PESQUIMAT , Revista de la F.C.M. de la
Universidad Nacional Mayor de San Marcos
Vol. XIII N°2, pp. 86-94, Lima - Perú. Diciembre 2010

EVALUACIÓN DE LAS PRINCIPALES CARACTERÍSTICAS HIDROGRÁFICAS Y FÍSICO QUÍMICAS RELACIONADAS A LA CALIDAD DEL MEDIO MARINO MEDINATE ESCALAMIENTO MULTIDIMENSIONAL

Ysabel Adriazola[1], *Ana María Cárdenas*[1], *Carlos Peche*[1] & *Victor García*[1]

Resumen. *El escalamiento multidimensional es un método estadístico de análisis de datos asociado a un conjunto de procedimientos que tienen como objetivo la representación de datos a través de una configuración de puntos cuando se conoce una determinada información sobre similaridades entre objetos. La historia de los métodos de escalamiento multidimensional comienzan con el trabajo de Torgeson en 1952, quien introdujo el término y esbozó las primeras ideas. En este trabajo se exponen los principales conceptos alrededor de este tema y se precisan sus aspectos conceptuales. Una aplicación a un problema respecto a las características hidrográficas y físico químicas relacionadas al medio marino peruano muestra alguna de las posibilidades del escalamiento multidimensional.*

Palabras clave: Escalamiento multidimensional, distancia euclideana, similaridades, característica hidrográficas, características físico-químicas.

EVALUATION OF THE MAJOR HYDROGRAPHIC AND PHYSICAL CHEMICAL CHARACTERISTICS RELATED TO THE QUALITY OF THE MARINE MULTIDIMENSIONAL SCALING

Abstract. *Multidimensional scaling is a statistical method for analyzing data associated with a set of procedures aimed at the representation of data through a configuration of points where certain information is known about similarities between objects. The history of multidimensional scaling methods begin with the work of Torgeson in 1952, who introduced the term and sketched the first ideas. In this paper we describe the main concepts around this issue and we need its conceptual aspects. An application to a problem regarding the physical and chemical hydrographic features related to the Peruvian marine samples of the possibilities of multidimensional scaling.*

Key Words. *Multidimensional scaling, Euclidean distance, similarities, Hydrographic, chemical and physical characteristics.*

[1]UNMSM, Facultad de Ciencias Matemáticas

1. Introducción

El escalamiento multidimensional, MDS (Multidimensional Scaling), es un método estadístico de análisis de datos que permite:

i) Obtener una estructura subyacente a los datos.

ii) Obtener una representación geométrica de los datos en un espacio de mínima dimensionalidad, de forma tal que sea accesible por simple inspección visual.

En el MSD cada objeto será representado como un punto en un espacio que es generalmente elucídelo (no necesariamente), y las disimilaridades entre los objetos, dij, serán representadas por las distancias, di, entre los puntos que representan esos mismos objetos de forma que se preserve la ordinalidad de los datos, es decir, se trata de obtener una configuración óptima, para lo cual se requiere definir una función para medir la desviación de la monotonicidad producida entre tas disimilaridades y las distancias. A esta función se le llama STRESS y el criterio utilizado para medir la desviación de la monotonicidad es en general el de mínimos cuadrados.

Shepard (1962) describe el MDS como un análisis orientado a reconstruir la configuración de un conjunto de puntos en un espacio métrico a partir de la información contenida en una matriz de datos, que puede estar conformada por los datos originales o por información cualitativa de los datos en base a la cual se obtiene una representación de los mismos que es estrictamente cuantitativa. Las pocas restricciones que se imponen a los datos es lo que otorga al MDS una gran aplicabilidad en diversas áreas. Así, este análisis ha sido ampliamente utilizado en investigaciones de mercado en el análisis de las preferencias de los consumidores, en segmentación de usuarios, evaluaciones de productos, etc., y, en general, en un amplio gama de investigaciones (Green y Canmone, 1970; Carrotl y Chang, 1974; Johnson, 1974).

Los axiomas en que se basan los modelos MDS son los de un espacio métrico, por ello los dij, la distancia entre los puntos i y j, cumplen los siguientes axiomas:

i) $d_{ij} \geq 0$, $\forall i,\ j$ Este axioma establece que la distancia entre dos puntos es no negativa.

ii) $d_{ij} = 0$, $i = j$ Establece que la distancia desde un punto a sí mismo es nula.

iii) $d_{ij} = d_{ij}$, $\forall i,\ j$ La distancia es una propiedad simétrica.

iv) $d_{ik} \leq d_{ij} + d_{ij}$, $\forall i,\ j$ Desigualdad triangular.

La distancia en que se basan los modelos MDS es una distancia Minkowski:

$$d_{ij} = \left[\sum_{\alpha}^{r} (x_{i\alpha} - x_{j\alpha}) \right]^{1/p} \tag{1.1}$$

donde:

i) d_{ij} es la distancia entre los puntos i y j.

ii) x_{ia} la es la coordenada del punto i en la dimensión α.

iii) x_{ja} la es la coordenada del punto j en la dimensión α.

iv) r = número de dimensiones y $p > l$.

A continuación se describirá el modelo métrico de Torgerson y el no métrico de Shepard-Kruscal. Torgerson (1952) supone que las disimilaridades están medidas en una escala de intervalos o de razón.

La disimilaridad, d_{ij} entre el estimulo i y el j será representada en un espacio multidimensional por la distancia euclídea, d_{ij} entre los puntos que representan a los objetos i y j, respectivamente, de forma que:

$$\delta_{ij} = f(d_{ij})$$

Donde:

$$d_{ij} = \left[\sum_{\alpha}^{r} (x_{i\alpha} - x_{j\alpha})^2\right]^{1/2} \tag{1.2}$$

es la distancia euclídea y f es una función lineal con pendiente positiva. En el caso más restrictivo se supone que $\delta_{ij} = d_{ij}$, Sea Δ la matriz de disimilaridades, δ_{ij}. A partir de ella obtenemos, Δ^* donde

$$\delta^*_{ij} = -\frac{1}{2}(\delta_{ij2}.\delta_{i2} - \delta_{ij2} + \delta_2) \tag{1.3}$$

Puesto que son los mismos objetos situados en las filas y en las columnas $i = j$. Δ^* es una matriz en la que la media de las filas y las columnas es cero. Torgerson demostró que si se satisface (1.3), entonces,

$$\delta^*_{ij} = \sum_{\alpha}^{r} x_{i\alpha}.x_{j\alpha} \tag{1.4}$$

es decir $\delta * ij$ es el producto escalar de los vectores que representan a los objetos i y j. Así,

$$\Delta^* = X.X^t \tag{1.5}$$

Puesto que Δ^* es simétrica, entonces es diagonalizable ortogonalmente y por tanto, es posible encontrar X, lo cual no es más que un problema de componentes principales.

Shepard (1962) propone un modelo MSD menos restrictivo que el de Torgerson, suponiendo que la función que relaciona las disimilaridades con las distancias espaciales es una función monótona. Así,

$$\delta_{ij} = f(d_{ij})$$

donde f es una función monótona creciente si la matriz es de disimilaridad, y decreciente si es de similaridad.

La solución que se obtiene, aún cuando solo se tiene en cuenta la relación de orden entre las disimilaridades, es en un espacio métrico único. Esto significa que dos matrices de disimilaridades distintas pero con la misma relación de orden entre sus elementos darán lugar a una misma solución, puesto que para ambas matrices se verificará:

i) $d_{ij} > d_{kj}$ $\delta_{ij} > \delta_{kj}$ (Criterio de monotonicidad fuerte)

ii) $d_{ij} > d_{kj}$ $\delta_{ij} \geq \delta_{kj}$ (Criterio de monotonicidad débil)

Kruskal se plantea la obtención de la solución como un problema de ajuste, a saber, tenemos unas disimilaridades y tratamos de encontrar una configuración de puntos tal que el ajuste entre las

disimilaridades y las distancias sea lo mejor posible. Para ello hay que definir un índice de ajuste, el STRESS, que en este caso no es de *bondad de ajuste* sino de error, definido como:

$$S = \frac{\sum_{ij}\left(d_{ij} - \widehat{d}_{ij}\right)^2}{\sum_{ij} d_{ij}^2} \tag{1.6}$$

donde los d_{ij} denominados *disparidades*, son los valores ajustados a las distancias y que están en un orden lo más similar posible al de los datos. La introducción de estos valores intermedios entre las disimilaridades y las distancias evita ejecutar operaciones aritméticas con las disimilaridades ya que se supone que éstas están en una escala ordinal. Knuskal introduce un método de ajuste denominado *regresión monótona.*

El procedimiento de cálculo de la solución final consiste en un proceso iterativo, que básicamente sigue los pasos siguientes:

i) Generar una configuración inicial de puntos en un espacio de dimensionalidad prefijada.

ii) Normalizar la configuración.

iii) Calcular las distancias entre cada par de puntos de esta configuración inicial.

iv) Realizar la regresión monótona. Las distancias son ajustadas mediante una función monótona obteniéndose las disparidades.

v) Obtener el STRESS entre las distancias y las disparidades.

vi) Si el STRESS es aceptablemente bajo, entonces esa configuración de puntos es la solución final.

vii) Si, por el contrario, el STRESS es alto entonces hay que mover la configuración de puntos en alguna dirección que conduzca a minimizar el STRESS. Esto se realiza calculando para cada punto la magnitud y la dirección del movimiento de acuerdo con,

$$X_{ia}^{t+1} = X_{ia}^{t} - K^t V^t$$

donde:

a) X_{ia}^t es la coordenada del punto i en la dimensión a en la iteración t.

b) K^t es un parámetro que expresa la magnitud óptima del movimiento.

$$\nu^t = \left[\frac{\alpha S}{\alpha X_{i\alpha}}\right]^t$$

es el gradiente de la función STRESS en la Iteración t.

viii) El proceso continúa como en ii).

2. PROBLEMAS FUNDAMENTALES QUE PUEDEN SURGIR UN ANÁLISIS MDS

i) La existencia de mínimos locales.

ii) La decisión acerca del número de dimensiones de la solución final.

iii) La interpretación de la solución.

2.1. Mínimos locales

Uno de los problemas que se puede presentar en el algoritmo de computación de la solución es que al minimizar la función STRESS el programa caiga en un mínimo local y pare las iteraciones antes de obtener una solución óptima. La mejor forma de evitar caer en mínimos locales es partir de una buena configuración inicial de puntos. Para ello algunos programas utilizan el análisis métrico de Torgerson para obtener la solución por componentes principales y tomarla como configuración inicial de puntos para realizar el análisis no métrico.

2.2. Número de dimensiones

La determinación del número correcto de dimensiones está relacionada al valor que tome el STRESS. Las funciones que se utilizan como STRESS, son las siguientes:

$$S_1 = \frac{\sum_{ij} \left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum_{ij} d_{ij}^2} \tag{1.7}$$

y

$$S_2 = \frac{\sum_{ij} \left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum_{ij} \left(d_{ij} - d_{**}\right)^2} \tag{1.8}$$

Kruskal (1964) sugiere que un valor del STRESS igual a 0.01 indica una solución *excelente*, entre 0.01 y 0,05 *buena*, entre 0.05 y 0.10 *pasable*, entre 0.10 y 0.15 *regular* y mayor de 0.15 indicada una solución *mala*.

Algunas veces pueden obtnerse soluciones degeneradas, esto es, cuando el número de puntos obtenidos es muy bajo comparándolo con el número de objetos que deben ser representados y por tanto aparecen las mismas coordenadas para diversos objetos. Una solución degenerada puede significar que la dimensionalidad del espacio es muy baja y que la configuración óptima de puntos debe buscarse en un espacio de mayor dimensionalidad. En general, el número correcto de dimensiones depende del número de objetos.

2.3. Interpretación de la solución

Es uno de los problemas con el que hay que enfrentarse cuando se realiza un MDS, no siempre es fácil interpretar la solución. En general se debe tener en cuenta que:

i) La ubicación de los ejes es arbitraria. Cualquier rotación de los ejes que ayude a dar interpretación a las dimensiones es válida siempre que esta rotación mantenga las distancias interpuntos. En general, los programas suelen rotar de forma que el primer eje abarque la mayor variabilidad.

ii) Si las dimensiones no tienen una interpretación clara, entonces es preferible ofrecer una interpretación localizando conglomerados de puntos. Se ha sugerido realizar un análisis de conglomerados jerárquico con la misma matriz de datos y buscar el grado de acuerdo entre los conglomerados obtenidos por este análisis y los conglomerados de puntos identificados a partir del MDS.

2.4. Métodos y resultados

Se aplicó el escalamiento multidimensional a fuentes de datos construidas en información publicada por IMARPE, que es la institución que genera y analiza información recolectada sobre la calidad ambiental del agua, sedimentos y organismos vivos, así como indicadores de contaminación por elementos o compuestos químicos orgánicos e inorgánicos para todo el litoral peruano.

Los lugares de muestreo lo constituyen diez estaciones de monitoreo de la Bahía del Callao con las siguientes características(Fuente-IMARPE):

i) La temperatura superficial, ha sido determinada a través de se registra con un termómetro de mercurio de balde y la del fondo con un termómetro de inversión kahlsico.

ii) El pH es determinado a través del método de potenciómetro por medio del EXTECH.

iii) Los nutrientes por medio del método de colorimétrico de STRICKLAND y PEARSONS. en diez puntos de monitoreo.

iv) El determinación del oxígeno disuelto a través del método titulométrico de WINKLER.

Se utilizará la fuente de datos DATE CALLAO SUPERFICIE2, para aplicar el MSD mediante el software estadístico SPSS.

La Figura 1, presenta el total de objetos incluidos en el análisis en base a los cuales se han obtenido 45 proximidades $\frac{n(n-1)}{2} = 45$.

Case Processing Summary		
Cases		0.0004
Sources		1
Objects		10
Proximities	Total Proximities	45^b
	Missing Proximities	0
	Active Proximities^a	45

donde

i) a: active proximities include all non-missing proximities.

ii) b: sum of all strictly lower triangular proximities.

La Figura 2, presenta el STRESS, la medida que nos indicará la adecuacidad de aplicar el MSD, en este caso el valor del STRESS es de 0.003, un valor muy próximo a cero, es decir se ha encontrado una solución excelente para el número de dimensiones que utilizará el MSD.

Stress and Fit Measures	
Normalized	0.0004
Stress I	0.193^a
Stress II	0.340^b
S Stress	0.0003^b
Dispersion Accounted For D.A.F	0.9996
Tucker's Coeficcient's of Congruence	0.9998

donde

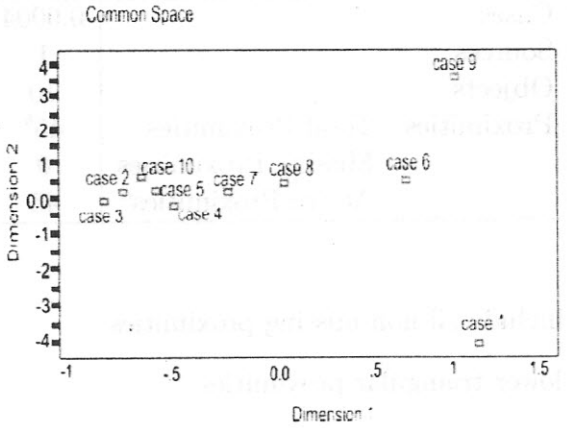
i) a: optimal scaling factor=1.000.

ii) b: optimal scaling factor=0.900.

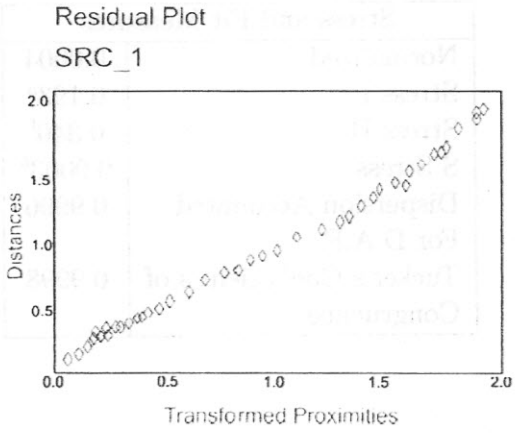
La Figura 3, presenta las dos dimensiones en el los objetos serán representados mediante el escalamiento multidimensional.

Common Space		
Final Coordinates		
	Dimension	
	1	2
Case 1	1.143	-0.0372
Case 2	-0.609	-0.031
Case 3	-0.603	-0.097
Case 4	-0.379	-0.043
Case 5	-0.435	-0.012
Case 6	0.696	0.082
Case 7	-0.175	0.043
Case 8	0.040	0.084
Case 9	0.964	0.299
Case 9	0.964	0.299
Case 10	-0.553	0.023

Figura 4. Representación del espacio común en que están representados los objetos utilizando el MDS.



La Figura 5, nos muestra que se verifica la relación lineal entre las distancias euclideanas y la proximidades



3. CONCLUSIONES

Puede observarse que en base a las características consideradas en el estudio: temperatura, oxígeno, pH, nitritos, nitratos y silicatos en la utilización del MDS, las zonas de monitoreo 2, 3, 4, 5, 8, y 10 son similares y estas con respecto a las zonas 1, 6, y 9 son disímiles. Puede observarse entonces una caracterización de las diez zonas de monitoreo en dos grupos.

En este proyecto se ha pretendido resaltar la importancia de utilizar métodos estadísticos multivariantes, en particular del escalamiento multidimensional, el cual permite analizar en este caso datos relacionados a las ciencias del mar.

Además puede señalarse que el MDS puede ser utilizado en muchas investigaciones conjuntamente con otros métodos multivariantes. Este trabajo también nos permite señalar que se deben realizar esfuerzos para llevar a cabo investigaciones multidisciplinarias en la que la estadística debe cumplir un rol importante.

REFERENCIAS BIBLIOGRÁFICAS

[1] Mardia K. V., Kent,J.T. & Bibby, J. M. (1979)., *Multivariate analysis.*, London: AcademicPress. Kruskal, J. B. (1964) Nonmetric muttidimensional scaling. Psychometnka, 29, 1-27.

[2] Shepard, R-N.(1962). *The analysis of proximities: Multidimensional scaling with unknown distance..* I y 11. Psychometrika, 27, 125- 140, 219-246.

[3] Torgenson, W. S. (1952).*Multidimensional scaling: Theoryand method.*, Psychometrika, 17, 401 -419.

[4] Bisauta Vinacua, (1998). *Análisis estadístico con SPSS para Windows.*, Estadística multivariante.