_{13} \leq \delta_{12}+\delta_{23}$ (5>1+2). Para calcular el valor mínimo de la constante aditiva *c* tendríamos que calcular todas las diferencias tal como se ha señalado anteriormente. En este caso se tendría que calcular 5-1-2=2. Estas diferencias las haríamos para todos los subíndices, obteniéndose que el valor mínimo de *c* es 2. La matriz de distancias sería en este caso:

$$D = \begin{pmatrix} 0 & 3 & 7 \\ 3 & 0 & 4 \\ 7 & 4 & 0 \end{pmatrix}$$

Una vez obtenida la matriz $D \in M_{nxn}$ es necesario transformarla en una matriz $B \in M_{nxn}$ de productos escalares entre vectores mediante la siguiente transformación:

$$b_{ij} = -\frac{1}{2}(d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d_{..}^2) \quad \text{donde:}$$

$$d_{i.}^2 = \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2 \qquad \text{(distancia cuadrática media por fila)}$$

$$d_{.j}^2 = \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2 \qquad \text{(distancia cuadrática media por columna)}$$

$$d_{..}^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2$$ (distancia cuadrática media de la matriz)

Una vez llegados a este punto, lo único que queda es transformar la matriz $B \in M_{nxn}$ en una matriz $X \in M_{nxm}$ tal que $B=X \cdot X'$, siendo X la matriz que nos da las coordenadas de cada uno de los n estímulos en cada una de las m dimensiones. Cualquier método de factorización permite transformar B en $X \cdot X'$.

En resumen el procedimiento consiste en transformar:

Δ (Proximidades) $\rightarrow D$ (Distancias) $\rightarrow B$ (Productos escalares) $\rightarrow X$ (coordenadas)

- Modelo de escalamiento no métrico.-

A diferencia del escalamiento métrico, el modelo de escalamiento no métrico no presupone una relación lineal entre las proximidades y las distancias, sino que establece una relación monótona creciente entre ambas, es decir, si $\delta_{ij} < \delta_{kl} \Rightarrow d_{ij} \leq d_{kl}$. Su desarrollo se debe a Shepard (1962) quién demostró que es posible obtener soluciones métricas asumiendo únicamente una relación ordinal entre proximidades y distancias. Posteriormente Kruskal (1964) mejoró el modelo. El procedimiento se basa en los siguientes apartados:

1) Transformación de la matriz de proximidades en una matriz de rangos, desde 1 hasta (n (n - 1))/2.
2) Obtención de una matriz $X \in M_{nxm}$ de coordenadas aleatorias, que nos da la distancia entre los estímulos.
3) Comparación de las proximidades con las distancias, obteniéndose las disparidades (d'_{ij}).
4) Definición del Stress.
5) Minimización del Stress.

Tanto para el modelo métrico como para el modelo no métrico es necesario obtener un coeficiente que nos informe sobre la bondad del modelo. Sabemos que las distancias son una función de las proximidades, es decir:

$$f: \delta_{ij}(x) \rightarrow d_{ij}(x)$$

De esta forma se tiene que $d_{ij}=f(\delta_{ij})$. Esto no deja ningún margen de error, sin embargo, en las proximidades empíricas es difícil que se dé la igualdad, con lo que generalmente ocurre que $d_{ij} \approx f(\delta_{ij})$. A las transformaciones de las proximidades por f se le denomina *disparidades*. A partir de aquí podemos definir el error cuadrático como:

$$e_{ij}^2 = (f(\delta_{ij}) - d_{ij})^2$$

Como medida que nos informa de la bondad del modelo podemos utilizar el *Stress* que Kruskal definió como:

$$Stress = \sqrt{\frac{\sum_{i,j}(f(\delta_{ij}) - d_{ij})^2}{\sum_{i,j} d_{ij}^2}}$$

Mientras mayor sea la diferencia entre las disparidades y las distancias, es decir, entre $f(\delta_{ij})$ y d_{ij}, mayor será el *Stress* y por tanto peor será el modelo. Por tanto, el *Stress* no es propiamente una medida de la bondad del ajuste, sino una medida de la no bondad o "maldad" del ajuste. Su valor mínimo es 0, mientras que su límite superior para n estímulos es $\sqrt{1-(2/n)}$.

Kruskal (1964) sugiere las siguientes interpretaciones del *Stress*:

- 0.2 → Pobre
- 0.1 → Aceptable
- 0.05 → Bueno
- 0.025 → Aceptable
- 0.0 → Excelente

También se suele utilizar una variante del *Stress* que se denomina *S-Stress*, definida como:

$$S - Stress = \sqrt{\frac{\sum_{i,j}(f(\delta_{ij})^2 - d_{ij}^2)^2}{\sum_{i,j}(d_{ij}^2)^2}}$$

Otra medida que se suele utilizar es el coeficiente de correlación al cuadrado (*RSQ*), que nos informa de la proporción de variabilidad de los datos de partida que es explicada por el modelo. Los valores que puede tomar oscilan entre 0 y 1, al ser un coeficiente de correlación al cuadrado. Valores cercanos a 1 indican que el modelo es bueno y valores cercanos a 0 indican que el modelo es malo. Su expresión es:

$$RSQ = \frac{\left[\sum_i \sum_j (d_{ij} - d_{..})(f(d_{ij}) - f(d_{..}))\right]^2}{\left[\sum_i \sum_j (d_{ij} - d_{..})^2\right]\left[\sum_i \sum_j (f(d_{ij}) - f(d_{..}))^2\right]}$$

La mayoría de los paquetes estadísticos tienen implementados tanto los algoritmos para obtener soluciones con MDS así como las medidas para determinar si el modelo es

adecuado o no[1]. En la actualidad todo los algoritmos implementados en los paquetes estadísticos son reiterativos, de forma que se alcance la mejor solución posible.

4. RELACIÓN ENTRE MDS Y OTRAS TÉCNICAS MULTIVARIANTES.

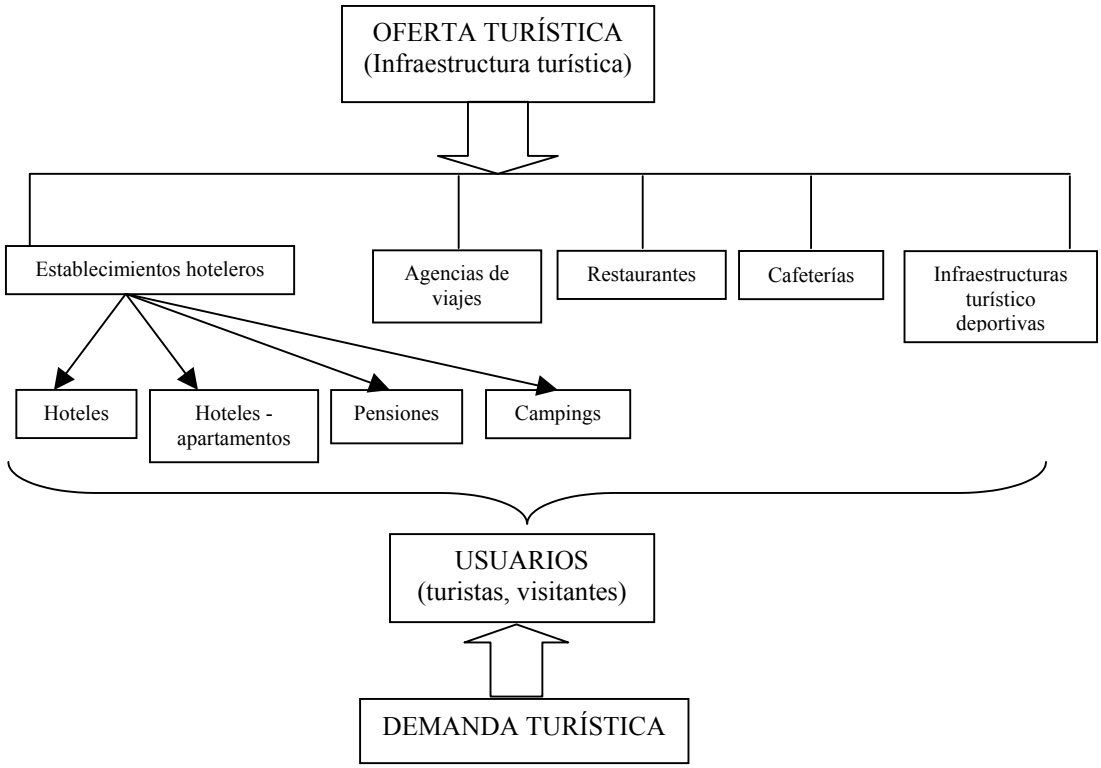
El MDS puede ser utilizado en muchas investigaciones junto a otras técnicas multivariantes, bien como una alternativa a dichas técnicas o bien como un complemento a las mismas. La utilización de cada una de ellas va a depender de los objetivos que se persigan en la investigación. Por tanto, no hay una técnica mejor que otra, sino que en algunos casos será más apropiado utilizar una técnica que en otros. Entre las ventajas de utilizar el MDS en comparación con otras técnicas multivariantes están:

- Los datos en MDS pueden estar medidos en cualquier escala, mientras que en el análisis factorial deben estar medidos en escala de razón o intervalo.
- El MDS proporciona soluciones para cada individuo, lo cual no es posible con el análisis factorial ni con el análisis cluster.
- En el MDS el investigador no necesita especificar cuáles son las variables a emplear en la comparación de objetos, algo que es fundamental en el análisis factorial y en el análisis cluster, con lo que se evita la influencia del investigador en el análisis.
- Las soluciones proporcionadas por MDS suelen ser de menor dimensionalidad que las proporcionadas por el análisis factorial (Schiffman, Reynolds y Young, 1981).
- En MDS pueden ser interpretados directamente las distancias entre todos los puntos, mientras que en el análisis de correspondencias solamente pueden ser interpretadas directamente las distancias entre filas o bien entre columnas.

5. APLICACIÓN DEL MDS AL SECTOR TURÍSTICO EN ANDALUCÍA.

No cabe duda de que el turismo es uno de los factores más importantes para Andalucía, constituyendo una de las fuentes de ingresos más importantes para la economía andaluza. Entre los elementos que forman parte del sistema turístico se encuentra la infraestructura turística, elemento que tiene gran importancia ya que de él depende en gran medida la capacidad de una zona para atraer los flujos turísticos. Mediante esta investigación se pretende analizar la infraestructura turística de Andalucía, con el objeto de identificar aquellas ciudades que sean más similares en relación a este aspecto, utilizando para ello el MDS. Los datos se han obtenido de la Encuesta de Coyuntura Turística de Andalucía (ECTA) y de la Encuesta de Ocupación Hotelera de la Junta de Andalucía, las cuales consideran que la infraestructura turística en Andalucía está formada por los siguientes elementos:

[1] El procedimiento de MDS implementado en SPSS es el programa ALSCAL (Alternating Least Squares SCALing), que fue desarrollado por Takane, Young y De Leew (1977) basándose en el algoritmo de mínimos cuadrados alternantes.



Partiendo de las dos encuestas señaladas anteriormente hemos obtenido para cada una de las provincias andaluzas los datos correspondientes al *número de establecimientos* y *número de plazas de hoteles*, *de hoteles-apartamentos*, *de pensiones*, *de camping, de agencias de viajes* (sólo número de establecimientos), *de restaurantes*, *de cafeterías* y *de infraestructuras rurales*[2]. Estos datos corresponden al año 2000.

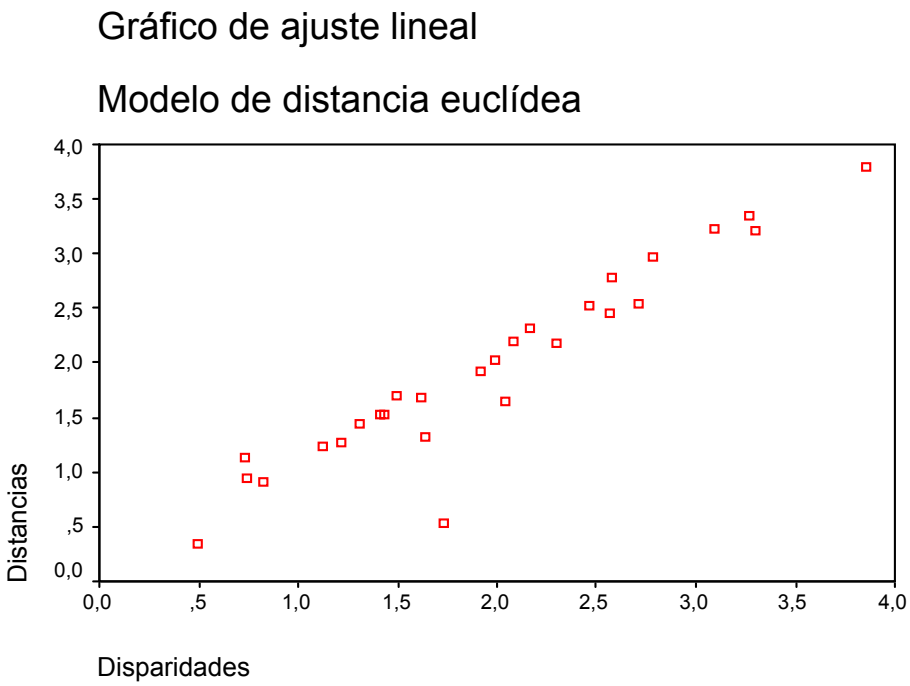
Con el fin de obtener una variable que nos informe de la capacidad turística en cada una de las provincias andaluzas se ha creado un ratio para cada una de las variables anteriores, excepto para la variable *número de agencias de viajes*, dividiendo el *número de plazas* entre el *número de establecimientos*. El siguiente paso ha sido obtener a partir de estos ratios una matriz de correlaciones entre ciudades. Finalmente, tenemos que hacer una última transformación de los datos, para convertirlos en distancias, a través de la fórmula de Coxon (1982):

$$d_{ij} = \sqrt{2(1 - r_{ij})}$$

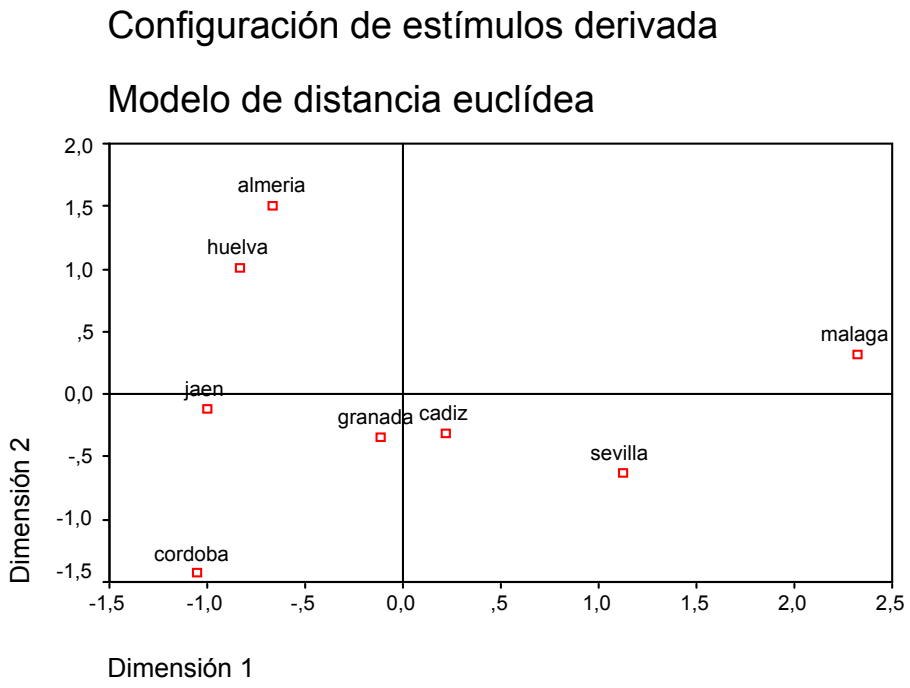
Esta matriz de distancias nos informa sobre las proximidades que existen entre las ciudades, en relación a la infraestructura turística. A partir de los datos obtenidos con la transformación de Coxon hemos aplicado un MDS, obteniéndose los siguientes resultados:

[2] La ECTA proporciona los datos referentes al número de establecimientos turísticos rurales y plazas de los mismos por provincias, en vez de los datos referentes a las infraestructuras turístico-deportivas.

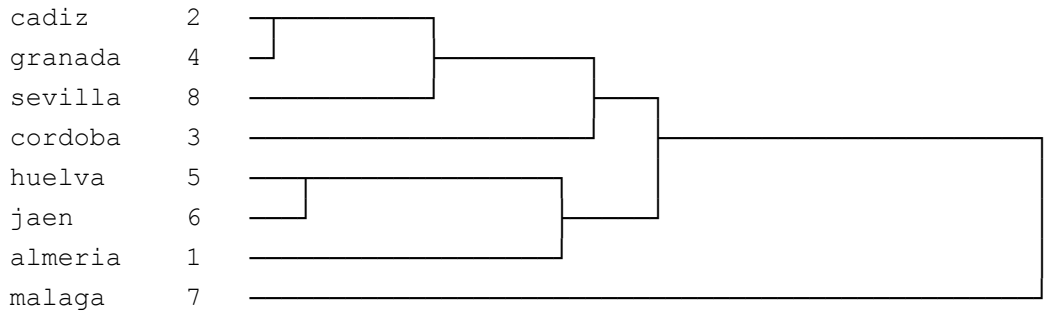
Los valores del *Stress* y del *RSQ* (0’13230 y 0’89424) nos indican que el ajuste de los datos es bueno. Un gráfico importante que nos informa si el modelo es adecuado o no es el gráfico de ajuste lineal. Si los datos se ajustan bien a una recta entonces el modelo es adecuado, ya que estamos suponiendo una relación lineal entre las distancias y las disparidades. En el gráfico podemos observar como los datos se ajustan bastante bien a una recta, por lo que el análisis es adecuado.



La configuración que se obtiene es la siguiente:



A partir de esta configuración podemos deducir que existen 4 agrupamientos de ciudades, referentes a su infraestructura turística. Por un lado está Málaga, por otro están Granada, Cádiz y Sevilla, por otro están Almería, Huelva y Jaén y por otro está Córdoba. Esto lo podemos corroborar a través de la aplicación de un Análisis Cluster a nuestros datos, obteniéndose los siguientes conglomerados:



Si seleccionamos cuatro conglomerados observaremos que se obtienen los mismos agrupamientos que hemos obtenido con el MDS. Así pues, el MDS puede ser una alternativa adecuada al Análisis Cluster.

Para la interpretación de las dos dimensiones obtenidas mediante el MDS podemos utilizar un Análisis Factorial, deduciéndose de ello que la dimensión 1 puede ser denominada como "servicios turísticos y capacidad de establecimientos hoteleros de prestigio" y que la dimensión 2 puede ser denominada como "capacidad de servicios de restauración y de establecimientos hoteleros económicos". Según la primera dimensión Málaga es la ciudad con más servicios turísticos y más capacidad de establecimientos hoteleros de prestigio, seguida de Sevilla, Cádiz y Granada y posteriormente del grupo formado por Almería, Huelva, Jaén y Córdoba. A partir de la segunda dimensión podemos deducir que Almería y Huelva son las ciudades con más capacidad de servicios de restauración y de establecimientos hoteleros económicos, seguidas de Málaga, Sevilla, Cádiz, Granada y Jaén, y en último lugar se encuentra Córdoba. Por tanto, el MDS puede servir como complemento a la interpretación de los datos en un Análisis Factorial.

6. CONCLUSIONES.

Con este trabajo se ha pretendido mostrar que la técnica de escalamiento multidimensional, a pesar de seguir siendo infrautilizada en muchas áreas, puede ser perfectamente utilizada en muchos casos, como alternativa a otras técnicas multivariantes o bien como complemento a las mismas. Para ello hemos visto las diferencias más importantes existentes entre el MDS y otras técnicas multivariantes como son el Análisis Factorial, el Análisis Cluster y el Análisis de Correspondencias.

A través del caso práctico realizado hemos visto que datos, que en un principio parece ser que están pensados para otro tipo de análisis, también pueden ser analizados a través de un escalamiento multidimensional.

BIBLIOGRAFÍA

- ARCE, C. (1993): *Escalamiento Multidimensional. Una* Técnica *Multivariante para el Análisis de Datos de Proximidad y Preferencia.* PPU, Barcelona.
- ARCE, C. (1994): *Técnicas de Construcción de Escalas Psicológicas*. Síntesis, Madrid.
- BORG, I. y GROENEN, P. (1997): *Modern Multidimensional Scaling*. Springer, New York.
- COXON, A. P. (1982): *The User's Guide to Multidimensional Scaling.* Heinemann Educational Books, London.
- GREEN, P. E. y CARMONE, F. J.(1969): *Multidimensional Scaling: An Introduction and Comparison of Nonmetric Unfolding Techniques*. Journal of Maketing Research, 6, 330-341.
- HAIR, J. F., ANDERSON R.E., TATHAM, R. L., BLACK, W. C. (1999): *Análisis Multivariante*. Prentice Hall, Madrid.
- KRUSKAL, J. B. (1964): *Nonmetric Multidimensional Scaling: A Numerical Method*. Psychometrika, 2, 115-129.
- LUQUE, T. (2000): *Técnicas de Análisis de Datos en Investigación de Mercados*. Pirámide, Madrid.
- REAL, J. E. (2001): *Escalamiento Multidimensional*. La Muralla, Madrid.
- SCHIFFMAN, S. S., REYNOLDS, M. L. y YOUNG, F. W. (1981): *Introduction to Multidimensional Scaling: Theory, Methods and Applications*. Academic Press, New York.
- SHEPARD, R. N. (1962): *The analysis of proximities: muldimensional scaling with an unknown distance function*. Psychometrika, 27, 125-140, 219-246.
- TAKANE, Y., YOUNG, F.W. y DE LEEW, J. (1977): *Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling feautures*. Psychometrika, 42, 7-67.
- TORGENSON, W. S. (1952): *Multidimensional Scaling: Theory and Method*. Psychometrika, 4, 401-419.
- YOUNG, G. y HOUSEHOLDER, A. S.(1938): *Discussion of a set of points in terms of their mutual distances*. Psychometrika, 3, 19-22.