PESQUIMAT, Revista de la F.C.M. de la
Universidad Nacional Mayor de San Marcos
Vol. XIII N°2, pp. 95-104, Lima - Perú, Diciembre 2010

CONSTRUCCIÓN Y GESTIÓN DE ÍNDICES MULTIVARIANTES PARA CARACTERIZAR LA CALIDAD DEL MEDIO MARINO COSTERO PERUANO

Ysabel Adriazola1[1], *Ana María Cárdenas*[1], *Olga Solano*[1],
Carlos Peche[1], *Victor García*[1]

Resumen: La principal objetivo de este estudio es la de construir indicadores basados en métodos estadísticos multivariantes considerando las variables; temperatura, oxígeno, pH, sólidos suspendidos totales las que permitiran caracterizar la calidad del medio marino en particular de la Bahía de Ferrol. Debido a la complejidad de las características relacionadas a la calidad medio marino, se ha utilizado métodos estadísticos multivariantes con el objetivo de reconocer algún patrón o estructura explicativa y, además, plantear hipótesis para estudios posteriores. En base al Análisis Factorial se ha construido un índice de calidad del medio marino en particular se presentan los resultados obtenidos en la Bahía de Ferrol.

Palabras clave: Análisis factorial, índice, características contaminantes, características físico-químicas.

CONSTRUCTION AND MANAGEMENT OF MULTIVARIATE INDICES FOR CHARACTERIZING THE QUALITY OF THE MARINE COASTAL PERÚ

Abstract: The main objective of this study is to build indicators based on multivariate statistical methods considering the variables temperature, oxygen, pH, total suspended solids which to characterize the quality of the marine environment in particular the Bay of Ferrol. Due to the complexity of the quality characteristics related to marine environment, we used multivariate statistical methods in order to recognize a pattern or explanatory structure, and also hypotheses for further study. Based on Factor Analysis has built a quality rating of the marine environment in particular are the results obtained in the Bay of Ferrol.

Key words: Factor analysis, index, pollutant characteristics, physical-chemical characteristics.

[1]UNMSM, Facultad de Ciencias Matemáticas

1. Introducción La principal objetivo de este estudio es el de construir indicadores basados en métodos estadísticos multivariantes considerando las variables más significativas; temperatura, y las relacionadas a la calidad de agua; oxígeno, pH, sólidos suspendidos totales las que permitiran caracterizar la calidad del medio marino en particular de la Bahía de Ferrol. Para llevar a cabo este proyecto fue necesario construir fuentes de datos basadas en los datos monitoreados por el Instituto del Mar del Perú (IMARPE), ésta institución posee una sede central la Provincia Constitucional del Callao, donde están ubicados los más importantes laboratorios de investigación. Para este estudio se considerará la información en la Bahía de Ferrol (2000).

Para la construcción del índice de la calidad del medio marítimo, se ha utilizado el análisis factorial. La aplicación de este método permite identificar las dimensiones e indicadores más significativos, este método tiene la capacidad de sintetizar información. El objetivo general del análisis multivariante es el de analizar simultáneamente un importante número de variables observables medidas en un conjunto de unidades de análisis. En particular, el análisis factorial es un método que tiene como objetivo reducir un conjunto de p variables aleatorias (interrelacionadas), en un conjunto de f factores latentes (independientes), de tal forma que los f factores siempre serán, en número, inferior a las p variables inicialmente consideradas. Los factores reflejan la síntesis de la información redundante de las variables. Este método será exitoso si cumple con dos requisitos básicos (Bisquerra, R., 1989):

i) El principio de parsimonia, que establece que todo modelo debe ser más simple que los datos en los que se basa.

ii) El número de factores elegidos debe ser interpretable.

El antecedente del análisis factorial (AF) se encuentra en el análisis de regresión planteados por Galton. En 1901 Pearson, discípulo de Galton, presentó el método de componentes principales, previamente al cálculo del análisis factorial. Sin embargo, el origen de este método se atribuye a Sperman quien en 1904 diera a conocer su trabajo sobre inteligencia, es decir, originariamnte el análisis factorial se vincula con aplicaciones en el campo de la Psicología proporcionando un modelo explicativo matemático a las teorías de capacidad y comportamiento.

Pero quien popularizaría el método sería Thurstone a través de la aplicación del análisis factorial para identificar y diferenciar los principales factores que intervienen en la inteligencia humana. Posteriormente Kauser lo transcribiría a un modelo matemático, el modelo de rotación varimax, al que posteriormente le sucederían otros. Han sido distintos los investigadores los que han participado de una u otra forma en la modelación de lo que hoy conocemos como métodos factoriales. En esta trayectoria de de lo que conocemos como análisis factorial puede distinguirse dos clasificaciones, el análisis factorial exploratorio y el análisis factorial confirmatorio. El análisis factorial exploratorio tiene por objeto explorar la dimensionalidad latente sobre un conjunto de variables expresadas, a través de sus factores comunes, cuya estructura debe ser lo más simple posible. El AF confirmatorio se realiza con un conocimiento previa de la estructura de los factores.

El modelo matemático del AF es semejante al modelo de regresión múltiple. En este caso, los factores no son variables simples, sino dimensiones conformadas por un conjunto determinado de variables, las cuales serán explicadas linealmente en función de los factores seleccionados. En el AF, si los factores son obtenidos en base a las variables originales, cada variable será expresada como una combinación lineal de factores no observables directamente. Se consideran un conjunto

de variables aleatorias X1,X2,...Xp, que serán explicadas por un conjunto de factores comunes f_1, f_2,...,f_m y p factores únicos u_1, u_2,...,u_p de acuerdo con el siguiente modelo factorial lineal,

$$\begin{cases} X_1 &= a_{11}f_1 + \cdots + a_{1m}f_m + d_1u_1 \\ X_2 &= a_{21}f_1 + \cdots + a_{2m}f_m + d_2u_2 \\ \vdots & \vdots \quad \vdots \\ X_p &= a_{p1}f_1 + \cdots + a_{pm}f_m + d_pu_p \end{cases}$$

En el modelo factorial se supone que los factores comunes y únicos están incorrelacionados y los m factores son en número inferior al número de variables. Los principios fundamentales del modelo lineal factorial lineal están dados por dos teoremas.

1.1 Teorema 1

i) La varianza total de una variable puede ser explicada en función de las varianzas independientes entre sí.

ii) La varianza no explicada corresponde a la varianza específica de la propia variable y a la varianza del error debido a errores aleatorios

$$S_i^2 = S_1^2 + S_2^2 + \cdots S_k^2 + S_u^2$$

consecuentemente

$$1 = a_{i1}^2 + a_{i2}^2 + \cdots a_{ik}^2 + a_{iu}^2$$

donde:

- $a_{i1}^2 + a_{i2}^2 + \cdots a_{ik}^2 = h_i^2$ es la proporción de varianza total explicada por los factores 1 a k.
- S_{im}^2 es la varianza no explicada, dada por la especificidad de la variable (E_i^2) y por las varianza del error (e_i^2).

$$1 = h_i^2 + S_{iu}^2 = h_i^2 + E_i^2 + e_i^2$$

1.1 Teorema 2

La proporción de la varianza total de una variable explicada por cada uno de los factores comunes puede ser expresada como un coeficiente de determinación (r^2). La raíz cuadrada de esta proporción es un coeficiente de correlación (saturación) entre la variable y el factor. De esta manera,

$$1 = a_{i1}^2 + a_{i2}^2 + \cdots a_{ik}^2 + E_i^2 + e_i^2$$

donde:

$$a_{i1}^2 + a_{i2}^2 + \cdots a_{ik}^2$$

son los coeficientes de correlación de la variable i con los factores $1, 2, \cdots, k$. Luego las correlaciones estimadas entre los factores y las variables pueden ser utilizadas como estimación de las correlaciones entre las variables. Es decir,

$$r_{xy} = \sum a_{xy}a_{yx}$$

La solidez matemática y computacional alcanzada por los métodos multivariantes ofrece una gran riqueza analítica garantizada a través de la realización de un análisis exploratorio de datos para tener conocimiento de cada una de las variables y comprobación de supuestos paramétricos, un análisis bivariante que permitirá conocer el nivel de relación entre las variables consideradas para llevar a cabo el AF.

El primer paso para llevar a cabo el AF es el análisis de la matriz de correlación, la que es identificada como una matriz de similitudes o proximidades. Para decidir si los datos se ajustan o no a un análisis factorial puede llevarse a cabo un test de esfericidad de Barlett o, el test de Kaiser-Meyer-Olkin (KMO). La significancia de estos test evidencian al AF como un método apropiado para interpretar la información contenida en las variables en estudio. Una vez decido por el AF se debe realizar el proceso de extracción de factores subyacentes a un conjunto inicial de variables. Los principales métodos para extraer los factores son los de máxima verosimilitud, mínimos cuadrados y componentes principales los más usados. El análisis de componentes principales se define como un método estadístico que permite transformar un conjunto de variables intercorrelacionada, en otro conjunto de variables no correlacionadas denominados factores. Sobre la base de la matriz de correlación el ACP obtendrá las las ecuaciones lineales que representam la transformación lineal de las variables originales en relación con las componentes resultantes. El primer factor (f_1) es la combinación que explica la mayor parte de la variabilidad de las variables. Obtenido este, sobre la variabilidad restante se elige el segundo factor principal (f_2) explica el máximo de variabilidad, f_1 y f_2 están incorrelacionados. Todo este proceso se basa en el indicador que recoge los autovalores de cada variable y puede ser interpretado como la variabilidad total explicada por el factor. Al trabajar con variables estandarizadas, la varianza total coincide con el número de variables consideradas en el estudio.

El MCP permite que el investigador pueda resolver el problema de la elección del número de los factores teniendo en cuenta lo siguiente,

i) Los factores son dispuestos jerárquicamente: de mayor a menor varianza total explicada.

ii) Seleccionar aquellos autovalores que superan la unidad, esto puede ser posible a pesar de su controversia teórica a través de la regla de Kaiser.

iii) Analizar el scree plot en el que los factores se ubican en el eje de las abcisas y los autovalores en el eje de las ordenadas. A través de este gráfico se distinguirán los factores con altos autovalores de aquellos con bajos autovalores, descartándose aquellos que estén situados por debajo del punto de inflexión.

De esta manera, cada una de las variables será expresada como una combinación lineal de los factores elegidos representada por la matriz factorial, esta matriz es una expresión de la matriz de correlaciones inicial en la que cada columna es un factor y las filas son las variables. Los f_{ij} son denominados los índices de correlación, estos son los pesos o cargas que indican el peso que cada variable asigna a cada factor. Cuando los pesos de las variables son altas es un indicador que esta variable se asocia perfectamente con el factor. Debe considerarse que la proporción de varianza explicada por el conjunto de factores comunes es medido a través de las comunalidades (h^2) calculada a partir de la matriz factorial y es el resultado de sumar el cuadrado de las ponderaciones factoriales de cada variable, es decir:

$$h_i^2 = a_{i1}^2 + a_{i2}^2 + \cdots a_{im}^2 \quad i = 1, 2, 3, ..., m$$

El rango de variación de las comunalidades es de 0 a 1, un valor próximo a 1 indica que la variable está siendo explicada por los factores de lo contrario valores próximos a 0 indican que los factores no explican la variabilidad de las variables.

La varianza total explicada resulta de sumar las comunalidades más el factor único:

$$1 = h_i^2 U_i$$

Debe además obtenerse la matriz de residuales con la finalidad de determinar si el AF se adecua a los datos, un criterio usado es el que considera que valores inferiores a 0.05 indican que el AF se adecua a los datos.

Con respecto a la rotación factorial, esta consiste en el proceso de hacer girar los ejes de las coordenadas, hasta que se aproximen a la nube de puntos de las variables, es decir a partir de sus pesos. La rotación factorial transforma la matriz factorial inicial en una matriz factorial rotada, la cual es una combinación lineal de la primera y por lo que el porcentaje de varianza explicada es la misma. El objetivo de la matriz factorial rotada es simplificar la interpretación factorial por lo que,

i) Debe tener pocos pesos altos y el resto deben estar próximos a cero.

ii) Una variable debe estar saturada en un solo factor.

iii) No debe existir factores con la misma distribución de pesos altos y bajos.

Una vez simplificada la matriz factorial a partir de la rotación de la misma, se interpreta los factores en función de las variables con las que se encuentra asociada, una de las principales críticas de este método es que débil debido a que no existe una única solución, puesto que las decisiones adoptadas están en función al modo de factorizar, al tipo de rotación escogido, al número de factores seleccionados, por lo que se sugiere (Biquerra Alzina):

i) Estudiar la composición de las saturaciones factoriales significativas de cada factor.

   a) A través de una representación gráfica de los ejes factoriales de dos en dos, la cual develará la estructura latente del factor, puesto que las variables saturadas de un factor aparecerán agrupadas.

   b) Ordenar las variables en función del peso de los factores sobre estas de tal manera que aparezcan agrupadas las variables con ponderaciones altas para el mismo factor.

   c) Eliminar las saturaciones bajas.

ii) Proponer un nombre a los factores. En esta etapa juega un rol importante el marco teórico en el que se sustenta el estudio, considerando además la experiencia del investigador.

2. Construcción del indicador para la evaluación de la calidad del medio marino basado en el análisis factorial El indicador es calculado en base a la combinación lineal de los factores elegidos para realizar el análisis y la raíz cuadrada de los respectivos autovalores.

2.1 Métodos y resultados

Se aplicó el AF a la fuente de datos construidas en base a la 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. Los lugares de muestreo lo constituyen 16 estaciones de monitoreo de la Bahía del Ferrol (2000) con las siguientes características(Fuente-IMARPE):

i) **Temperatura:** A nivel de fondo se ha presentado valores entre 15.3 y 16.2, ascendentes de norte a sur.

ii) **Oxígeno:** Los valores fueron muy bajos (inferiores a 0.30ml/l), registrándose anoxia en gran parte de las estaciones evaluadas; las concentraciones detectadas fluctuaron entre 0.14 y 1-15 ml/l, correspondiendo a la estación 15 el máximo valor.

iii) **Potencial de iones hidronio o pH:** Los valores fluctuaron entre 7.30 y 7.61, el menor valor se ubicó en la estación 1 y el máximo en la estación 11; predomina la isolinea de 7.5 al centro y sur de la bahía. Los valores de pH son normales, no detectándose alteraciones por la influencia de las descargas, principalmente de tipo orgánico.

iv) **Sulfuros:** Los valores fueron elevados tanto en superficie como en fondo, propio de aguas anóxicas, fluctuando entre 0.80 y 55.89 $ug - atH2S - S/l$. La marcada anoxia originada por el consumo de oxígeno utilizado en la descomposición de deshechos orgánicos, origina evidentemente la formación de gases tóxicos como sulfuros, cuyo valor excede al establecido por Ley General de Aguas (1983), que es de 0.002 mg/l o 0.0669 $ug - at$H2S$-S/1$.

v) **Sólidos totales de suspensión STS:** Los valores fluctuaron entre 11.5 y 312mg/l; correspondiendo la máxima concentración a la estación 10, la que se ubica frente a la zona de descarga pesquera. La distribución de estos sólidos registró una tendencia decreciente conforme se aleja de la costa.

El valor mínimo detectado se encuentra en la estación 15 en cuyas proximidades se encontraron flujos de mayor intensidad. Se utilizará la fuente de datos ferrol-contaminante.sav, para aplicar el AF mediante el software estadístico SPSS.

vi) A continuación, en el Cuadro N°1 se presentará la matriz de correlación de las variables consideradas en el estudio; temperatura (z_{tem}), pH (z_{pH}), oxígeno (z_{oxig}), sulfuros (z_{sulfu}), sólidos totales suspendidos (z_{sts}).

Cuadro N°1: Matriz de correlación de las variables consideradas en el estudio.

Correlation Matrix					
	Z_{tem}	z_{ph}	z_{oxig}	z_{sulfu}	z_{sts}
Z_{tem}		0,638	0,671	0,107	0,750
z_{ph}	0,638		0,189	0,167	0,321
z_{oxig}	0,671	o,189		0,013	0.069
z_{sulfu}	0,107	0,167	0,013		0,149
z_{sts}	0,750	0,321	0.069	0,149	

Puede observarse que las correlaciones mas altas se encuentran entre la temperatura y el oxígeno, así como entre el pH y la temperatura.

vii) Para determinar si el AF es aplicable a estos datos se procede a obtener la medida de adecuacidad, KMO y a realizar el test de Barlett, que evidencian que es posible proceder con el AF.

Cuadro N°2: Medida de adecuacidad KMO y el test de Barlett.

KMO and Bartlett's Test	
Kaiser Meyer Olkin Measure of Sampling Adequacy	0.606
Bartlett Test of Aprox Chi Square	21,343
Sphericity df	10
Sig.	0,019

El siguiente resultado presenta las comunalidades, que es el porcentaje de la varianza de una variable que contribuye a la correlación con las otras variables

Cuadro N°3: Comunalidades de las variables consideradas en el estudio.

Communalities		
	Initial	Extration
Z_{tem}	1,000	0,672
z_{ph}	1,000	0,419
z_{oxig}	1,000	0,762
z_{sulfu}	1,000	0,767
z_{sts}	1,000	0,922

La comunalidad más alta está asociada a la variable sólidos totales suspendidos.

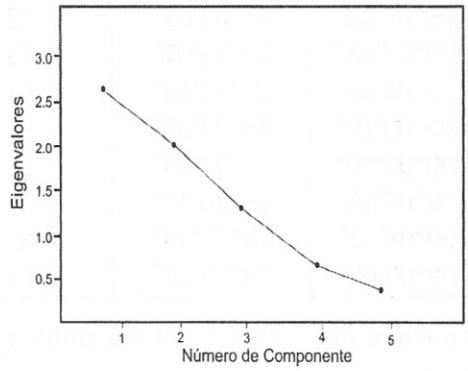
viii) El siguiente cuadro proporciona información relacionada a número de factores que serán considerados en el análisis, en este caso con los dos primeros factores explican el 70.83 % de la varianza total explicada.

Cuadro N°4: La siguiente figura muestra el scree plot, a través del cual se decide considerar solamente dos factores.

Componente	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,158	43,165	43,165	2,158	43,165	43,165	1,806	36,129	36.129
2	1,383	27,668	70,833	1,383	27,668	70,833	1,735	34,704	70,833
3	0,819	16,385	87,218						
4	0,487	9,737	96,955						
5	0,152	3,045	100,000						

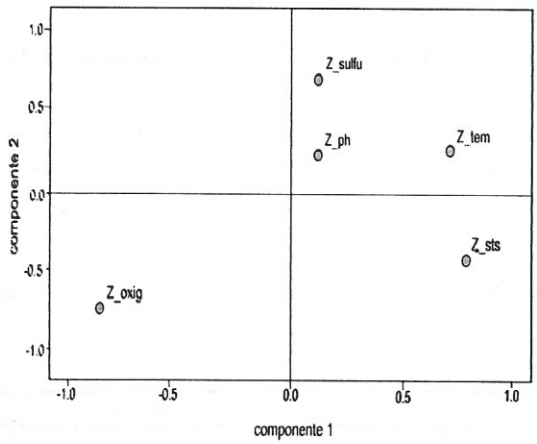
Autovalores >1 % de varianza explicada

En esta figura se puede observar tres factores no serán considerados en el análisis, debido a que están situados por debajo del punto de inflexión. El siguiente resultado muestra la importancia de las variables con los dos factores. La temperatura, el oxígeno y los sólidos totales suspendidos conforman un primer factor y el pH y los sulfuros un segundo factor. Resulta difícil designar un nombre, sin embargo puede observarse que el factor 1 está directamente relacionado con las variables temperatura y sólidos totales suspendidos e inversamente relacionado con el oxígeno. Asimismo, el factor 2 tiene asociadas a las variables pH y sulfuros, el el factor que evidencia el deterioro ambiental.



Rotated Component Matrix		
	Componentes	
	1	2
Z_{tem}	0,737	0,359
z_{ph}	0,071	0,643
z_{oxig}	$-0,720$	$-0,493$
z_{sulfu}	0,111	0,868
z_{sts}	0,853	$-0,442$

Represtación gráfica de las dos componentes principales



Obtención del índice sintético. Se obtendrá a través de la siguiente combinación lineal

$$I = factor1 * \sqrt{autovalor1} + factor2 * \sqrt{autovalor2}$$

A partir de la aplicación de este índice se jerarquizan los 16 puntos de monitoreo evaluados en la Bahía de Ferrol. Posición jerárquica de las estaciones monitoreadas en la Bahía de Ferrol

	POSICIÓN			
ESTACIÓN	LATITUD	LONGITUD	POSICIÓN	GRUPO
1	09°04′23″	78°36′36″	2	1
2	09°04′25″	78°36′50″	2	1
3	09°04′47″	78°37′08″	3	1
4	09°04′55″	78°36′38″	6	2
5	09°05′52″	78°36′38″	5	1
6	09°05′20″	78°35′37″	9	2
7	09°06′10″	78°34′40″	13	3
8	09°07′31″	78°36′02″	8	2
9	09°07′33″	78°34′52″	12	3
10	09°07′15″	78°34′07″	15	3
11	09°08′08″	78°34′13″	14	3
12	09°09′10″	78°34′37″	16	3
13	09°06′30″	78°35′40″	11	3
14	09°07′15″	78°36′47″	7	2
15	09°08′21″	78°37′00″	4	1
16	09°09′08″	78°35′54″	10	2

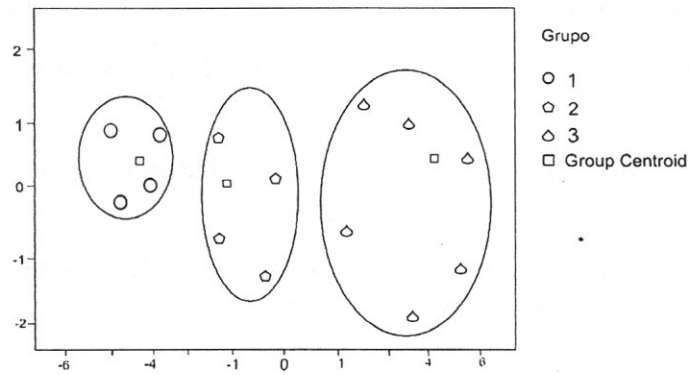
De acuerdo a la posición alcanzada por cada uno de los puntos de monitoreo, se puede realizar una clasificación en tres grupos:

i) 1 = estación con bajo deterioro ambiental (puntos de monitoreo que ocupan los 5 primeras posiciones).

ii) 2 = estación con mediano deterioro ambiental (puntos de monitoreo que ocupan la posición 6 hasta la posición 10).

iii) 3 = estación con alto deterioro ambiental (puntos de monitoreo que ocupan la posición 11 hasta la posición 10).

Gráficamente puede apreciarse que a través del AF permite una perfecta clasificación de los puntos de monitoreo de acuerdo a la variable grupo



El Análisis Factorial, ha permitido la construcción de un índice que permite observar como se configuran los puntos de monitoreo de la Bahía de Ferrol. Los puntos de monitoreo 1, 2, 3, 5 y 15 conformando un grupo con bajo deterioro ambiental, los puntos 4, 6, 8, 14 y 16 conformando un grupo con mediano deterioro ambiental y un tercer grupo conformado por los puntos de monitoreo 7, 9, 10, 11, 12 y 13. Además, puede señalarse que el AF puede ser utilizado en diversas áreas del conocimiento 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.

[2] Kruskal, J. B. (1964). Nonmetric muttidimensional scaling. Psychometnka, 29, 1-27.

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

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

[6] Bisauta Vinacua, (1998) Análisis estadístico con SPSS para Windows. Estadística multivariante. Fuentes de publicación-IMARPE.