UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

ANALISIS DE CORRESPONDENCIAS MULTIPLES COMO UNA TECNICA PARA EL ESTUDIO DE DATOS CUALITATIVOS

T E S I S

Que para obtener el
TITULO DE ACTUARIO
p r e s e n t a :

CLAUDIA LARA PEREZ SOTO

CIUDAD UNIVERSITARIA, 1990

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

INDICE

INTRODUCCION

Cuando las ciencias naturales o sociales cuentan con información de un determinado fenómeno, recurren a la estadística para tratar de describirlo en términos conocidos por medio de los datos, con el propósito de estudiar los resultados obtenidos.

En estadística existen métodos que se crearon con fines descriptivos. Muchas veces el fenómeno, sobre todo en ciencias sociales, puede ser descrito en términos de variables nominales o categóricas (cualitativas). Estas variables presentan problema cuando se trata de obtener algo a partir de ellas. Para ello existen técnicas especiales para este tipo de variables.

Muchos de los métodos para variables cualitativas están basados en supuestos probabilísticos para poder ser aplicados. Estos supuestos muchas veces no se cumplen y por lo tanto no tiene sentido aplicarlos en la realidad. Sin embargo existen metodos que no requieren de estas suposiciones como lo son todos aquellos que conforman el Análisis de Datos.

Para analizar la estructura de asociación entre variables cualitativas existen diferentes posibilidades como son las Medidas de Asociación que son índices que resumen la información contenida en una tabla de contingencia con el fin de evaluar la relación entre dos o más variables.

Otra alternativa son los Modelos Log-Lineales los cuales modelan la estructura de asociación entre variables para explicar el comportamiento del fenómeno estudiado.

Por otra parte está el Análisis de Correspondencias basado en el álgebra lineal que proporcionan representaciones gráficas de la estructura de asociación entre variables, así como escalamientos multidimensionales para los niveles de cada variable analizada.

Estos tres métodos son complementarios entre sí y resultan igualmente importantes en cuanto a su aplicación.

El propósito de este trabajo es explicar el Análisis de Correspondencias haciendo uso del álgebra lineal y de la geometría.

En el Capítulo I se introducen de una manera muy general conceptos básicos que son requeridos para poder entender el Análisis de Correspondencias Simples. Este análisis es planteado utilizando la Descomposición en Valores Singulares de una matríz.

En el Capítulo II se explica el Análisis de Correspondencias por medio de un ejemplo real para el caso de dos variables. Se dan elementos para la interpretación del Análisis de Correspondencias Simples y se plantea la necesidad de generalizarlo para el caso de Q - variables.

En el Capítulo III se construye toda la herramienta necesaria para el manejo de Q - variables cualitativas y se presenta el Análisis de

Correspondencias Multiples.

En el Capítulo IV se presenta un ejemplo practico del Análisis de Correspondencias Múltiples así como su interpretación.

El trabajo termina con las conclusiones del tema desarrollado tanto de la base practica como teorica, planteando así rasgos generales de dicha técnica.

CAPITULO I

VARIABLES Y ALGUNOS CONCEPTOS GEOMETRICOS EN EL ESPACIO MULTIDIMENSIONAL

Tipos de Variables y su Presentación

En el Análisis de Correspondencias y en general en los métodos multivariados, se trabaja con datos provenientes de un conjunto de individuos observados respecto a un conjunto de variables.

Todo fenómeno a investigar cuenta con determinadas características que se especifican en aquellas propiedades que son de interés para que a cada individuo se le clasifique o sea " medido " desde cierto punto de vista. Estas características particulares se les denomina variables.

Por individuo se entiende la entidad elemental sobre la cual se observa un cierto fenómeno a estudiar.

Una variable es una trasformación que va de un conjunto de individuos a un conjunto de valores. Sea V un subespacio del conjunto de valores, entonces:

i) Si V es un subespacio que está contenido en $\mathbb{R}$ entonces la variable X es cuantitativa (por ejemplo: estatura, peso, etc).

ii) Si V solo tiene estructura de orden, X

es ordinal, es decir todas aquellas que establecen cierta jerarquía (por ejemplo: grado de estudios, semestre que se cursa, etc.)

iii) Si V no tiene ninguna estructura en particular, X es nominal, es decir son aquellas a las que se les asigna un " nombre " o " etiqueta " (por ejemplo: estado civil, sexo, etc.)

El conjunto de valores que toman las variables con respecto a un fenómeno observado se les llama datos.

Existen muchas formas de presentar los datos pero solamente se mencionarán las que se utilizarán dentro de este trabajo. Estas son :

Matrices de Individuos por Variables. Es la representación matricial en donde los renglones serán correspondientes a los individuos y las columnas a las variables, es decir :

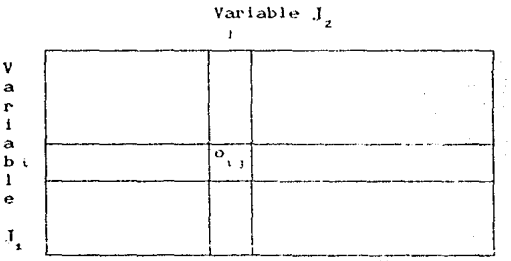
El elemento x_{ij} de la matriz X representa el valor que toma el individuo i con respecto a la variable j.

Variables

$$\begin{array}{c} \text{Individuos} \\ \left[\begin{array}{ccccc} x_{11} & \cdots & x_{1j} & \cdots & x_{1K} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{iK} \\ \vdots & & \vdots & & \vdots \\ x_{I1} & & x_{Ij} & & x_{IK} \end{array}\right] \end{array}$$

Tabla de Contingencia. Es una clasificación múltiple, esto es, dada una muestra de una población con

I individuos donde cada uno es descrito por un determinado tipo de atributos. Todos los individuos con la misma descripción son contados y clasificados de acuerdo a los criterios de interes, este valor pertenece a una celda de la tabla formando así una Tabla de Contingencia, construida generalmente con el fin de establecer las relaciones entre las variables. Para el caso de dos variables, los renglones de esta tabla corresponden a las categorías de la variable J_1 y las columnas a la variable J_2, entonces en la celda (i, j) se encontrará el número de individuos que son clasificados tanto en la categoría J_{1i} como a la categoría J_{2j}.



Matriz indicador. Una tabla de contingencia puede ser presentada de otra manera bajo las siguientes condiciones :

Si la variable J_1 tiene p categorías, entonces la matriz Z_1 asociada a la variable J_1 está dada como :

$$
Z_1 = \begin{array}{c} \\ 1 \\ \\ \\ \\ i \\ \\ \\ I \end{array}
\begin{array}{c}
\begin{array}{cccccc} 1 & 2 & & j & & p \end{array} \\
\begin{bmatrix}
1 & 0 & \cdots & 0 & \cdots & 0 \\
0 & 1 & & 0 & & 0 \\
\vdots & \vdots & & \vdots & & \vdots \\
0 & 1 & & 0 & & 0 \\
\vdots & \vdots & & \vdots & & \vdots \\
0 & 0 & & 1 & & 0
\end{bmatrix}
\end{array}
$$

donde

$$
z_{1_{ij}} = \begin{cases} 1 & \text{si el i-ésimo individuo pertenece a la categoría j} \\ 0 & \text{en otro caso} \end{cases}
$$

de manera análoga para la variable J_2 con q categorías, se obtiene la matriz Z_2. De tal forma que :

$$N = Z_1^T Z_2$$

es una matriz cuyas entradas son idénticas a las correspondientes a una tabla de contingencia. Cada matriz Z_1 y Z_2 genera una partición sobre el conjunto de individuos, lo que esta garantizando por el hecho de que en cada renglón exista solamente un uno.

Lo que justifica el uso de este tipo de matrices es el que permite asociar valores numéricos (0 y 1) a variables cualitativas, lo que hace posible aplicar ciertas técnicas que se verán más adelante.

Bases y dimensiones

Definición 1.- Un espacio lineal R se llama n - dimensional si en él se pueden encontrar n vectores

linealmente independientes.

Definición 2.- Toda colección de n vectores linealmente independientes de un espacio n - dimensional R se llama base de este espacio.

Como ejemplo se tiene lo siguiente:

En el segundo semestre de 1985, el grupo de profesores de Estadística llevó acabo una encuesta para la enseñanza de estadística descriptiva, con el manejo de datos reales. Algunas de las variables obtenidas de la encuesta son:

X_2.- Estatura X_3.- Peso

Se tomó la información de tres personas elegidas al azar X_a, X_b y X_c. Cada pareja de datos es expresada por un vector. De esta manera a cada individuo se le asocia un punto en el plano cartesiano o en $\mathbb{R}^2$ (Fig. 1), por lo que se tiene lo siguiente:

$$X_a = [\ 180 \quad 63\ ] \qquad X_b = [\ 172 \quad 60\ ] \qquad X_c = [\ 174 \quad 61\ ]$$

Por la definición 2, los vectores X_a, X_b y X_c pueden a su vez ser expresados como una combinación lineal de otros, por ejemplo utilizar los vectores e_1 e_2 donde:

$$e_1 = [\ 1 \quad 0\ ] \qquad\qquad e_2 = [\ 0 \quad 1\ ]$$

al conjunto $\{ e_1, e_2 \}$ se denomina base canónica. Cualquier vector en $\mathbb{R}^2$ puede ser expresado como combinación lineal de los vectores canónicos, ejemplo de ello es expresar a X_a como sigue:

REPRESENTACION GRAFICA DE VECTORES
Estatura vs. Peso

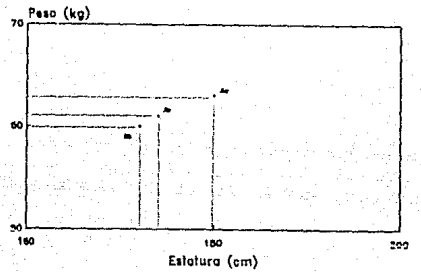


Fig. 1

$$X_a = 180\ e_1 + 63\ e_2$$

Los vectores canónicos tienen longitud[1] igual a 1 y son perpendiculares, si los vectores cumplen con las características se les llaman ortonormales[2].

Por otro lado, cualquier vector puede ser representado como combinación de otros vectores base, que no necesariamente son de norma uno y que además no son únicos. Un ejemplo de ellos es expresar a X_a de la siguiente manera:

$$X_a = 18\ b$$

---

1 Definida para cualquier vector $\mathbf{v} = [\ a,\ b\ ]$ como $\| \mathbf{v} \| = (\ a^2 + b^2\ )^{1/2}$. También denominada como norma de un vector.

2 STRANG. . pag. 77.

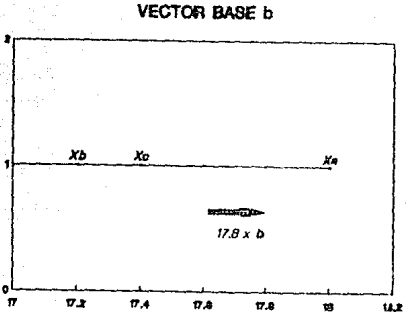


Fig. 2

donde $\mathbf{b} = [\ 10\ \ 3.5\ ]$. Al analizar la Fig. 1 se observa que los tres puntos se encuentran sobre una línea recta, por lo que pueden ser expresados como múltiplos del vector **b**, es decir, X_a , X_b y X_c como :

$$X_a = 18\ \mathbf{b}\ ,\ X_b = 17.2\ \mathbf{b}\ \ y\ \ X_c = 17.4\ \mathbf{b}$$

Entonces los puntos X_a, X_b y X_c pertenecen al subespacio de dimensión uno que está definido por el vector base $\mathbf{b}$[3]. De manera gráfica quedaría representado como en la Fig. 2.

Centroide

En general, otra manera de representar a un vector

3 Por definición de vector base. HOFFMAN, Kenneth. (1982). pags. 40 y 41.

como combinación lineal es el de un vector fijo más el múltiplo de un vector base. Este vector fijo se propone como el vector de medias o el centroide, el cual tiene en cada una de sus entradas el promedio aritmético de cada una de las entradas de los I vectores muestrales, es decir:

$$\bar{X} = (1 / I) [X_1 + X_2 + \ldots + X_I]$$

Los ejemplos comunes de ello son todos aquellos que se construyen en Regresión Lineal, tanto simple como múltiple porque siempre el " ajuste " de una línea recta estimada a un determinado conjunto de datos debe pasar por el vector de medias (Fig. 3)[4].

Al igual que en Regresión Lineal, un fenómeno no es descrito por dos variables unicamente sino que siempre se contempla la posibilidad de que sean más variables. Lo que hace pensar que no solamente se limita al espacio de dos dimensiones sino de 3 o más. Ahora bien, se plantea la siguiente pregunta:

¿ Se podrá encontrar en un espacio de J dimensiones, un subespacio de dimensión menor, tal que sea lo " más cercano posible " - en el sentido de minimizar la distancia - al conjunto de puntos representados en el espacio J ?.

Como se mostrará más adelante, la respuesta es afirmativa ya que para cualquier matriz puede encontrarse la Descomposición de Valores Singulares

---

4 JOHNSTON, J. pag. 19.

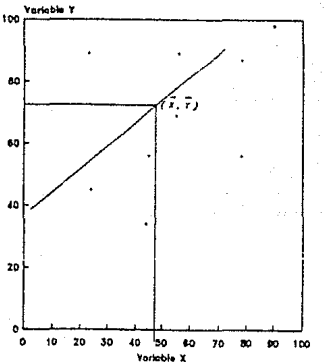


Fig. 3

(SVD)[3], lo que garantiza encontrar vectores base y además ortonormales. Pero antes, es necesario involucrar otros conceptos.

Distancia y Plano Euclidiano

El poder decir que dos observaciones son " parecidas " establece una noción de distancia. La definición euclidiana de distancia al cuadrado entre dos puntos es :

$$d^2 (X , Y) = \sum_i (X_i - Y_i)^2$$

Ahora, retomando el ejemplo anterior que involucraba las variables de estatura y peso, se graficó

---

3 NOBLE, Benjamin. (1977). pags. 323 a la 330.

como si fueran puntos con igual unidad de medida (Fig. 1), por lo que al tratar de interpretar la distancia calculada resulta difícil puesto que la diferencia en estatura en cm de dos personas suele ser mayor que la de peso en kg, por lo que la estatura contribuye más a la distancia. Al hacer un cambio de escala, es decir, que la estatura sea dada en metros en lugar de centímetros, el peso dominará . Por esto es necesario elegir una métrica que pondere cada variable de acuerdo a su variabilidad. Por esta razón se toma la definición de distancia entre dos observaciones como :

$$d^2 (X , Y)_P = \sum_i p_{ii} (X_i - Y_i)^2 \qquad (1)$$

$$= (X - Y)^T D_p (X - Y)$$

donde D_p es la matriz diagonal que determina la métrica del espacio J.

Para que la distancia esté bien definida, debe cumplir con lo siguiente :

Para todo X, Y y Z en el espacio de referencia :

i) $d (X, Y) \geq 0$

ii) $d (X, Y) = 0 \Leftrightarrow X = Y$

iii) $d (X, Y) = d (Y , X)$

iiii) $d (X, Y) \leq d (X, Z) + d (Z , Y)$

y la matriz D_p en (1) debe cumplir con las siguientes propiedades :

a) simétrica $p_{ij} = p_{ji}$

b) definida positiva :
$X^T P X = 0 \Leftrightarrow X = 0$ y $X^T P X > 0$ si $X \neq 0$

Cualquier matriz que cumpla con estas propiedades representa una métrica en un espacio determinado. Sea D_p una métrica en donde D_p es una matriz diagonal con las p_i's positivas:

$$D_p = \begin{bmatrix} p_1 & & & 0 \\ & p_2 & & \\ & & \ddots & \\ 0 & & & p_n \end{bmatrix}$$

De acuerdo con esta métrica y suponiendo que los vectores son perpendiculares, dado que se espera encontrar este tipo de vectores (por la DVS), el producto interno entre dos variables está dado por :

$$\langle X , Y \rangle_{D_p} = X^T D_p Y = \sum p_i X_i Y_i$$

y la norma o longitud como :

$$\| X \|_{D_p} = [\langle X , Y \rangle_{D_p}]^{\frac{1}{2}}$$

Continuando con el ejemplo, sea D_s la matriz de las desviaciones estándar de las variables, donde las desviaciones de las variables son :

$$DS_E = ((\sum (X_{i2} - \bar{X}_2)^2) / 1)^{\frac{1}{2}} = 3.40$$

$$DS_p = ((\sum (X_{i3} - \bar{X}_3)^2) / 1)^{\frac{1}{2}} = 1.25$$

Geométricamente, los vectores son graficados en unidades originales pero su producto escalar se calcula por medio de la expresión :

$$X_i^T (D_s^2)^{-1} X_j \quad i \neq j \qquad \begin{matrix} i = a, b, c \\ j = a, b, c \end{matrix}$$

donde $(D_s^2)^{-1}$ es la inversa de la matriz diagonal de varianzas, que es el factor de peso en el espacio euclidiano.

Como se mencionó antes, es ventajoso trabajar con bases ortonormales. Para el ejemplo, si se calcula la norma de los vectores de una base ortogonal no es 1. Para trabajar con una base ortonormal es necesario considerar que los vectores bases sean multiplicados (en el ejemplo), por la desviación estándar correspondiente, es decir:

$$3.40\ e_1 \qquad\qquad 1.25\ e_2$$

de tal forma que :

$$3.40\ e_1^T (D_s^2)^{-1}\ 3.40\ e_1 = 1$$

$$1.25\ e_2^T (D_s^2)^{-1}\ 1.25\ e_2 = 1$$

$$X_i = (X_{i2}/3.40)\ 3.40\ e_1 + (X_{i3}/1.25)\ 1.25\ e_2 \qquad i = a, b\ y\ c$$

En general se piensa que dado que se puede encontrar la DVS, la matriz debe construirse de tal manera que tambien involucre pesos que pondere cada variable, tal que se siga conservando la propiedad de ortonormalidad, al igual que DVS será explicado más adetante. Esto no solamente se realiza para $\mathbb{R}^2$, sino tambien para $\mathbb{R}^n$.

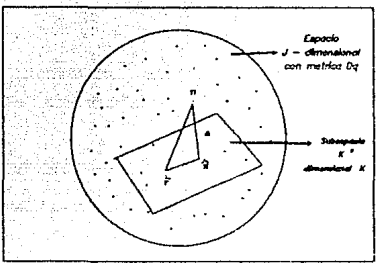


Fig. 4

Subespacios Optimos

No solamente se trata de encontrar los subespacios de menor dimensión, sino que además sea el más cercano al conjunto de puntos.

Se define cercanía entre el conjunto de puntos Y al subespacio S como :

$$\psi (S, Y_1, Y_2, \ldots, Y_I) \equiv \sum_{i=1}^{K} w_i \, d_i^2 \qquad (2)$$

donde d_i^2 es el cuadrado de la distancia mínima entre cada pundo de Y y S y las w_i's son los factores de ponderación, por lo tanto mayores que cero, de d_i^2 tales que $\sum_i w_i = 1$.

Para aclarar un poco, en la Fig. 4 se tiene el conjunto de puntos en el espacio euclidiano, incluyendo

un subespacio de menor dimensión ($\leq K$), cuya intersección con el espacio es no vacia , donde $\hat{Y}_i$ es el punto en el subespacio que es el más cercano a Y_i y de distancia ponderada mínima.

Proposición 1.- El centroide $\bar{Y}$ es el punto donde se minimiza ψ.

Demostración .-

Suponga que cualquier subespacio K del espacio euclidiano ponderado J no contiene a $\bar{Y}$. Sea S', el subespacio óptimo, un punto cercano a Y_i es Y'_i. Entonces la suma de las distancias cuadradas ponderadas de los puntos a S' son :

$$\psi (S', Y_1, Y_2, \ldots, Y_I) \equiv \sum_{i=1}^{K} w_i (Y_i - Y'_i)^T D_p (Y_i - Y'_i)$$

Sea $\bar{Y}'$ el punto más cercano a $\bar{Y}$ y sea $t = \bar{Y} - \bar{Y}'$ que es la traslación de $\bar{Y}'$ a $\bar{Y}$. Por último sea

$$\hat{Y}_i = Y'_i + t \quad \Rightarrow \quad \hat{Y}_i - Y'_i = t$$

Sumando y restando $\hat{Y}_i$ a $\psi (S', Y_1, Y_2, \ldots, Y_I)$

$$\sum_{i=1}^{K} w_i (Y_i - \hat{Y}_i + \hat{Y}_i - Y'_i)^T D_p (Y_i - \hat{Y}_i + \hat{Y}_i - Y'_i)$$

$$\psi (S', \underline{Y}) = \sum_{i=1}^{K} w_i (Y_i - \hat{Y}_i)^T D_p (Y_i - \hat{Y}_i) + \sum_{i=1}^{K} w_i (\hat{Y}_i - Y'_i)^T D_p (\hat{Y}_i - Y'_i) + 2 \sum_{i=1}^{K} w_i (Y_i - \hat{Y}_i)^T D_p (\hat{Y}_i - Y'_i)$$

Analizando cada término se tiene que:

$$\sum_{i=1}^{K} w_i \, (\hat{Y}_i - Y_i')^T D_p \, (\hat{Y}_i - Y_i') = t^T D_p \, t$$

$$\sum_{i=1}^{K} w_i \, (Y_i - \hat{Y}_i)^T D_p \, (\hat{Y}_i - Y_i') = 0$$

ya que

$$\sum_{i=1}^{K} w_i \, (Y_i - \hat{Y}_i) = \sum_{i=1}^{K} w_i \, Y_i - \sum_{i=1}^{K} w_i \, \hat{Y}_i$$

$$= \bar{Y} - \sum_{i=1}^{K} w_i \, (Y_i' + t) = \bar{Y} - \bar{Y}' - t = 0$$

por lo tanto

$$\psi (S', Y_1, Y_2, \ldots, Y_I) = \sum_{i=1}^{K} w_i \, (Y_i - \hat{Y}_i)^T D_q \, (Y_i - \hat{Y}_i) + t^T D_q \, t$$

$$= \psi (S, Y_1, Y_2, \ldots, Y_I) + \| t \|^2_{D_p}$$

Lo que muestra que

$$\psi (S', Y_1, Y_2, \ldots, Y_I) \geq \psi (S, Y_1, Y_2, \ldots, Y_I)$$

ya que $\| t \|^2_{D_p} \geq 0$, por lo tanto S es un subespacio óptimo ■

Las aproximaciones $\hat{Y}_i$ a Y_i serían de la forma:

$$\hat{Y}_i = \bar{Y} + \sum_{r=1}^{K} f_{ir} \, v_r$$

donde $v_1, v_2, \ldots, v_K$ son vectores base del subespacio S y f_{ir} constantes. Como se conoce a $\hat{Y}i$ y $\bar{Y}$ entonces solo falta determinar $v_1, v_2, \ldots, v_K$, que es lo que se hará a continuación.

Descomposición de Valores Singulares (DVS)[6]

El propósito, como se mencionó anteriormente, es encontrar, en un espacio de dimensión J, un subespacio de dimensión K < J que se aproxime lo mejor posible al conjunto dado de puntos en el espacio de dimensión J.

Para encontrar la solución que minimiza a la función de cercanía ψ de (2), para cualquier subespacio de dimensión K, se utilizará los conceptos de Descomposición de Valores Singulares (DVS) y la matriz de aproximación de bajo rango que se expondrá a continuación:

Sea $A_{I\times J}$ una matriz cualquiera de rango K (= Rango (A) - 1), por medio de la DVS A se puede descomponer como:

$$A_{I\times J} = U_{I\times K}\, D_{\alpha\ K\times K}\, V^T_{K\times J} = \sum_{r}^{K} \alpha_r u_r v_r^T$$

donde $U^TU = V\ V^T = I$ y D_α es la matriz diagonal con $\alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_K > 0$. La matriz ortonormal U con vectores columna $u_1, u_2, \ldots, u_k$ se denominan vectores singulares izquierdos, v son base ortonormal de las columnas de la matriz A. Similarmente, la matriz ortonormal V con vectores renglón $v_1, v_2, \ldots, v_k$ se

---

6 GREENACRE. (1984). Capítulo I.

llaman vectores singulares derechos, son base ortonormal de los renglones de A. Los elementos $\alpha_1, \alpha_2, \ldots, \alpha_k$ de D_α son los valores singulares de A.

Las matrices $F \equiv U D_\alpha$ y $G \equiv V D_\alpha$ contienen los renglones y columnas de A con respecto a sus vectores base de V y U.

Esta matriz A puede aproximarse por matrices de rango menor que K, es decir:

$$A_{(k^*)} = U_{(k^*)} D_{\alpha(k^*)} V_{(k^*)}$$

donde $U_{(k^*)}$, $D_{\alpha(k^*)}$ y $V_{(k^*)}$ son submatrices propias de U, D_α y V. Esta forma de aproximar es óptima bajo el criterio de :

$$\min \| A - X \|^2 = \min_X \sum_{i,j} (a_{ij} - x_{ij})^2$$

$$= \min \text{traza} \mid (A - X)^T (A - X) \mid$$

Cuando se toman en cuenta pesos se tiene que la DVS generalizada con respecto a las metricas :

Ω matriz de pesos por columna	Φ matriz de pesos por renglón

entonces $A_{(k^*)}$ es descompuesta como :

$$A_{(k^*)} = N_{(k^*)} D_{\mu(k^*)} M_{(k^*)}$$

donde $N^T \Omega N = M^T \Phi M = I$, los vectores $m_1, m_2, \ldots, m_{k^*}$ de $M_{(k^*)}$ definen las bases ortonormales del subespacio óptimo y las coordenadas de los vectores $Y_i - \bar{Y}$ con respecto a estas bases, son los renglones de:

$$F_{(k^*)} \equiv N_{(k^*)} \; D_{\mu(k^*)}$$

Las DVS (generalizada) de la solución requerida para cualquier dimensión K :

K = 1 Primer par de vectores singulares y primer valor singular proporciona solución óptima.

K = 2 Primero y Segundo par de vectores y valores singulares asociados proporcionan solución óptima.

y así sucesivamente. Esta " sucesión " de dimensiones conducen a los vectores base $m_1, m_2, \ldots, m_{k^*}$ a los ejes de los renglones de Y.

El cuadrado de los valores singulares dan una idea de como la matriz es bien representada a lo largo de los ejes . La variación total de A es cuantificada por su norma al cuadrado:

$$\| A \|^2_{\Phi \; \Omega} = \sum_{i=1}^{K} w_i \; a_i^T \; D_p \; a_i = \sum_{i=1}^{K} \mu_i^2$$

similarmente la variación de $A_{(k^*)}$

$$\| A_{(k^*)} \|^2_{\Phi \; \Omega} = \sum_{i=1}^{k^*} \mu_i^2$$

y la variación no explicada:

$$\| A - A_{(k^*)} \|^2_{\Phi \; \Omega} = \sum_{i=k^*+1}^{K} \mu_i^2$$

la cual ya está minimizada. La variación explicada de

$A_{(k^*)}$, expresada por el porciento τ_{k^*} de la variación total, es usada para cuantificar la calidad de la aproximación:

$$\tau_{k^*} \equiv 100 \left(\sum_{r}^{k^*} \alpha_r^2 \; / \; \sum_{r}^{K} \alpha_r^2 \right)$$

Cuando $a_i = Y_i - \bar{Y}$,la variación de $A_{(k^*)}$ es la suma de cuadrados de distancias ponderadas de los vectores Y_i a su centroide $\bar{Y}$ lo cual es la inercia o el total de la inercia del grupo de vectores. Conforme a la variación explicada y no explicada, el K - ésimo eje principal informa la cantidad μ_k^2 de la inercia total, que es descompuesta a lo largo de los ejes.

Variables Cualitativas

Todo lo anterior es tomando en cuenta una matriz de individuos por variables pero el propósito de este trabajo es aplicarlo a variables que sean de tipo cualitativo, en particular a una Tabla de Contingencia. Algunos de los conceptos antes mencionados serán modificados de tal manera que al hacer uso de las variables cualitativas, sean interpretados de la mejor manera posible.

Al igual que con los individuos, al definir la " cercanía " entre dos variables o p variables, es necesario elegir una métrica conveniente. De acuerdo con la métrica elegida, el producto interno entre dos variables está dado como :

$$\langle X_1, X_2 \rangle_{D_p} = X_1^T D_p X_2$$

la norma por:

$$\| X_i \|_{D_p} = (< X_i, X_i >_{D_p})^{\frac{1}{2}}$$

y la distancia como :

$$d^2 (X_1, X_2) = (X_1 - X_2)^T D_p (X_1 - X_2)$$

Si se calcula el coseno del ángulo entre estas variables tiene la siguiente propiedad :

$$\cos \theta = \frac{< X_1 , X_2 >_{D_p}}{(\| X_1 \|^2 \| X_2 \|^2)^{\frac{1}{2}}} =$$

$$\frac{\text{cov} (X_1 , X_2)}{(\text{var} (X_1) \text{ var} (X_2))^{\frac{1}{2}}} = \text{corr} (X_1 , X_2)$$

por lo que el *cos* θ puede ser interpretado como la correlación entre las variables X_1 y X_2.

En particular, si los datos son centrados con respectoa al vector de medias, tanto el producto interno como la norma pueden ser interpretados como la covarianza y varianza respectivamente, considerando a los pesos como probabilidades.

CAPITULO II

ILUSTRACION DEL ANALISIS DE CORRESPONDENCIAS SIMPLE

A fin de entender el Análisis de Correspondencias Simple, se explicará de manera general la relación que existe entre el este y la DVS, después mediante un ejemplo se explicará cada uno de los pasos que comprende el análisis. Esto se hará con la finalidad de motivar la generalización de dicho análisis, bajo ciertas condiciones que se verán más adelante.

Descomposición de Valores Singulares (DVS) y el Análisis de Correspondencias

Como se mencionó anteriormente el Análisis de Correspondencias es la representación gráfica de variables del espacio de dimensión J en un subespacio de dimensión K. Existen distintos enfoques del Análisis pero el que se utiliza en este trabajo es a partir de la DVS. Para ello se tiene lo siguiente:

Sea **N** la matriz de una la Tabla de Contingencia. La matriz de correspondencias se define como :

$$P = (1 / n_{..}) N$$

donde $n_{..} = 1^T N 1$. Los vectores de las sumas totales por renglón y columna son denotados y calculados de la siguiente manera :

$$r \equiv P1 \qquad c \equiv P^T 1$$

y sus correspondientes matrices diagonales denotadas como :

$$D_r = \text{diag}\ (\ r\ ) \qquad D_c = \text{diag}\ (\ c\ )$$

Las matrices de perfiles renglón y columna serán :

$$R \equiv D_r^{-1}\ P \qquad C \equiv D_c^{-1}\ P^T$$

Suponga que R es de rango K, entonces se pueden encontrar K vectores ortonormales tales que:

$$R = F\ B^T$$

donde $B = (\ b_1,\ b_2,\ \ldots\ ,\ b_k\ )$ son los ejes principales del perfil renglón y F sus coordenadas. De manera análoga :

$$C = G\ A^T$$

con $A = (\ a_1,\ a_2,\ \ldots\ ,\ a_k)$ que son los ejes principales del perfil columna y G son las coordenadas. Al pre-multiplicar a las expresiones anteriores por D_r y D_c respectivamente se tendrá lo siguiente:

$$D_r\ R = P \qquad D_c = P^T$$

con lo que encontrar A y B son problemas interrelacionados. La DVS da la pauta para encontrar estos dos conjuntos de vectores. Esta herramienta algebraica provee las matrices L, M, D_μ tales que :

$$P = L\ D_\mu\ M^T$$

tal que

$$L^T D_r^{-1} L = M^T D_c^{-1} M = I$$

las columnas de la matriz L son K vectores ortonormales
bajo la métrica D_r^{-1} y constituyen una base ortonormal para los renglones de P, mientras que para las columnas de M son K vectores ortonormales bajo la métrica D_c^{-1} siendo una base ortonormal para las columnas de la matriz P^T. Por último D_μ es la matriz diagonal de valores singulares tal que $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_k > 0$.

Como se mencionó M contiene a las columnas que representa una base de R. Entonces sea F las coordenadas respectivas para dicha matriz, es decir:

$$FM = R$$

como la matriz M esta compuesta por K vectores ortonormales bajo la métrica D_c^{-1} de $P = D_r R$ entonces:

$$F = R D_c^{-1} M$$

de manera similar se tiene que si G denota las coordenadas de los renglones de C respecto a L se tendria :

$$G = C D_r^{-1} L$$

donde F y G son las coordenadas de los perfiles renglón y columna en el subespecio de dimensión K. M y L representan los ejes principales respectivamente.

Con estos resultados se tendrán las representaciones gráficas de los perfiles renglón y

columna. Debe aclararse que estas representaciones están referidas a subespacios diferentes, pero como uno de los propósitos del Análisis de Correspondencias es la representación simultanea de ambos conjuntos, se muestran expresiones que permitirán establecer tal representación.

Las matrices F y G están relacionadas mediante la siguiente expresión :

$$F = D_r^{-1} P D_c^{-1} M$$

sustituyendo $P = L D_\mu M^T$ en la expresión anterior se tiene que:

$$F = D_r^{-1} L D_\mu M^T D_c^{-1} M = D_r^{-1} L D_\mu$$

de la misma manera para G se tiene :

$$G = D_c^{-1} M D_\mu$$

Como

$$F = D_r^{-1} P D_c^{-1} M = D_r^{-1} P G D_\mu^{-1} = R G D_\mu^{-1}$$

y para

$$G = C F D_\mu^{-1}$$

que son las llamadas Fórmulas de Transición que permiten hacer que las coordenadas del conjunto de perfiles se puedan expresar en función del otro conjunto. Estas fórmulas son las que permiten hacer la superposición de gráficas de los dos tipos de perfiles para la interpretación conjunta y por lo tanto del Análisis de

Correspondencias.

El ejemplo

El grupo de la Biología de Campo del semestre 88-2 junto con el Equipo de Buceo de la Facultad de Ciencias, realizó el siguiente trabajo: " ESTUDIOS BASICOS PARA EL ESTABLECIMIENTO DE RECOMENDACIONES DE CONSERVACION DE LOS SUSTRATOS BENTONICOS ARRECIFALES DEL PUERTO DE VERACRUZ, MEXICO ", eligiendo el arrecife denominado el " Cabezo "[7], el más grande de todo el sistema arrecifal veracruzano, para luego compararlo con los cercanos a él y tambien a los del Caribe mexicano. Este arrecife se dividió en tres zonas principales[8]:

1. Punta Valiente (PUV)
2. Centro (CEN)
3. Punta del Aguila (PUA)

cada zona se dividió en nueve subzonas por presentar distintas características para el desarrollo de diferentes especies:

1. Arrecife Frontal Exterior (AFE)
2. Arrecife Frontal Interior (AFI)
3. Transición Barlovento (TBA)
4. Rompiente Arrecifal (ROM)
5. Arrecife Posterior (POS)
6. Parches (PAR)
7. Transición Sotavento (TSO)
8. Cementerio de A.Cervicornis (CEM)
9. Platos de Hexacorales (PLA)

---

7 PADILLA, Claudia. (1989).

8 LARA, Mario. (1989).

	PUV	CEN	PUA	T.Renglón
AFE	11	11	14	36
AFI	13	14	9	36
TBA	4	4	2	10
ROM	4	5	0	9
POS	2	4	3	9
PAR	7	6	2	15
TSO	5	5	9	19
CEM	9	14	0	23
PLA	10	10	0	20
T.Columna	65	73	39	177

Tabla 1

Las tres regiones del arrecife en que se dividió para el muestreo son caracterizadas por el número de especies, que es una medida de diversidad.

Entre los puntos a comparar en el estudio, se eligió la abundancia de especies en cada zona y subzona para medir la riqueza específica del arrecife. Es decir, se quería observar en cuales zonas el desarrollo de diferentes especies es mayor. Como dentro de cada especie existen diferentes clases, se cuantifico para cada especie el número de clases diferentes. Los datos proporcionados para este trabajo fueron los de corales y se encuentran en la Tabla 1.

Análisis por Renglón (una dimensión)

Se calcularon las frecuencias relativas por cada renglón (Tabla 2). Se hicieron comparaciones sencillas en el sentido de analizar que zona permite el mayor desarrollo de diferentes tipos de coral, bajo ciertas

características (subzonas).

Se encontró que PUV y CEN, en la subzona PLA son las que permiten mayor desarrollo, pero al analizar el Total por Renglón de la Tabla 2, solamente están " descritas " en un 11.3 % con respecto a las demás subzonas. Entonces es necesario analizar desde otro punto de vista cómo es la asociación de las zonas y subzonas.

Si se piensa que cada subzona puede representarse como un punto en el espacio de 3 dimensiones (esto por ser 3 zonas), entonces puede encontrarse en vector base que trate de explicar el porque de dicha asociación. Para encontrar el vector base, se utiliza el concepto de DVS tomando en cuenta el peso de cada una de las variables. Entonces sean:

$$N = \begin{bmatrix} 11 & 11 & 14 \\ 13 & 14 & 9 \\ 4 & 4 & 2 \\ 4 & 5 & 0 \\ 2 & 4 & 3 \\ 7 & 6 & 2 \\ 5 & 5 & 9 \\ 9 & 14 & 0 \\ 10 & 10 & 0 \end{bmatrix} \qquad r = \begin{bmatrix} 0.203 \\ 0.203 \\ 0.056 \\ 0.051 \\ 0.051 \\ 0.085 \\ 0.107 \\ 0.130 \\ 0.113 \end{bmatrix}$$

$$c = \begin{bmatrix} 0.367 \\ 0.412 \\ 0.220 \end{bmatrix}$$

$$D_r = \text{diag}\ (\ r\ ) \qquad D_c = \text{diag}\ (\ c\ )$$

	PUV	CEN	PUA	T.Renglón
AFE	0.306	0.306	0.389	0.203
AFI	0.361	0.389	0.250	0.203
TBA	0.400	0.400	0.200	0.056
ROM	0.444	0.556	0	0.051
POS	0.222	0.444	0.333	0.051
PAR	0.467	0.400	0.133	0.085
TSO	0.263	0.263	0.474	0.107
CEM	0.391	0.609	0	0.130
PLA	0.500	0.500	0	0.113
T.Columna	0.367	0.412	0.220	1

Tabla 2

$$R = D_r N$$

en donde N es la matriz de datos observados, R es la matriz de perfíl por renglón (Tabla 2), c es el vector centroide, r es el vector de pesos por renglón, D_r es matriz diagonal con elementos iguales a r, y D_c es matriz diagonal con elementos igual a c.

Para poder calcular la DVS de R es necesario contar con una métrica que en este caso sera D_c^{-1}, para poder interpretar variables adecuadamente. Se calcularon las coordenadas en el subespacio de dos dimensiones que son obtenidas por:

$$F_{(2)} = \tilde{N}_{(2)} D_{\mu(2)}$$

donde $N_{(2)}$ y $D_{\mu(2)}$ son submatrices propias de la DVS generalizado para $R - 1c^T$, esto es:

$$R - 1c^T = N D_\mu M^T$$

donde

$$N^T D_r N = M^T D_c^{-1} M = I$$

y son las siguientes:

$$F_{(2)} = \begin{bmatrix} 0.409 & 0.013 \\ 0.073 & 0.013 \\ -0.045 & 0.056 \\ -0.534 & 0.024 \\ 0.258 & 0.226 \\ -0.200 & 0.147 \\ 0.613 & 0.007 \\ -0.542 & 0.144 \\ -0.526 & 0.102 \end{bmatrix}$$

Por la matriz de aproximación de bajo rango se puede graficar un solo eje, el de valor singular más grande.

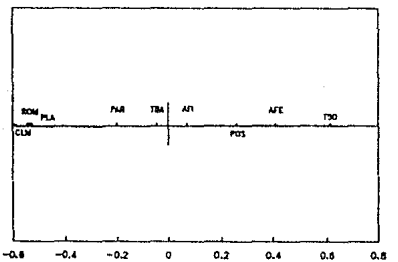
En la Gráfica 1 se pueden distinguir 3 grupos formados por:

1) ROM, CEM, PLA 2) PAR, TBA, AFI
3) POS, AFE, TSO

utilizando la Tabla 1 para explicar porque se situan así, se encontró que para el primer conjunto, en la columna de PUA no existe ningun tipo de coral por no estar definida en su totalidad la zona.

Para el segundo conjunto, junto con la Tabla 2, se observa que las frecuencias relativas son similares entre si, siendo más " pequeñas " en la columna de PUA.

CORRESPONDENCIAS EN UNA SOLA DIMENSION
Subzonas del arrecife "Cabezo"



Gráfica 1

Esto quiere decir que a pesar de estar definidas las subzonas, el desarrollo de PUA no es estable.

Por último para el tercer conjunto, se observa que las subzonas estan bien definidas en todas las subzonas y zonas (Tabla 1), que con respecto a sus frecuencias relativas (Tabla 2) y a su posición en el eje, entre mayor número de diferentes tipos de coral exista en PUA se aleja del orígen.

Análisis por Columna (una dimensión)

Una vez teniendo el análisis por renglones es lógico pensar en uno para columnas y observar las posibles relaciones que existan entre las zonas, es decir calcular la DVS para N^T.

Se calcularon tambien las frecuencias relativas por columna para la matriz N^T. Al analizar la Tabla 3 se

	AFE	AFI	TBA	ROM	POS	PAR	TSO	CEM	PLA	T.R
PUV	.169	.200	.062	.062	.031	.108	.077	.138	.154	.367
CEN	.151	.215	.055	.068	.062	.082	.068	.215	.137	.412
PUA	.359	.231	.051	0	.077	.051	.231	0	0	.220
T.C	.203	.203	.056	.051	.051	.085	.107	.130	.113	1

Tabla 3

puede observar que las mejores características para el desarrollo de corales son las que se encuentran en AFE situadas en PUA, pero la " descripción " de esta zona es la más " pequeña " de todas (total por región, Tabla 3). Nuevamente se trata de describir la estructura de las variables para analizarlas.

Al igual que en el análisis anterior, se necesita calcular nuevamente el centroide, la métrica y los pesos, entonces:

Sean :

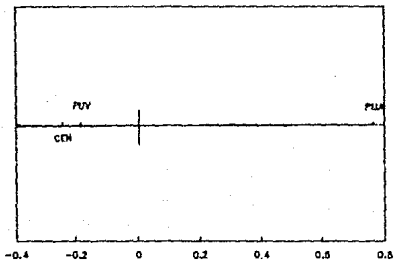
$$N^T = \begin{bmatrix} 11 & 13 & 4 & 4 & 2 & 7 & 5 & 9 & 10 \\ 11 & 14 & 4 & 5 & 4 & 6 & 5 & 14 & 10 \\ 14 & 9 & 2 & 0 & 3 & 2 & 9 & 0 & 0 \end{bmatrix}$$

$$C = \text{diag}\,(c)\, N^T$$

donde N^T es la matriz transpuesta de datos observados, C es la matriz de perfil por columna.

El centroide ahora será r, los pesos el vector c y la métrica es definida por D_r^{-1}. Así se tiene definido nuevamente todos los elementos para el cálculo de DVS

CORRESPONDENCIAS EN UNA SOLA DIMENSION
Zonas del arrecife "Cabezo"



Gráfica 2

para el perfil por columna.

Las coordenadas de los perfiles columna con respecto al subespacio de dos dimensiones son generadas por la matriz $G_{(2)}$:

$$G_{(2)} = \tilde{M}_{(2)} D_{\mu(2)}$$

donde $\tilde{M}_{(2)}$ y $D_{\mu(2)}$ son submatrices aproximadas de la DVS generalizada de $C - 1r^T$ tal que :

$$C - 1r^T = \tilde{M} D_{\mu} \tilde{N}^T$$

donde

$$\tilde{M} D_c \tilde{M}^T = \tilde{N} D_r^{-1} \tilde{N} = I$$

por lo tanto :

$$G_{(2)} = \begin{bmatrix} -0.183 & 0.114 \\ -0.246 & -0.095 \\ 0.765 & -0.012 \end{bmatrix}$$

Se utiliza nuevamente el concepto de matriz de aproximación de bajo rango y se toma en cuenta un solo eje, para graficar. En la Gráfica 2 se puede observar que CEN y PUV se encuentran muy cercanos entre si y ambos al orígen, mientras que PUA esta muy alejado de las anteriores. Con ayuda de la Tabla 3, la zona mejor descrita es CEN seguida por PUV, a grosso modo esto se debe a que entre ellas son " similares " en el sentido de que tienen todas las subzonas definidas. El desarrollo para corales se mantiene más o menos estable, por las cantidades encontradas en cada una de las subzonas. Como PUA no cuenta con un comportamiento estable, se separa completamente de las demás.

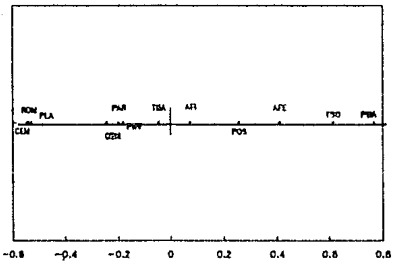
No se ha hablado de los valores singulares de cada análisis pero es importante saber que para ambos estos son iguales(0.1662 y 0.0086). La calidad de representación gráfica está dada en términos de la suma de cuadrados de valores singulares que es el total de la inercia. Como se tiene graficado unicamente el eje 1, correspondiente al valor singular mayor, este explica a las zonas y subzonas en un 95 % relativamente.

Análisis Conjunto (una dimensión)

Por medio de las Fórmulas de Trancisión, se pueden graficar ambos perfiles de manera conjunta, así las ventajas y desventajas serán contempladas en una sola gráfica.

El poder analizar la gráfica de manera conjunta es un problema visto desde dos puntos de vista distintos.

CORRESPONDENCIAS EN UNA SOLA DIMENSION
Zonas y Subzonas del arrecife "Cabezo"



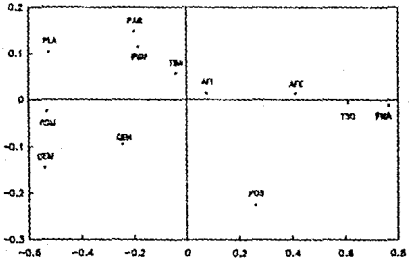
Gráfica 3

Entonces en la Gráfica 3 se observa lo siguiente:

PAR, TBA y AFI se situan cercanas a CEN y PUV mientras que POS, AFE y TSO se encuentran distribuidas a lo largo del eje 1 (lado positivo), con PUA en el extremo. Del lado contrario, como se mencionó anteriormente, CEM ROM y PLA se encuentran agrupadas. Debido a la situación que se presenta en la gráfica, el eje 1 es el que separa las zonas totalmente definidas, con las características predominates, de la que no lo esta.

El subespacio graficado es de una dimensión y su interpretación fue sencilla. Pero siempre se espera poder tener la mayor información posible contenida en el análisis. Como en este ejemplo se contempla un subespacio de dos dimensiones es posible contar con toda la información en una sola gráfica.

CORRESPONDENCIAS RENGLON Y COLUMNA
Zonas y Subzonas del arrecife "Cabezo"



Gráfica 4

Subespacio de Dos Dimensiones

Los puntos graficados en el subespacio de dos dimensiones se encuentra en la Gráfica 4.

Por renglón :

Al ser analizada esta gráfica, junto con la Tabla 2, se observó que las subzonas PLA, ROM y CEM son similares en el sentido de que no se encontró ninguna especie en la última columna. Por otro lado las subzonas PAR, TBA, AFI se encuentran alineados, esto se debe que la proporción de diferentes tipos de coral es similar. AFE y TSO se situan alrededor del eje 1 porque en la última columna se tiene mayor proporción y estan situados de manera ascendente. Sucede lo mismo con AFI, pero existiendo mayor relación entre los dos AFE y TSO porque ambos tienen la misma proporción en las dos primeras columnas. Por último POS se encuentra aislada

de las demás subzonas, que apesar de estar definida en todas las zonas proporcionalmente es " pequeña " (Total por renglón, Tabla 2) y poco representativa con respecto a la zonación.

Por columna:

Tomando en cuenta la Tabla 3 y la Gráfica 4 se puede observar que CEN y PUV se encuentran del lado izquierdo de la gráfica, como se mencionó anteriormente es porque ambas están caracterizadas por todas las subzonas lo que se considera como zonas bien definidas. Entre ellas la mejor es CEN. Por otra parte PUA se encuentra en el extremo derecho de la gráfica muy cercana al eje 1, lo que se puede observar es que el haber graficado ambos ejes no repercutio en esta zona.

Conjunto:

Analizando la Gráfica 4 de manera conjunta se puede ver como cuatro conjuntos. CEN es el más representativo porque a su alrededor se encuentran las subzonas ROM, CEM, AFI, y TBA, son las que permiten mayor desarrollo de tipos de coral (Tabla 1). PUV, al igual que CEN, a su alrededor se encuentran más cercanas a esta las subzonas AFI, TBA, PAR y PLA. Ahora, tanto en CEN y PUV comparten las subzonas AFI y TBA esto se debe a que existen cantidades similares en ambas zonas como se puede ver en la Tabla 1 y proporcionalmente son más o menos parecidas. PUA tiene a su alrededor a AFE y TSO, que son las que tiene mayor cantidad de tipos de corales en esta zona, principalmente TSO que en proporción es mayor que AFE y por lo tanto es la más cercana a PUA. POS es una subzona aislada sin que predomine en alguna zona determinada.

Existen 3 conjuntos de puntos que predominan principalmente en la gráfica. Todos vistos desde el punto de vista de zonas le corresponden que subzonas. Esto se debe a que además de existir un centroide para zonas y subzonas, existe un centro de atracción en cada conjunto al cual se le denomina baricentro, el cual determina el dominio de cada conjunto.

Habiendo analizado las subzonas y zonas, interesa saber que interpretación se le puede dar a los ejes. Entonces, el eje 1 (horizontal) puede ser visto como el que mide las frecuencias de los diferentes tipos de coral encontrados, es decir a mayor cantidad de tipos de coral estarán situados alrededor del origen y de lo contrario a lo largo del eje, dependiendo de la entrada de cada renglón. Al eje 2 (vertical) puede ser interpretado como el que separa las zonas y subzonas más representativas de las que no lo son.

La calidad de representación en esta gráfica es del 100 %, 95.1 % del eje 1 y 4.9 % del eje 2. lo cual era de esperarse porque a lo más esta matriz proporciona un subespecio de dos dimensiones[9].

Interpretación Biológica

El grupo de la Biología de Campo al analizar las gráficas dio la siguiente interpretación :

Esto se debe a la situación geográfica en la que se encuentra el Arrecife: en PUY las subzonas estan bien definidas porque es la parte del arrecife más protegida de la depositación de sedimentos terrigenos provenientes del continente, a pesar del choque de agua contra la

9 Ver DVS Capitulo I.

	PUV	CEN	PUA	ADF
AFE	11	11	14	9
AFI	13	14	9	11
TBA	4	4	2	4
ROM	4	5	0	6
POS	2	4	3	5
PAR	7	6	2	5
TSO	5	5	9	8
CEM	9	14	0	15
PLA	10	10	0	8

Tabla 5

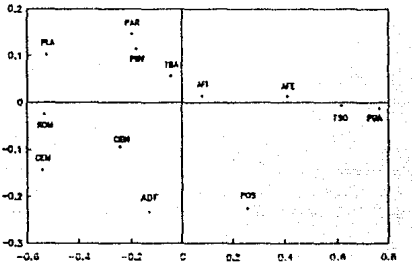
zona y por lo tanto el desarrollo de diferentes especies de corales es mayor. El caso extremo es PUA por la existencia de un canal y por el aporte de terrígenos que provienen de la desembocadura de Alvarado, las subzonas CEM y PLA no existen y en su lugar se encontraron bancos de arena. En CEN existe una situación similar a PUV, siendo afectado por el canal pero no en extremo, como en PUA , por lo que no existen muchas especies diferentes.

Variables Suplementarias

Al analizar las diferencias en el número de especies de coral encontrados en cada subzona del arrecife y uno de los propósitos es el de comparar al arrecife el " Cabezo " con los cercanos a él. Los datos con los que se contaban pertenecen al arrecife " Anegada de Afuera[10] " que en comparación con el " Cabezo " es menor en longuitud y en anchura, pero cuenta con las mismas subzonas y puede ser considerado, en proporción,

10 LARA, Mario. (1989).

CORRESPONDENCIAS CON V. SUPLEMENTARIA
Zonas y Subzonas



Gráfica 5

al tamaño de las zonas del arrecife. Los datos de "Anegada de Afuera (ADF) " se encuentran en la Tabla 5 junto con los datos originales.

Se espera que el análisis indique diferencias con respecto a la zona PUA por no encontrarse todas las subzonas definidas por la existencia del canal antes mencionado, y sea semejante ya sea a PUV o CEN.

Para comparar, se utiliza el concepto de variables suplementarias, no importando si es por renglón o por columna porque son tratados de manera similar. Estas variables suplementarias quedan representadas unicamente por su centriode, sin tomar participación en el análisis. La manera de calcular las coordenadas de esta variable es la siguiente:

$$f_{(2)_s} = c^T G_{(2)} D^{-1}_{\mu(2)}$$

La gráfica 5 contempla todo el análisis realizado anteriormente junto con la participación de la variable suplementaria, en este caso por columna.

Una vez teniendo la Gráfica 5, se encontró que los datos de la variable suplementaria " Anegada de Afuera (ADF) ", son semejantes a CEN que comparándolos con el análisis de diversidad, fuente de datos, y se muestra que las subzonas que caracterizan a la zona y al arrecife son las variables CEM y ROM, ya que en estas subzonas se encuentra el mayor número de especies de corales para la región profunda y somera respectivamente.

Todo el análisis anterior fue realizado para dos variables. Sin embargo siempre se plantea la necesidad de que un fenómeno no puede ser descrito unicamente con dos variables sino que es necesario el involucrar más variables y ver si de esta manera se puede interpretar de una forma más adecuada y cercana a la realidad.

CAPITULO III

ANALISIS DE CORRESPONDENCIAS MULTIPLES

En este capítulo se mostrará lo que constituye el Análisis de Correspondencias Múltiples, partiendo de dos variables para luego generalizarlo a Q variables. Esto se hará relacionando a la Tabla de Contingencia con la Matriz Indicadora. A partir de esa relación se hará todo el análisis para la Matríz Indicadora de Q variables.

El Análisis

El análisis se hará de manera similar a como se procedió en el Capítulo anterior, es decir :

i) Análisis por Columna
ii) Análisis por Renglón
iii) Análisis Conjunto

i) Análisis por Columna

Se había denotado el número de renglones y columnas para la Tabla de Contingencia N por J_1 y J_2 [11] respectivamente. Para la matriz indicadora Z se tienen I

---

11 Ver Capitulo I

renglones y ($J_1 + J_2$) columnas que particionadas quedan como:

$$Z = [\ Z_1 \quad Z_2\ ]$$

de tal forma que :

$$N = Z_1^T \quad Z_2$$

En el análisis para N se hace uso de pesos por renglon y columna, para Z los pesos correspondientes serán

$$r^z = 1/I \qquad y \qquad c^z = 1/2 \begin{bmatrix} r \\ c \end{bmatrix}$$

para la matriz c^z se divide entre 2 por estar contando dos veces a los individuos y a las variables. La matriz de correspondencias y sus matrices diagonales son definidas de la siguiente manera:

$$P^z = (1/2I)\ Z$$

$$D_r^z = (1/I)\ I$$

$$D_c^z = (\ 1/\ 2\ ) \begin{bmatrix} D_r & 0 \\ 0 & D_c \end{bmatrix}$$

donde P^z es tal que la suma de sus elementos es 1, D_r^z y D_c^z son las matrices diagonales correspondientes a los pesos de renglón y columna respectivamente para Z.

Como lo que se trata de hacer es mostrar la relación que existe entre el análisis de N y Z, se utilizarán una redefinición de las Fórmulas de

Transición. La razón de no utilizar las coordenadas tal como fueron calculadas es porque para N se tiene que las Fórmulas de Transición de renglones a columnas y de columnas a renglones son:

$$G\,D_{\mu} = C\,F \qquad\qquad F\,D_{\mu} = R\,G$$

respectivamente.

En particular para la transición de renglones a columna se tiene[12]:

$$D_c^{-1} P^T F = G\,D_{\mu}.$$

Pre - multiplicando por $D_r^{-1}\,P$ la expresión anterior queda

$$D_r^{-1}\,P\,D_c^{-1}\,P^T F = D_r^{-1}\,P\,G\,D_{\mu} \qquad (1)$$

$$F = D_r^{-1}\,P\,G\,D_{\mu}$$

pero como $F\,D_{\mu} = R\,G$ entonces

$$F = F\,D_{\mu}^2 \qquad (2)$$

De manera análoga para G se tiene que

$$G = G\,D_{\mu}^2$$

y como consecuencia :

$$F^T D_r F = D_{\lambda}\ (= D_{\mu}^2) \qquad\qquad G^T D_r G = D_{\lambda}\ (= D_{\mu}^2)$$

12 Recordar que $R = D_r^{-1}\,P$ y $C = D_c^{-1}\,P^T$.

Dichas coordenadas tienen cambios de escala que varían conforme a los valores singulares y para algunos resultados es necesario que tengan un comportamiento uniforme.

Para evitar estos cambios de escala, es conveniente mostrar la relación entre los dos análisis en términos de matrices coordenadas estándar[19] que se definen como :

$$\Phi = F\, D_{\mu}^{-1} \qquad\qquad \Gamma = G\, D_{\mu}^{-1}$$

donde las matrices Φ y Γ son las matrices coordenadas estándar de renglones y de columnas respectivamente, tal que :

$$\Phi^{T} D_{r} \Phi = I \qquad\qquad \Gamma^{T} D_{c} \Gamma = I.$$

Como se están analizando las columnas para Z, realizando las mismas operaciones que en (1) y tomando en cuenta (2), se tiene que:

$$(D_{c}^{-1} P^{T} D_{r}^{-1} P) \Gamma = \Gamma D_{\lambda}$$

donde las Fórmulas de Transición de coordenadas estándar son definidas de manera similar que para F y G, sustituyendo en (1) lo correspondiente a Z se tendrá:

$$\left[2 \left[\begin{array}{cc} D_{r}^{-1} & 0 \\ 0 & D_{c}^{-1} \end{array} \right] (1/2I) Z^{T} - (1/2I) Z \right] \Gamma^{Z} = \Gamma^{Z} D_{\lambda}^{Z}.$$

simplificando

---

18 GREENACRE (1984), pags. 93 y 94.

$$(2I)^{-1}\begin{bmatrix} D_r^{-1} & 0 \\ 0 & D_c^{-1} \end{bmatrix}\begin{bmatrix} Z_2^T Z_1 & Z_2^T Z_2 \\ Z_2^T Z_1 & Z_2^T Z_2 \end{bmatrix}\begin{bmatrix} \Gamma_1^z \\ \Gamma_2^z \end{bmatrix} = \begin{bmatrix} \Gamma_1^z \\ \Gamma_2^z \end{bmatrix} D_\lambda^z$$

Del Análisis de Correspondencias Simples sabe que :

$$P = (1/I)\ N = (1/I)\ Z_1^T Z_2$$

de donde

$$Z_1^T Z_1 = I\ D_r \quad y \quad Z_2^T Z_2 = I\ D_c$$

por lo que se tienen las siguientes ecuaciones:

$$\Gamma_1^z + D_r^{-1} P \Gamma_2^z = 2 \Gamma_1^z D_\lambda^z \qquad (3)$$

$$D_c^{-1} P^T \Gamma_1^z + \Gamma_2^z = 2 \Gamma_2^z D_\lambda^z \qquad (4)$$

pre - multiplicando a (3) por $D_c^{-1} P^T$ y sustituyendo $D_c^{-1} P^T \Gamma_2^z$ de (4) se tiene que :

$$D_c^{-1} P^T D_r^{-1} P \Gamma_2^z = \Gamma_2^z (2 D_\lambda^z - I)(2 D_\lambda^z - I) \qquad (5)$$

de manera similar se pre-multiplica a (4) por $D_r^{-1} P$ y usando de (3) $D_r^{-1} P \Gamma_1^z$ se obtiene que:

$$D_r^{-1} P D_c^{-1} P^T \Gamma_1^z = \Gamma_1^z (2 D_\lambda^z - I)(2 D_\lambda^z - I) \qquad (6)$$

Las ecuaciones (5) y (6) son matrices similares que en (2) por lo tanto las soluciones de N tambien son para estas ecuaciones :

$$(D_c^{-1} P^T D_r^{-1} P)\Gamma = \Gamma D_\lambda \qquad y \qquad (D_r^{-1} P D_c^{-1} P^T)\Phi = \Phi D_\lambda .$$

Como las coordenadas están sujetas a diferentes reescalamientos a lo largo de los ejes, los valores singulares estan relacionados. En particular para un cierto valor λ, se tiene que:

$$\lambda = (2\lambda^2 - 1)^2 \qquad (7)$$

$$\lambda^2 = (1/2)(1 \pm \lambda^{1/2}) \qquad (8)$$

Existe una solución trivial de las matrices solución que es $\lambda^{\circ}_{Z_1} = \lambda^{\circ}_{Z_2} = 1$ con $\lambda^{\circ} = 1$[14]. Esto es debido a que Z_1 y Z_2 son matrices tales que :

i) tiene exactamente un 1 en cada renglón
ii) La suma de los vectores columna es igual al vector 1.

Si $J_2 \leq J_1$ tal que J_2 es la dimensión de la matriz N, entonces se sabe que la solución trivial $\lambda = 1$. Para Z se tienen dos soluciones a través de (8), estas se dan si $\lambda = 0$ entonces $\lambda^2 = \pm 1$. por lo que se tienen $J_1 + J_2 - 2$ soluciones no triviales.

Análisis por Renglón

Para Z, los I perfiles renglón son vectores en el espacio de $(J_1 + J_2)$ dimensiones, como cada variable tiene sus respectivos niveles de clasificación, existen diferentes subgrupos de renglones que a su vez definen una columna de la matriz indicadora.

Geométricamente es imposible obtener puntos que se situen en medio de las categorías de respuestas y en

14 GREENACRE, (1984).

donde simultáneamente el grupo de centroides coincida con el punto columna correspondiente. Es por esta razón que se demostrará que las coordenadas calculadas para los puntos renglón son únicamente promedios de cada uno. Se utilizará la Fórmula de Transición de columnas a renglones por tener las primeras ya calculadas, entonces:

$$F = D_r^{-1} P G D_\mu^{-1} = R G D_\mu^{-1}$$

lo que implica que para Z :

$$F^z = R^z G^z (D_\mu^z)^{-1} = R^z \begin{bmatrix} G_1^z \\ G_2^z \end{bmatrix} (D_\mu^z)^{-1}$$

donde $R^z = I P^z = I (1/2I)Z = (1/2) Z$ por lo que el i - ésimo renglón de la matriz F^z es:

$$f_i^z = (D_\mu^z)^{-1} (1/2) (g_j^z + g_{j'}^z) \qquad (9).$$

Como G_1^z y G_2^z son respectivamente idénticas a F y G reescaladas[15] para N, y haciendo uso de nuevo de la Fórmula de Transición se tiene que:

$$G_1^z = R G_2^z D_\mu^{-1} \Rightarrow R G_2^z = G_1^z D_\mu \qquad (10)$$

donde R y D_μ^{-1} pertenecen al análisis de N.

El centroide $g_{(j)}$, para todos los j puntos, es el promedio de (9) de los i rangos sobre los renglones con respuestas (j , j') para $j' = 1, 2, \ldots, J_2$, donde j fija determina el promedio de los términos g_j^z. El promedio de g_j^z es el j - ésimo renglón de $R G_2^z$ en (10)

15 Esto es por el resultado mostrado anteriormente donde $F = F D_\lambda$ y $G = G D_\lambda$.

tal que :

$$\tilde{f}^{z}_{(j)} = (D^{z}_{\mu})^{-1} (1/2) (\varepsilon^{z}_{j} + D_{\mu} \varepsilon^{z}_{j})$$

→

$$\tilde{f}^{z}_{(j)} = (1/2) (D^{z}_{\mu})^{-1} (I + D_{\mu}) \varepsilon^{z}_{j}$$

como $D^{z}_{\lambda} = (D^{z}_{\mu})^{2}$, sustituyendo (6) se tiene que :

$$\tilde{f}^{z}_{(j)} = D^{z}_{\mu} \varepsilon^{z}_{j}$$

$$\therefore \quad \tilde{\varepsilon}^{z}_{j} = (D^{z}_{\mu})^{-1} f^{z}_{(j)}$$

que es la Fórmula de Transición de renglones a columnas.

Análisis Conjunto

En los resultados presentados se utiliza la relación que existe entre los análisis por renglón y columna de Z por medio de las Fórmulas de Tranción. En esta parte se dará un breve resumen de lo mostrado y algunos resultados que se consideran importantes para Q variables[16]. Para ello se tiene lo siguiente:

Se genera una matriz indicadora de Q variables

$$Z = [Z_{1} \quad Z_{2} \quad \ldots \quad Z_{Q}]$$

---

16 Valido tambien para dos variables.

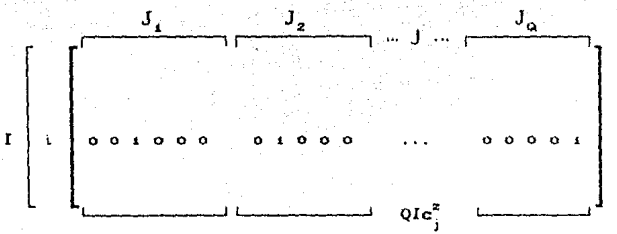


Fig. 1

con I renglones y $J = J_1 + J_2 + \dots + J_Q$ columnas, donde la q - ésima matriz de Z tiene J_q columnas. Se tienen QI unos dispersos en Z, I en cada submatriz Z_q y el resto son ceros.

Los elementos de cada renglón Z_q suman 1 mientras que la suma de todos los renglones de Z es Q (Fig. 1). La suma por columnas $1^T Z$ muestran la distribución marginal de respuestas sobre todas las categorías. Al igual que en los anteriores, el vector peso c^z de las columnas de Z es dado por:

$$c^z = (1/QI) Z^T 1$$

y por subgrupos es

$$c_q^z = (1/QI) Z_q^T 1$$

Se demostró que las coordenadas siguen siendo calculadas de la misma manera, la relación que existe entre los valores singulares de un análisis a otro y su generalización consiste unicamente en tomar en cuenta la actual matriz Z. Aún falta de demostrar otro tipo de resultados que son enumerados y demostrados al mismo

tiempo. Los resultados son los siguientes :

(a) La suma de los pesos de las columnas Z_q es $1/Q \ \forall \ q$.

Demostración .-

Partiendo de que $1^T Z_q 1 = I$ que es el número de unos de cada matriz Z_q, los pesos de las columnas de Z_q suman QI, entonces :

$$I / Q I = 1 / Q$$

Así cada variable discreta recibe el mismo peso, el cual es distribuido sobre todas las categorías de acuerdo a las frecuencias de respuesta ∎

(b) El centroide de los perfiles columna de Z_q es el centro de la gráfica que es el de todos los perfiles columna.

Demostración .-

Los pesos de las columnas de Z_q es

$$c_q^z = (1 / Q I) Z_q^T 1$$

y el centroide es vector de medias, entonces

$$Z_q \ c_q^z / 1^T c_q^z = ((1/QI) Z_q Z_q^T 1)/(1^T (1/QI) Z_q^T 1)$$

$$= (1/1^T Z_q^T 1) 1 = (1/I) 1$$

donde $Z_q Z_q^T = I$. Así cada subconjunto de categorías es balanceada en el centro ∎

(c) La inercia total de los perfiles columna y renglón es:

$$in(J) = J/Q - 1$$

(d) La inercia de los perfiles columna de Z_q es

$$in(J_q) = (J_q - 1) / Q$$

(e) La inercia de una categoría particular (j) es:

$$in(j) = 1/Q - c_j^2$$

Para demostrar estos tres últimos incisos primero se probará e) y como consecuencia d) y c).

Demostración .-

Para e) :

La inercia se define como la suma de las distancias al cuadrado del punto al centroide por el peso. Entonces para Z la distancia entre las columnas es proporcional porque el vector de pesos es constante. Para variables cualitativas se tiene unicamente un determinado rango de valores, en este caso solamente dos: 0 y 1, entonces la distancia será 0 ó (1 / QIc$_j^2$) en la j - ésima columna de Z. Al calcularla las distancias para los renglones se tiene que :

$$d_{ir}^2 = (0 - 1/1)^2 / (1/I)$$

y para las columnas

$$d_{jc}^2 = ((1/(QIc_j^2)) - (1/I))^2 / (1/I)$$
$$= (1/I) ((1/(Qc_j^2)) - 1)$$

la suma de las distancias es :

$$\Sigma\ d^2_{ir} = I - QIc^2_j\ (\ 1/I\ ) = 1 - Qc^2_j$$

y

$$\Sigma\ d^2_{jc} = QIc^2_j\ ((\ 1/(QIc^2_j\ ))\ -\ 1\ )^2.$$

Por lo tanto

$$\Sigma\ d^2_{ir} + \Sigma\ d^2_{jc} = 1/Qc^2_j\ (\ 1 - Qc^2_j\ )$$

entonces la inercia para una categoría en particular :

$$in\ (\ j\ ) = c^2_j\ (\ 1/Qc^2_j\ (\ 1 - Qc^2_j\ )) = (\ 1/Q - c^2_j\ )$$

con lo que se prueba e) ∎

Esto quiere decir que la inercia contribuída por la categoría aumenta las respuestas decreciendo esta categoría de 1/Q en 1/Q unidades.

La inercia de los perfiles renglón de Z_q es :

$$in\ (\ J_q\ ) = J_q\ (\ 1/Q\ ) - (\ 1/Q\ ) = (\ J_q - 1\ )$$

con lo que se prueba d) ∎

Por lo que la inercia explicada por la variable discreta aumenta linealmente con el número de categorías de respuesta.

Y por último, la inercia total esta dada por :

$$in\ (\ J\ ) = (\ J/Q\ ) - 1$$

por lo que se prueba c) ■

Continuando con los resultados se tiene :

(f) El número de dimensiones no triviales con inercia positiva es a lo más J - Q.

Demostración .-

Cada conjunto J_q de perfiles columna tiene el mismo centroide y el mismo subespacio de dimensionalidad menor o igual a $J_q - 1$. Entonces, para todo el conjunto J la dimensionalidad será menor o igual que $\sum_{q=1}^{J_Q} (J_q - 1) = J - Q$ ■

(g) El centroide del grupo de renglones con una respuesta en común y el punto columna representa la respuesta.

Demostración .-

Utilizando la Fórmula de Transición de columnas a renglones se tiene :

$$F^z = R^z G^z (D_\lambda)^{-1/2}$$

donde el i - ésimo perfil renglón en R^z es un vector de ceros y Q valores de 1/Q que indican las respuestas de $J_1, J_2, \ldots, J_Q$[17], entonces como f_i^z es el promedio

$$f_i^z = (1/Q)(g_{j1}^z + \ldots + g_{jQ}^z)$$

---

17 Por la definición de R^z.

de los puntos columna para la primera variable ($J_1 = J$), el vector es de la forma :

$$(1/Q) (\mathbf{e}_j^{zT} + j \text{ - ésimo renglón de } R_{12}\, G_2^z + \ldots + j \text{ - ésimo renglón de } R_{1Q}\, G_Q^z)$$

Al igual que los resultados anteriores, estos pueden extenderse al caso multivariado.

Matrices Indicador Multivariadas

Una vez hecho todo el análisis tanto de N como de Z para dos variables y viendo las relaciones que existen entre ellas, lo más conveniente es generalizar para Q variables. Cuando se involucran más de dos variables tanto la matriz indicadora como la Tabla de Contingencia presentan diferencias drásticas. Estas diferencias se deben a que precisamente por querer involucrar más de dos variables, se aumenta la dimensión del espacio. Como se ha venido mencionando, el propósito del Análisis de Correspondencias es que al encontrar un subespacio de dimensión menor, para que este pueda ser representado graficamente, se pierde información de las variables por lo tanto los porcentajes de la inercia en los ejes son bajos y la descripción no es del todo satisfactoria.

Matriz de Burt

La matriz simétrica $Z^T Z$ de $J \times J$, la cual es llamada la matriz de Burt y que tiene la siguiente estructura:

$$B_{JxJ} = \begin{bmatrix} Z_1^T Z_1 & Z_1^T Z_2 & \cdots & Z_1^T Z_Q \\ Z_2^T Z_1 & \cdots & \cdots & \\ \vdots & \vdots & & \vdots \\ Z_Q^T Z_1 & \cdots & \cdots & Z_Q^T Z_Q \end{bmatrix}.$$

Las submatrices $Z_q^T Z_{q'}$, con $q \neq q'$, son tablas de contingencia de dos variables q y q'. Cada submatriz de la diagonal $Z_q^T Z_q$ es la matriz diagonal de la suma de las columnas de la matriz Z_q la cual se denota como:

$$Z_q = Q \; I \; c_q^z$$

La matriz B es semidefinida positiva por lo que el análisis de correspondencias produce dos grupos idénticos de coordenadas para renglones y columnas[18] que son las mismas que para las columnas de Z. La prueba es la siguiente:

Por

$$(G^z \; R^z) \; \Gamma^z = \Gamma^z \; D_\lambda^z$$

donde $(\Gamma^z)^T \; D_c^z \; \Gamma^z = I$, $R^z = (1/Q)Z$ y $G^z = (Q \; I \; D_c^z)^{-1} \; Z^T$, como los pesos de las columnas de B son idénticos a los pesos de las columnas de Z, entonces la ecuación de valores propios anterior queda como:

---

18 Recordar que la Descomposición de Valores Singulares es la generalización de la Descomposición de Valores Propios. GREENACRE, (1984).

$$(Q^2 I D_c^B)^{-1} Z^T Z \Gamma^Z = \Gamma^Z D_\lambda^Z$$

lo cual es la Fórmula de Transición de B por lo que $\Gamma^Z = \Gamma^B$ y $D_\lambda^Z = (D_\lambda^B)^{1/2}$ lo que implica que para un valor λ^Z particular :

$$\lambda^B = (\lambda^Z)^2$$

y relacionada con N sería:

$$\lambda^B = (1/4) (1 \pm \lambda^{1/2})^2$$

El objetivo de la matriz indicadora Z es equivalente a la de la matriz de Burt y es el de ilustrar que estos análisis podrían ser descritos mejor en bivariante que multivariado[19].

La suma de renglones de cada matriz $Z_q^T Z_q$, para todo q y q', es el vector $Q I c^Z$ y sabiendo que la suma de renglones de $Z^T Z$ es $Q^2 I c^Z$ da la matriz R^B de los perfiles renglón, tiene la siguiente forma:

$$R^B = (1/Q) \begin{bmatrix} I & R_{12} & \cdots & R_{1Q} \\ R_{21} & \cdots & \cdots & \vdots \\ \vdots & & & \vdots \\ R_{Q1} & \cdots & \cdots & I \end{bmatrix}$$

donde $R_{qq'}$ es la matriz de perfiles renglón de la Tabla de Contingencia de dos variables $Z_q^T Z_{q'}$, existiendo solamentente una Transición de coordenadas columna a ellas mismas:

$$\Gamma^B = R^B \Gamma^B (D_\lambda^B)^{-1/2}$$

---

19 De hecho el paquete utilizado en el Capitulo IV trabaja bajo este concepto.

donde Γ^B puede ser particionada dentro de Q grupos de renglones Γ^B_q, q = 1, 2, ... , Q ; en tal caso:

$$\Gamma^B_q = (1/Q) \left(\Gamma^B_q + \sum_{q \neq q'} R_{qq'} \Gamma^B_{q'} \right) \left(D^B_\lambda \right)^{-1/2}$$

para q = 1, 2, ... , Q ; agrupando términos en Γ^B_q y recordando que $\Gamma^Z = \Gamma^B$ y $D^Z_\lambda = \left(D^B_\lambda \right)^{1/2}$ se tiene la siguiente expresión de coordenadas de las categorías de la variable q en términos de éstas para las otras variables en el análisis de correspondencias de Z:

$$\Gamma^Z_q \left(Q D^Z_\lambda - I \right) = \left(\sum_{q \neq q'} R_{qq'} \Gamma^Z_{q'} \right) \qquad (11)$$

Si Q = 2 se tienen las ecuaciones (3) y (4). Si Q > 2, debe ser equivalente a la matriz con dos variables.

Ahora, suponga que Q variables se pueden dividir en subgrupos de Q_1 y $Q_2 = Q - Q_1$ variables respectivamente, tal que cada subgrupo sea apareado en forma independiente uno del otro, y sin perdida de generalidad se tiene que :

$$Z^T_q Z_{q'} = I \, c_q \, c^T_{q'} \qquad \text{para } q, q' = 1, 2, \ldots, Q_1, \; q \neq q'$$

$$\text{y para } q, q' = Q_1 + 1, \ldots, Q, \; q \neq q'$$

donde c_q y $c_{q'}$ son los pesos renglón y columna de la tabla $Z_{qq'}$, teniendo que es una matriz de $\left(J_1 + \ldots + J_{Q_1} \right) \times \left(J_{Q_1+1} + \ldots + J_Q \right)$

$$\begin{bmatrix} Z_1^T Z_{Q_1+1} & \cdots & Z_1^T Z_Q \\ Z_2^T Z_{Q_1+1} & \cdots & \vdots \\ \vdots & & \vdots \\ Z_{Q_1}^T Z_{Q_1+1} & \cdots & Z_{Q_1}^T Z_Q \end{bmatrix} \qquad (12)$$

Las ecuaciones que definen las coordenadas columna de Z son iguales a (11) con los términos del lado derecho subdivididos en dos grupos. Si q' es el mismo grupo que q entonces $R_{qq'} = 1$ $c_{q'}^T$, lo que implica que $R_{qq'} \Gamma_{q'}^z = 0$ por ser el centroide de las columnas de $Z_{q'}$. Así por el lado derecho de (11) se encuentran variables solamente del otro grupo, resultando las Fórmulas de Transición entre coordenadas de cada grupo:

para $q = 1, \ldots, Q_1$:

$$\Gamma_q^z \left(Q D_\lambda^2 - I \right) = \left(\sum_{q'=Q_1+1}^{Q} R_{qq'} \Gamma_{q'}^z \right) \qquad (13)$$

para $q = Q_1+1, \ldots, Q$:

$$\Gamma_q^z \left(Q D_\lambda^2 - I \right) = \left(\sum_{q'=1}^{Q_1} R_{qq'} \Gamma_{q'}^z \right) \qquad (14)$$

factorizando (13) y (14) a $(1/Q_2)=(Q_1^{1/2}/Q_2^{1/2})(Q_1^{1/2}Q_2^{1/2})$ que a su vez es asociado con $D_\lambda^{1/2}$ tal que se tiene la siguiente relación :

$$Q_1^{1/2} Q_2^{1/2} D_\lambda^{1/2} = Q D_\lambda^2 - I$$

La relación entre las coordenadas estandarizadas es un poco más complicado de determinar, por ejemplo, los pesos asociados con los primeros Q_1 grupos de

columnas de Z suman Q_1/Q, donde los pesos asociados con los renglones de la matriz (12) suman 1. De (13) y (14) se sabe que :

para $q = 1, \ldots, Q_1$:

$$Q_2^{1/2}\ \Gamma_q = \beta\ \Gamma_q^z$$

donde β es un escalar constante. Para la estandarización de Γ_q y Γ_q^z →

$$\beta = (Q_1^{1/2} Q_2^{1/2} / Q^{1/2})$$

tal que

$$\Gamma_q = (Q_1/Q)^{1/2}\ \Gamma_q^z$$

y similarmente para $q = Q_1+1, \ldots, Q$;que son las coordenadas columna:

$$\Gamma_q = (Q_1/Q)^{1/2}\ \Gamma_q^z$$

CAPITULO IV

APLICACION DEL ANALISIS DE CORRESPONDENCIAS MULTIPLES

En el Capítulo III se mostró la teoría del Análisis de Correspondencias Múltiples, en éste se dará la siguiente aplicación:

Cierta Compañía de Productos Higiénicos que realiza su venta a tiendas de autoservicio, decidió hacer un estudio para encontrar el prototipo del " vendedor ideal " en cuanto a su productividad en ventas. Este estudio se hizo con el propósito de contratar personal con las características relevantes de dicho vendedor.

Para ello el departamento de personal, aplicó diversas pruebas a un número determinado de empleados, en las que se analizaron características físicas, intelectuales, psicológicas, de actitudes, etc. En base a las pruebas realizadas no se encontró evedencia alguna para determinar que características se se requería para garantizar una alta producción de ventas.

Se les propuso utilizar el Análisis de Correspondencias Múltiples para analizar si con ésta técnica se podría establecer una diferencia evidente

entre las características y con ello tratar de encontrar las que definen al " vendedor ideal ".

Se tomó una muestra de 45 personas y se decidió tomar en cuenta solo aquellas características que se creyeron necesarias para desempeñar el trabajo. Estas unicamente fueron 21, todas ellas de tipo cualitativo, la primera corresponde a los años de estudio, las 19 restantes son resultados de pruebas psicológicas y por último la que indica si pertenecen al grupo de alta o baja venta, de acuerdo a criterios establecidos por ellos mismos. Las variables o características estan dadas en la Tabla 1.

Variables	Etiqueta	Rango
Años de Estudio	ADE	3 a 16
Tarea 1	T1	2 a 12
Tarea 2	T2	1 a 15
Tarea 3	T3	4 a 13
Tarea 4	T4	1 a 12
Tarea 5	T5	2 a 13
Tarea 6	T6	1 a 17
Coef. Intelectual	CI	50 a 103
Meili	MEI	1 a 8
Moss	MOS	17 a 63
CMP -	CMP-	24 a 107
CMP +	CMP+	60 a 143
Actitudes	ACT	3 a 20
Vigor	VIG	1 a 13
Impulsivo	IMP	4 a 18
Dominante	DOM	3 a 20
Estable	EST	3 a 17
Sociable	SOC	8 a 19
Reflexivo	REF	4 a 15
Planeción	PLA	8 a 25
Grupo	GRU	1 a 2

Tabla 1

Para las primeras 20 variables se consideraron 3 clases y 2 para la última, esto se estableció así por

limitantes del paquete utilizado (STATICF. 70 variables a lo más), formando 62 clases en total. A cada una de la clases se les asignó un etiqueta empezando con las dos primeras letras de las variables y la clase correspondiente (1, 2 y 3), es decir ADE será identificado con respecto a sus clases como AD1, AD2 y AD3, a excepción de las tareas y CMP. Las variables de Tarea y CMP serán identificadas como sigue:

Las tareas con el número de tarea a identificar y su correspondiente clase como 01, 02 y 03 , por ejemplo las clases de Tarea 2 le corresponden las etiquetas 201, 202 y 203.

Para las CMP se tomará en cuenta la primera letra de la variable,el signo correspondiente y la clase, es decir las clases de CMP+ se identificarán como C+1, C+2 y C+3.

A los 45 individuos se les asignó el numero correspondiente a la entrega de sus exámenes al realizar las pruebas. Serán identificados con dicho número y un punto a su izquierda, dicha identificación se muestra en la Tabla 2 junto con las variables y con los resultados de las pruebas codificados en las clases correspondientes.

Es importante aclarar que la variable GRU será considerada como variable suplementaria porque como lo que interesa es encontrar las características ideales del vendedor, se decidió tomar el promedio como lo " ideal ".

Una vez establecido como serían identificados tanto los individuos como las variables y determinando

Variables / Individuos	ADE	T1	T2	T3	T4	T5	T6	CI	MEI	MOS	CMP-	CMP+	ACT	VIG	IMP	DOM	ESM	SOC	REF	PLA	GRU
.1	3	1	3	1	1	2	1	1	3	1	1	2	1	3	2	3	2	2	2	2	1
.2	3	2	2	2	1	2	2	1	3	3	3	3	3	3	2	1	1	1	2	1	1
.3	1	2	3	3	1	1	3	3	2	1	2	2	1	1	3	3	3	3	3	3	1
.4	2	1	2	2	3	3	2	2	3	2	3	1	3	2	1	3	2	3	2	2	1
.5	1	1	1	1	1	1	1	2	1	1	1	3	3	1	2	1	3	1	2	1	1
.6	3	1	2	2	1	1	2	2	2	2	1	1	1	2	2	1	1	1	1	2	1
.7	3	3	3	2	2	3	2	3	3	3	1	3	3	1	3	1	2	3	2	1	1
.8	3	3	3	2	1	1	2	3	1	1	3	3	1	1	2	1	2	3	2	1	1
.9	1	1	1	3	1	1	1	2	1	1	1	3	2	1	1	1	2	1	3	1	1
.10	3	1	2	2	1	2	1	2	2	3	3	1	2	1	1	2	2	2	3	2	1
.11	3	3	3	3	3	3	3	3	3	3	2	3	2	2	3	3	3	3	3	2	1
.12	3	2	3	3	3	2	3	3	3	3	2	3	1	3	3	2	1	3	1	2	1
.13	3	2	2	1	1	2	1	1	2	1	2	2	1	2	1	1	1	1	1	3	1
.14	2	1	2	2	2	3	3	2	3	2	2	1	1	3	3	1	1	1	3	3	1
.15	3	3	3	3	3	3	3	3	3	3	2	2	2	3	2	3	3	3	2	3	1
.16	2	3	3	3	3	3	3	3	3	3	1	1	3	2	3	3	2	3	1	3	1
.17	2	1	1	2	3	1	2	1	2	1	2	2	1	2	1	1	1	1	1	2	1
.18	1	2	2	3	3	3	2	3	2	1	3	2	1	3	2	2	2	2	2	1	1
.19	2	3	2	2	3	2	2	2	1	2	2	1	1	1	2	1	1	3	1	2	1
.20	3	3	3	2	3	3	3	3	3	2	2	3	1	2	1	1	3	3	1	2	1
.21	1	1	1	1	1	2	2	1	1	2	3	2	3	2	1	2	1	2	3	2	1
.22	3	3	3	3	3	3	2	3	3	2	3	2	3	3	3	3	1	3	1	3	1
.23	1	2	1	1	1	1	1	1	1	1	3	1	2	1	1	1	1	1	2	1	2
.24	2	2	1	3	2	3	2	2	2	2	1	2	3	1	3	1	2	1	2	3	2
.25	3	2	2	2	2	2	2	2	2	2	3	3	2	2	3	2	2	1	3	2	2
.26	3	3	3	3	3	3	3	3	3	3	3	2	1	3	3	3	3	3	2	2	2
.27	3	2	3	3	3	3	3	3	3	2	2	2	3	1	2	2	1	3	2	3	2
.28	1	2	2	2	2	2	1	1	3	3	3	1	2	1	1	2	3	2	1	2	2
.29	2	1	1	2	2	2	2	1	1	2	3	1	2	2	3	1	1	3	1	1	2
.30	1	1	1	1	1	2	2	1	1	1	1	3	1	1	1	2	2	1	2	2	2
.31	1	1	1	2	1	1	2	1	2	2	1	1	2	2	2	1	3	1	1	3	2
.32	3	1	1	1	1	1	1	1	1	1	1	1	2	3	2	2	3	2	2	3	2
.33	2	2	2	3	2	3	3	2	3	3	1	3	3	1	3	2	2	1	2	3	2
.34	1	3	3	3	2	2	3	3	3	3	2	3	2	3	1	1	1	2	1	1	2
.35	2	2	2	3	2	3	3	3	2	3	3	1	2	1	2	2	1	2	3	1	2
.36	1	1	1	1	2	1	1	1	2	2	3	1	2	2	3	3	2	2	3	3	2
.37	1	3	2	3	2	3	3	1	1	2	2	2	2	2	2	2	2	2	2	2	2
.38	2	1	1	1	1	1	1	1	2	3	1	2	3	3	3	3	3	3	2	1	2
.39	1	1	2	3	2	2	3	3	2	2	1	3	3	1	2	3	3	2	2	3	2
.40	2	1	1	2	1	1	1	1	1	1	3	3	2	1	2	3	1	2	3	2	2
.41	3	2	3	3	2	3	3	2	3	3	2	3	3	3	2	2	2	3	3	3	2
.42	2	2	3	3	2	3	2	2	3	1	2	1	1	2	2	3	2	1	2	1	2
.43	2	2	2	3	1	3	1	2	3	3	1	1	1	3	3	2	3	3	2	2	2
.44	2	1	1	3	2	2	3	2	2	3	1	3	3	2	1	2	1	3	1	3	2
.45	2	2	2	3	2	1	2	2	3	2	2	2	3	2	2	2	2	2	2	2	2

Tabla 2

quien sería la variable sumplemetaria, se procedió a realizar el Análisis de Correspondencias Múltiples por medio del paquete STATICF proporcionando las gráficas correspondientes a dicho análisis.

Al igual que en el análisis anterior (CAPITULO II), se analizarán 3 tipos de gráficas : por individuos (perfil renglón), por variables (perfil columna) y conjunta (perfiles renglón y columna).

Como se explicó anteriormente, el Análisis de Correspondencias Múltiples es una técnica para la representación gráfica de varios grupos de variables situados en un espacio de dimensión J para reducirla a un subespacio de dimensión K^*, esto es con la idea de analizar la estructura de asociación de las variables e individuos. En este caso el subespacio de dimensión K^* será de 3, esto es gracias a la matriz de aproximación de bajo rango y serán los que tengan la mayor representatividad con respecto a los valores singulares.

En general, las gráficas se analizarán junto con los datos de los individuos y variables dependiendo de la ubicanción dentro de las mismas.

Primera Gráfica

Por Individuos :

Para analizar a los individuos se tomó en cuenta las respuestas de estos conforme se encuentran en cada cuadrante (Tabla 3).

Al analizar el cuadrante *I* junto con la Tabla 3(*I*) se puede ver que la clase que predomina en la mayoría de

.2	3	2	2	2	1	2	2	1	3	3	3	3	3	3	2	1	1	1	2	1	1
.6	3	1	2	2	1	1	2	2	2	2	1	1	1	2	2	1	1	1	1	2	1
.10	3	1	2	2	1	2	1	2	2	3	3	1	2	1	1	2	2	2	3	2	1
.13	3	2	2	1	1	2	1	1	2	1	2	2	1	2	1	1	1	1	1	3	1
.17	2	1	1	2	3	1	2	1	2	1	2	2	1	2	1	1	1	1	1	2	1
.21	1	1	1	1	1	2	2	1	1	2	3	2	3	2	1	2	1	2	3	2	1
.24	2	2	1	3	2	3	2	2	2	2	1	2	3	1	3	1	2	1	2	3	2
.25	3	2	2	2	2	2	2	2	2	2	3	3	2	2	3	2	2	1	3	2	2
.28	1	2	2	2	2	2	1	1	3	3	3	1	2	1	1	2	3	2	1	2	2
.29	2	1	1	2	2	2	2	1	1	2	3	1	2	2	3	1	1	3	1	1	2
.31	1	1	1	2	1	1	2	1	2	2	1	1	2	2	2	1	3	1	1	3	2
.4	2	1	2	2	3	3	2	2	3	2	3	1	3	2	1	3	2	3	2	2	1
.7	3	3	3	2	2	3	2	3	3	3	1	3	3	1	3	1	2	1	1	1	1
.14	2	1	2	2	2	3	3	2	3	2	2	1	1	3	3	1	1	1	3	3	1
.19	2	3	2	2	3	2	2	2	1	2	2	1	1	1	2	1	1	3	1	2	1
.20	3	3	3	2	3	3	3	3	3	2	2	3	1	2	1	1	3	3	1	2	1
.33	2	2	2	3	2	3	3	2	3	3	1	2	2	1	2	3	3	1	2	3	2
.35	2	2	2	3	2	3	3	3	2	3	3	1	2	1	2	2	1	2	3	1	2
.37	1	3	2	3	2	3	3	1	1	2	1	2	2	2	2	2	2	2	2	2	2
.42	2	2	3	3	2	3	2	2	3	1	2	1	1	2	2	3	2	1	2	1	2
.44	2	1	1	3	2	2	3	2	2	3	1	3	3	2	1	2	1	3	1	3	2
.45	2	2	2	3	2	1	2	2	3	2	2	2	3	2	2	2	2	2	2	2	2
.3	1	2	3	3	1	1	3	3	2	1	2	2	1	1	3	3	3	3	3	3	1
.11	3	3	3	3	3	3	3	3	3	3	2	3	2	2	3	3	3	3	3	2	1
.12	3	2	3	3	3	2	3	3	3	3	2	3	1	3	3	2	1	3	1	2	1
.15	3	3	3	3	3	3	3	3	3	3	2	2	2	3	2	3	3	3	2	3	1
.16	2	3	3	3	3	3	3	3	3	3	1	1	3	2	3	3	2	3	1	3	1
.18	1	2	2	3	3	3	2	3	2	1	3	2	1	3	2	2	2	2	2	1	1
.22	3	3	3	3	3	3	2	3	3	2	3	3	3	3	3	1	3	1	3	3	1
.26	3	3	3	3	3	3	3	3	3	3	3	2	1	3	3	3	3	3	2	2	2
.27	3	2	3	3	3	3	3	3	3	2	2	2	3	1	2	2	1	3	2	3	2
.34	1	3	3	3	2	2	3	3	3	3	2	3	2	3	1	1	1	2	1	1	2
.39	1	1	2	3	2	2	3	3	2	2	1	3	3	1	2	3	3	2	2	3	2
.41	3	2	3	1	2	3	3	2	3	3	2	3	3	3	2	2	2	3	3	3	2
.43	2	2	2	3	1	3	1	2	3	3	1	1	1	3	3	2	3	3	2	2	2
.1	3	1	3	1	1	2	1	1	2	1	1	2	1	3	2	3	2	2	2	2	1
.5	1	1	1	1	1	1	1	2	1	1	1	3	3	1	2	1	3	1	2	1	1
.8	3	3	3	2	1	1	2	3	1	1	3	3	1	1	2	1	2	3	2	1	1
.9	1	1	1	3	1	1	1	2	1	1	1	3	2	1	1	1	2	1	3	1	1
.23	1	2	1	1	1	1	1	1	1	1	3	1	2	1	1	1	1	1	2	1	2
.30	1	1	1	1	1	2	2	1	1	1	1	3	1	1	1	2	2	1	2	2	2
.32	3	1	1	1	1	1	1	1	1	1	1	1	2	3	2	2	3	2	2	3	2
.36	1	1	1	1	2	1	1	1	2	2	3	1	2	2	3	3	2	2	3	3	2
.38	2	1	1	1	1	1	1	1	2	3	1	2	3	3	3	3	3	3	2	1	2
.40	2	1	1	2	1	1	1	1	1	1	3	3	2	1	2	3	1	2	3	2	2

3(*I*)

3(*II*)

3(*III*)

3(*IV*)

Tabla 3

las variables es la 2, seguida de 1. Los individuos entre más alejados del centro y cercanos a los ejes se encuentran, como lo son .21 y .25, predomina la clase 1 y 2 respectivamente. Los que se encuentran en medio como lo es .17 y .10 conservan la observación anterior pero no es tan marcada como en los individuos antes mencionados.

Para los individuos que se situan cercanos entre ellos, como lo son específicamente los individuos .6 y .29, al comparar las variables en las cuales coiciden (8 en total), se puede decir que son personas de características muy similares en el sentido que varian unicamente entre los niveles 1 y 2 y por eso se explica su cercanía (ver Gráfica 1).

En el cuadrante *II* (Tabla 3(*II*) y gráfica 1), la clase que sigue dominando es la 2 seguida ahora de 3 y que al igual que en el cuadrante *I*, los individuos alejados del centro y cercanos al los ejes, predomina alguna de las clases ya mencionadas, esto se puede observar en la gente con estiqueta .20 y .19. Al agrupar a los individuos con etiquetas que van de .4 a .20 y por otro lado los que van de .33 a .45, dentro del cuadrante, se puede observar que este último es " pequeño " y " cercano " los ejes. Este " pequeño " conjunto cuenta con la peculiaridad de que para la última variable se tiene la clase 2, y la 1 es el conjunto " grande " que se menciona antes.

Para el cuadrante *III*, la mayoría de los individuos contenidos en él, se encuentran cercanos a lo largo del eje horizontal y los restantes muy alejados tanto del centro como de los ejes, por lo que este cuadrante muestra mayor dispersión que en los cuadrantes

anteriores. Al analizar la Tabla 3 (*III*) se puede ver que ahora la clase 3 es la que domina aquí . Al igual que en los otros cuadrantes se tienen individuos alejados del centro y cercanos al eje, dominandos por alguna clase. En este caso para los individuos .3 y .16 predomina laclase 3 pero con la característica de que en .16, las primeras 10 variables y para .3 en las últimas 6. La similitud entre individuos, para este cuadrante es " grande ", pues forman pequeños grupos mejor " definidos " que en los cuadrantes anteriores.

El último cuadrante muestra a los individuos sin ninguna estructura en particular. El comportamiento que guardan los cuadrante *I*, *II* y *III* no se representa en *IV* a no ser la simulitud entre individuos. En la Tabla 3 (*IV*) se observa gran dominio de la clase 1 y es en general el que menor cantidad de individuos representa para esta gráfica.

Por Variables :

En la Gráfica 2 se puede observar que las clases de variables se van situando en cuadrantes muy específicos. La mayoría de las variables de clase 1 se encuentran en el *IV* cuadrante, lo mismo sucede con la clase 3 en el *III* y la clase 2 en el *I* y *II*.

Al agrupar todas las variables en cada una de sus clases, se puede observar que al trazar en círculo en el centro de la gráfica se encuentran solo las variables y clases que se podrían considerar como características del personaje, pues GRU en sus dos clases se encontraría dentro del dicho circulo.

Como se menciono anteriormente, la variable GRU se

consideró suplementaria y apesar de que sus clases se encuentran " cercanas " al centro, puede establecerse cierta diferencia pues se localizan en cuadrantes totalmente opuestos. Las variables más cercanas a GR1 y GR2 son :

GR1 : AC1, C+3, 102, AC3
GR2 : RE3, IM2, DO2, ES2

Conjunta :

En la gráfica 3 se puede observar pequeños conjuntos tanto de individuos o variables como de ambos, lo cual resulta un poco engorroso para analizar; sin embargo de esos conjutos existen algunos, con respecto a las variables, que se repiten, es decir :

En la parte superior del eje 2 al igual que en la parte media e inferior, principalmente en los extremos del eje 1, se encuentran las variables : MOS, T2, T3, T4, T6, y CI formando dichos conjuntos con respecto a su clase y a su alrededor los individuos que en un momento dado los caracterízan. De estos tres conjuntos se puede decir que además de que son los mejor representados gráficamente, pues entre más alejados se encuentren del centro mejor, son caracteristicas (esto es con respecto a las interpretaciones que se les puede dar), que no son determinantes para el objetivo principal del análisis. De lo demás aún no se podría dar una interpretación convincente de su relación.

Con lo que respecta a los ejes se puede decir que para el eje 1 (horizontal), es el que indica que las clases de las variables se encuentran ordenadas a lo largo de él, es decir, del lado izquierdo se encuentran

la mayoría de las variables de clase 3, en medio las de clase 2 y por último las de 3.

Al eje 2 (vertical) su interpretación es la siguiente : se puede decir que es el que delsiga completamente individuos y/o varibles que no se asocian con ninguna de otra clase, por lo antes mencionado además de que en la parte superior se encuentra la mayor parte de las variables de clase 2 y por debajo las demás; se puede decir que se encarga de separar los casos intermedios de los extremos, en el sentido de características muy particulares.

Esta gráfica explica el 24 % de inercia total, correspondiendo a sus valores singulares con 0.31 para el eje 1 y 0.17 para eje 2.

Segunda Gráfica

Por Individuos :

En el cuadrante *I* de la gráfica 4 se puede observar que los individuos no presentan mayor variabilidad entre ellos, se mantienen relativamente cercanos a los ejes (principalmente al horizontal), existiendo tambien similitud entre ellos. Para los que se encuentran en los extremos de los ejes sigue conservando la misma propiedad que en los anteriores, como se observa en los individuos .8 y .24. La clase que predomina en este cuadrante es la 1 seguida de 2 , esto es la patron inverso del cuadrante *I* en la primera gráfica.

El cuadrante *II* presenta disperción en los individuos, la mayoría de ellos se distribuyen a lo

.5	1	1	1	1	1	1	2	1	1	1	3	3	1	2	1	3	1	2	1		1	
.9	1	1	1	3	1	1	1	2	1	1	1	3	2	1	1	1	2	1	3	1		1
.10	3	1	2	2	1	2	1	2	2	3	3	1	2	1	1	2	2	2	3	2		1
.24	2	2	1	3	2	3	2	2	2	2	1	2	3	1	3	1	2	1	2	3		2
.25	3	2	2	2	2	2	2	2	2	2	3	3	2	2	3	2	2	1	3	2		2
.28	1	2	2	2	2	2	1	1	3	3	3	1	2	1	1	2	3	2	1	2		2
.32	3	1	1	1	1	1	1	1	1	1	1	1	2	3	2	2	3	2	2	3		2
.36	1	1	1	1	2	1	1	1	2	2	3	1	2	2	3	3	2	2	3	3		2
.38	2	1	1	1	1	1	1	1	2	3	1	2	3	3	3	3	3	3	2	1		2
.3	1	2	3	3	1	1	3	3	2	1	2	2	1	1	3	3	3	3	3	3		1
.14	2	1	2	2	2	3	3	2	3	2	2	1	1	3	3	1	1	1	3	3		1
.16	2	3	3	3	3	3	3	3	3	3	1	1	3	2	3	3	2	3	1	3		1
.18	1	2	2	3	3	3	2	3	2	1	3	2	1	3	2	2	2	2	2	1		1
.27	3	2	3	3	3	3	3	3	3	2	2	2	3	1	2	2	1	3	2	3		2
.33	2	2	2	3	2	3	3	2	3	3	1	3	3	1	3	2	2	1	2	3		2
.35	2	2	2	3	2	3	3	3	2	3	3	1	2	1	2	2	1	2	3	1		2
.37	1	3	2	3	2	3	3	1	1	2	1	2	2	2	2	2	2	2	2	2		2
.39	1	1	2	3	2	2	3	3	2	2	1	3	3	1	2	3	3	2	2	3		2
.41	3	2	3	1	2	3	3	2	3	3	2	3	3	3	2	2	2	3	3	3		2
.42	2	2	3	3	2	3	2	2	3	1	2	1	1	2	2	3	2	1	2	1		2
.43	2	2	2	3	1	3	1	2	3	3	1	1	1	3	3	2	3	3	3	2		2
.44	2	1	1	3	2	2	3	2	2	3	1	3	3	2	1	2	1	3	1	3		2
.45	2	2	2	3	2	1	2	2	3	2	2	2	3	2	2	2	2	2	2	2		2
.4	2	1	2	2	3	3	2	2	3	2	3	1	3	2	1	3	2	3	2	2		1
.7	3	3	3	2	2	3	2	3	3	3	1	3	3	1	3	1	2	1	1	1		1
.11	3	3	3	3	3	3	3	3	3	3	2	3	2	2	3	3	3	3	3	2		1
.12	3	2	3	3	3	2	3	3	3	3	2	3	1	3	3	2	1	3	1	2		1
.15	3	3	3	3	3	3	3	3	3	3	2	2	2	3	2	3	3	3	2	3		1
.19	2	3	2	2	3	2	2	2	1	2	2	1	1	1	2	1	1	3	1	2		1
.20	3	3	3	2	3	3	3	3	3	2	2	3	1	2	1	1	3	3	1	2		1
.22	3	3	3	3	3	3	2	3	3	2	3	2	3	3	3	3	1	3	1	3		1
.26	3	3	3	3	3	3	3	3	3	3	3	2	1	3	3	3	3	3	2	2		2
.34	1	3	3	3	2	2	3	3	3	3	2	3	2	3	1	1	1	2	1	1		2
.1	3	1	3	1	1	2	1	1	2	1	1	2	1	3	2	3	2	2	2	2		1
.2	3	2	2	2	1	2	2	1	3	3	3	3	3	3	2	1	1	1	2	2		1
.6	3	1	2	2	1	1	2	2	2	2	1	1	1	2	2	1	1	1	1	2		1
.8	3	3	3	2	1	1	2	3	1	1	3	3	1	1	2	1	2	3	2	1		1
.13	3	2	2	1	1	2	1	1	2	1	2	2	1	2	1	1	1	1	1	3		1
.17	2	1	1	2	3	1	2	1	2	1	2	2	1	2	1	1	1	1	1	2		1
.21	1	1	1	1	1	2	2	1	1	2	3	2	3	2	1	2	1	2	3	2		1
.23	1	2	1	1	1	1	1	1	1	1	3	1	2	1	1	1	1	1	2	1		2
.29	2	1	1	2	2	2	2	1	1	2	3	1	2	2	3	1	1	3	1	1		2
.30	1	1	1	1	1	2	2	1	1	1	1	3	1	1	1	2	2	1	2	2		2
.31	1	1	1	2	1	1	2	1	2	2	1	1	2	2	2	1	3	1	1	3		2
.40	2	1	1	2	1	1	1	1	1	1	3	3	2	1	2	3	1	2	3	2		2

4(*I*)

4(*II*)

4(*III*)

4(*IV*)

Tabla 4

largo de los ejes mas o menos cerca. Ahora son dos clases las que predominan en el cuadrante y son 2 y 3, sin encontrar mucha diferencia entre las frecuencias observadas para las primeras 20 variables en ambas, 115 y 111 respectivamente (Tabla 4(*II*)). Los 4 individuos de baja venta (.3, .14, .16, .18) son los que más cercanos al centro.

En el cuadrante *III*, los individuos se encuentran muy alejados unos de otros, sin observar que en algunos de ellos existe similitud (.22, .12 y .11, .26). La clase que domina es la 3 en todas las variables y las clases que dominan a los individuos extremos (los ubicados lejanos al centro y cercanos a los ejes), .19 y .15 son 2 y 3 respectivamente. Para .20 , individuos medios, las clases 1 y 2 se encuentran equilibradas, es decir, en la Tabla 4(*III*) para este individuo existe el mismo número de clases 1 y 2 .

Si se comparan los cuadrantes *IV* de la primera y segunda gráfica, existe menos disperción en esta última, se puede hablar de individuos extremos y la similitud es mayor. La clase que predomina es la 1 (Tabla 4(*IV*)). Si se toma encuenta las clases de la última variable, el conjunto de alta es pequeño y alejado del centro, mientras que el de baja se encuentra muy disperso y situado alrededor de los ejes.

Por Variables :

En la gráfica 5 se puede observar que existen tambien pequeños conjuntos de variables de los cuales existe uno que es el mayor de todos ellos, el cercano a GR2 y se encuentra en la parte superior del eje 3 (vertical). La mayoría de las variables, incluyendo las

clases se encuentran en los cuadrantes *I*, *II* y *IV*, gran parte de las variables de clase 3 en el *II* cuadrante, clase 2 y 1 en el *IV* y clase 1 y 2 en el *IV*. En el cuadrante *III* , las variables que hay son principalmente de clase 3 y se encuentran bastante alejadas del centro.

Con lo que respecta a las variables suplementarias, aquí se encuentran mejor representadas, por encontrarse alejadas del centro y al igual que en la gráfica anterior en lados contarios. GR2 en el *I* y GR1 en el *III*. Se hablaba del conjunto de variables más grande que se encuentran en la parte superior del eje 3, en esta gráfica como que las variables cercanas a el GR2 se podría decir que son características que se requiere la persona para garantizar venta alta. Estas variables son : CI2, ES2, ZO2, RE2 y AD2; pueden tambien tomarse encuenta : C-1, ME1 y VI1.

Con lo que respecta a GR1 existe una variables que se determina como caracteristica trascendental que es AC1. Después podría considerarse las variables PL2 y ES1.

Conjunto :

Al analizar la gráfica 6 se puede observar que realmente en la parte superior, se encuentran tanto los individuos y variables como sus respectivas clases, que se consideran de venta alta y en la parte inferior lo contrario. Entonces el eje 3 es quien realmente determina el objetivo principal del análisis.

Existen 5 conjuntos relativamente " grandes ", de los cuales 3 se encuentra en la parte superior y 2 en la inferior que serían los marcados en la gráfica. De estos

5 grupos solamente interesan 2 que en particular son los que se encuentran a lo largo del eje 3, determinando las características e individuos que establecen dicha diferencia. Para el GR2 se tiene que las variables son : CI2, ES2, 202, RE2 y AD2, y posiblemente DO2 AC3 102, 402 y PL3, las qué más contribuyen y los individuos son : 24, 37, 44, 18 principalmente y posiblemente 45, 39, 35 y 43.

Para GR1 se tiene que las variables son : AC1, ES1, PL2, VI2 y 602, tal vez 502.

Esta Gráfica explica tanto a los individuos como a las variables en un 23 % de la inercia total, el valor singular correspondiente a el eje 3 es de 0.15.

Tercera Gráfica

Por Individuos :

Debido a que ya se tomaron encuenta la situación de los individuos con respecto a todos los ejes, se tiene la siguente para la tercera gráfica :

I.- Existe mayor variabilidad entre los individuos que en los demás cuadrantes *I*. Al igual que en la segunda gráfica predomina la clase 2. La similitud entre individuos no es tan " fuerte " como en las demás. No existen individuos extremos.

II.- Los individuos estan más cercanos unos a otros . Predominan las clases 3 y 1, principalmente en las últimas 10 y 10 primeras variables respectivamente. Existen individuos extremos y la similitud de los individuos es mayor que en *I* . La mayoría de los

.10	3	1	2	2	1	2	1	2	2	3	3	1	2	1	1	2	2	2	3	2	1	
.14	2	1	2	2	2	3	3	2	3	2	2	1	1	3	3	1	1	1	3	3	1	
.24	2	2	1	3	2	3	2	2	2	2	1	2	3	1	3	1	2	1	2	3	2	
.25	3	2	2	2	2	2	2	2	2	2	3	3	2	2	3	2	2	2	1	3	2	
.28	1	2	2	2	2	2	1	1	3	3	3	1	2	1	1	2	3	2	1	2	2	
.33	2	2	2	3	2	3	3	2	3	3	1	3	3	1	3	2	2	1	2	3	2	5(*I*)
.35	2	2	2	3	2	3	3	3	2	3	3	1	2	1	2	2	1	2	3	1	2	
.37	1	3	2	3	2	3	3	1	1	2	2	2	2	2	2	2	2	2	2	2	2	
.42	2	2	3	3	2	3	2	2	3	1	2	1	1	2	2	3	2	1	2	1	2	
.44	2	1	1	3	2	2	3	2	2	3	1	3	3	2	1	2	1	3	1	3	2	
.45	2	2	2	3	2	1	2	2	3	2	2	2	3	2	2	2	2	2	2	2	2	
.3	1	2	3	3	2	2	3	2	3	3	2	2	3	3	3	3	3	3	3	3	1	
.5	1	1	1	1	1	1	1	2	1	1	1	3	3	1	2	1	3	1	2	1	1	
.9	1	1	1	3	1	1	1	2	1	1	1	3	2	1	1	1	2	1	3	1	1	
.16	2	3	3	3	3	3	3	3	3	3	1	1	3	2	3	3	2	3	1	3	1	
.18	1	2	2	3	3	3	2	3	2	1	3	2	1	3	2	2	2	2	2	2	1	
.27	3	2	3	3	3	3	3	3	3	2	2	2	3	1	2	2	1	3	2	3	2	5(*II*)
.32	3	1	1	1	1	1	1	1	1	1	1	1	2	3	2	2	3	2	2	2	3	
.36	1	1	1	1	2	1	1	1	2	2	3	1	2	2	3	3	2	2	3	3	2	
.38	2	1	1	1	1	1	1	1	2	3	1	2	3	3	3	3	3	3	3	2	1	
.39	1	1	2	3	2	2	3	3	2	2	1	3	3	1	2	3	3	2	2	3	2	
.41	3	2	3	1	2	3	3	2	3	3	2	3	3	3	2	2	2	3	3	3	2	
.43	2	2	2	3	1	3	1	2	3	3	1	1	1	3	3	2	3	3	2	2	2	
.1	3	1	3	1	1	2	1	1	2	1	1	2	1	3	2	3	2	2	2	2	1	
.8	3	3	3	2	1	1	2	3	1	1	3	3	1	1	2	1	2	3	2	1	1	
.11	3	3	3	3	3	3	3	3	3	3	3	2	3	2	2	3	3	3	3	3	2	
.12	3	2	3	3	3	2	3	3	3	3	2	3	1	3	3	2	1	3	1	2	1	
.15	3	3	3	3	3	3	3	3	3	3	3	2	2	2	3	2	3	3	3	2	3	
.22	3	3	3	3	3	3	2	3	3	2	3	2	3	3	3	3	1	3	1	3	1	5(*III*)
.23	1	2	1	1	1	1	1	1	1	1	3	1	2	1	1	1	1	1	1	2	1	
.26	3	3	3	3	3	3	3	3	3	3	3	2	1	3	3	3	3	3	3	2	2	
.30	1	1	1	1	1	2	2	1	1	1	1	3	1	1	1	2	2	1	2	2	2	
.34	1	3	3	3	2	2	3	3	3	3	2	3	2	3	1	1	1	2	1	1	2	
.40	2	1	1	2	1	1	1	1	1	1	3	3	2	1	2	3	1	2	3	2	2	
.2	3	2	2	2	1	2	2	1	3	3	3	3	3	3	2	1	1	1	2	2	1	
.4	2	1	2	2	3	3	2	2	3	2	3	1	3	2	1	3	2	3	2	2	1	
.6	3	1	2	2	1	1	2	2	2	2	1	1	1	2	2	1	1	1	1	2	1	
.7	3	3	3	2	2	3	2	3	3	3	1	3	3	1	3	1	2	1	1	1	1	
.13	3	2	2	1	1	2	1	1	2	1	2	2	1	2	1	1	1	1	1	3	1	
.17	2	1	1	2	3	1	2	1	2	1	2	2	1	2	1	1	1	1	1	2	1	5(*IV*)
.20	3	3	3	2	3	3	3	3	3	2	2	3	1	2	1	1	3	3	1	2	1	
.21	1	1	1	1	1	2	2	1	1	2	3	2	3	2	1	2	1	2	3	2	1	
.29	2	1	1	2	2	2	2	1	1	2	3	1	2	2	3	1	1	3	1	1	2	
.31	1	1	1	2	1	1	2	1	2	2	1	1	2	2	2	1	3	1	1	3	2	

Tabla 5

individuos se encuentran a lo largo del eje vertical.

III.- Los individuos no están tan alejados unos de otros. Predomina la clase 3 siendo de poca las clases 1 y 2 pero en cantidad similar. La similitud entre individuos es mayor que en los anteriores, pero cercanos a los ejes y relativamente al centro.

IV.- Presenta a los individuos muy alejados de ellos. Existe poca similitud . Predomina la clase 1 en las 20 variables y la situación de los individuos es inversa a las del cuadrante *I*.

Por Variables :

Como se puede observar en la gráfica 8, prevalece solamente un conjunto, en los cuadrante *I* y *II*, mientras que las demás se distribuyen en conjuntos " pequeños " y dispersos en todos los cuadrantes. El conjunto " grande " es intregrado por las variables y clases :

*	DO2	IM3	PL3	AD1
*	102	AC2	303	
*	AC3	IM2	SO3	
*	ES2	603	C-1	
*	RE3	VI1	PL1	
*	503	MO2	RE2	

y GR2. Cercanas a este último son las marcadas con "*".

Conjunta :

La gráfica 9 muestra que las variables antes mencionadas en la gráfica 8 y ME3, C+3 y ME2, junto con los individuos .9, .39, .41, .18, .36, .43, .27, .16,

.37, .28, .42, .10, y .35, como parte del conjunto más " grande " de dicha gráfica. Los demás conjuntos son relativamente pequeño y podría considerarse en un momento dado irrelevantes para el objetivo del análisis. Con respecto a GR1 en ningun momento muestra cambio pues a su alrededor solamnete permanece junto a ella AC1. Los ejes 2 y 3 explican a los individuos y variables en un 15 % total.

Conclusiones de las Gráficas

El análisis de las tres gráficas llevan a concluir lo siguiente :

La mayoría de individuos y variables que realmente tienen algún significado en el objetivo principal son :

Para GR2

Variables : RE3, DO2, ES2, CI2, CI2, 202 y RE2.

Individuos : .24 y .43.

Para GR1

Variables : AC1.

Individuos : Ninguno.

por lo que se puede concluir de este análisis que el prototipo del vendedor " ideal " en términos de los niveles que lo caracterízan tiene: muy responsable, medio dominante, medio estable, medio inteligente y dominar la Tarea 2 (202) medianamente. Con respecto a

las personas que definitivamente no se deben de aceptar son todas aquellas que el nivel de actividad sea nulo, independientemente de los otros niveles en los que se encuentran con respecto al resto de las variables.

El eje 2 no muestra mayor importancia para determinar la venta alta o baja. Esto es porque para la primera y tercera gráfica la variable GRU no mostraba mayor trascendancia al ser decrita por este eje, en ambas gráficas las clases de GRU se encontraban totalmente representadas por los ejes 1 y 3.

La gráfica que mayor importancia e información da con respecto al objetivo es la segunda, por lo que realmente fué la que determinó el prototipo de persona " ideal ", tanto de venta alta como baja. Esta gráfica describe en un 23 % al personaje en cuestion, lo cual se considera " bueno " en el sentido de que al bajar de un espacio de dimensión 60 a un subespacio de 3 dimensiones es de aportación significativa.

Comentarios Adicionales

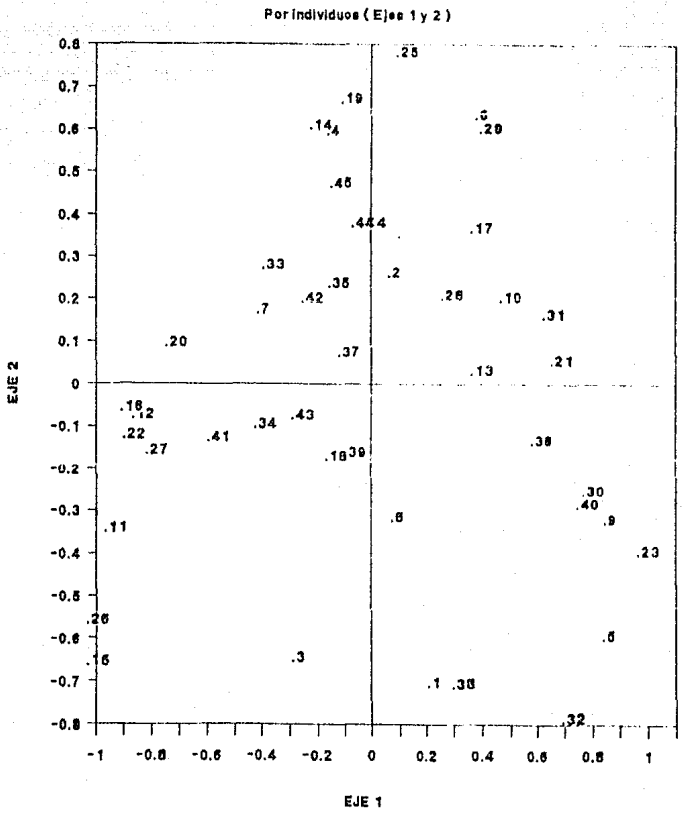
Después de analizar las gráficas y dar los resultados obtenidos, los encargados del proyecto no se mostraron del todo satisfechos por lo que se realizó nuevamente el análisis pero tomando a la variable GRU dentro de las activas. Dicho cambio no fué significativo, pues aún después de que la variable se consideró activa, las gráficas no mostraron mucho cambio, las variables siguieron siendo las mismas (Gráfica 10, ejes 1 y 3).

Se continuó trabajando en el proyecto pero tomando en cuenta el análisis realizado, es decir, las variables

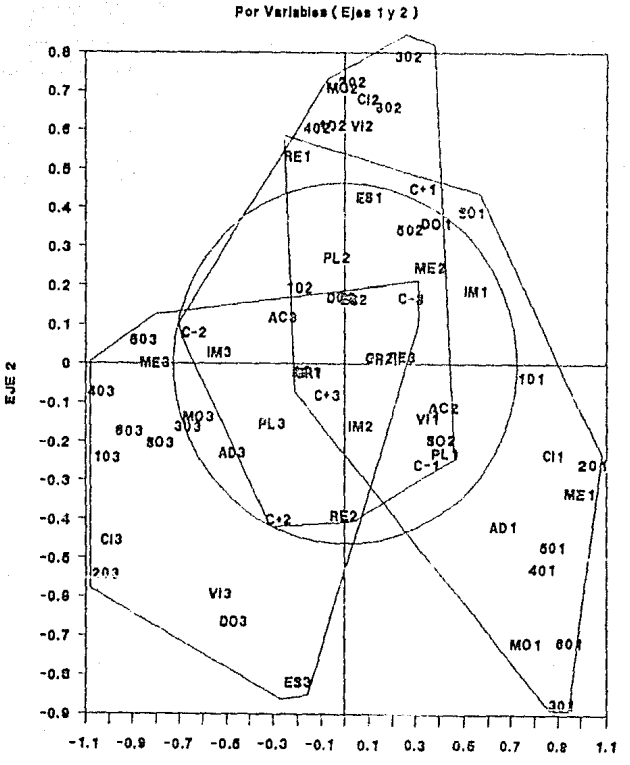
que se les indicó que eran significativas para establecer diferencia, sirvieron para concluir que del 64 % de la gente que se les aplicaba dichas pruebas tenían un margen de error del 3 % por lo que consideraron que era un buen punto de partida para predecciones futuras.

En general el análisis fue muy limitado pues solamente se obtuvo la información que se necesitaba de manera inmediata, pero se podrían haber analizado mucho más.

A. C. M. POR RENGLON

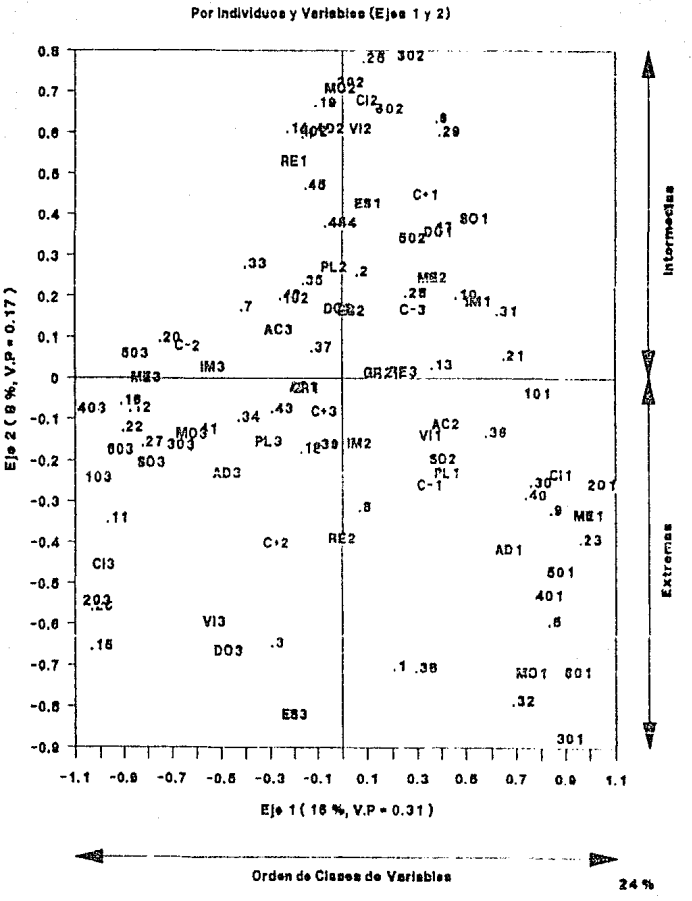


A. C. M. POR COLUMNA



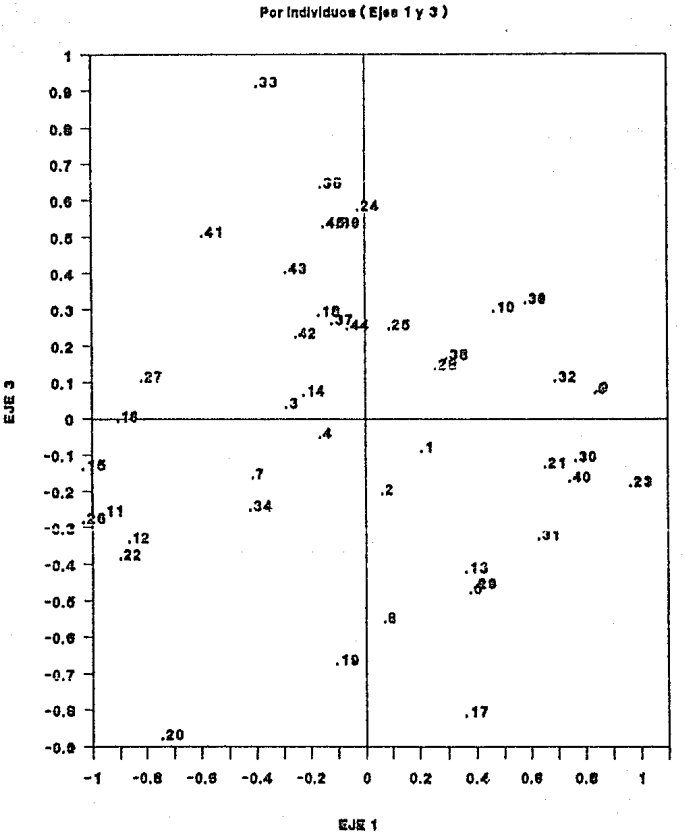
GRAFICA # 2

A. C. M. CONJUNTA



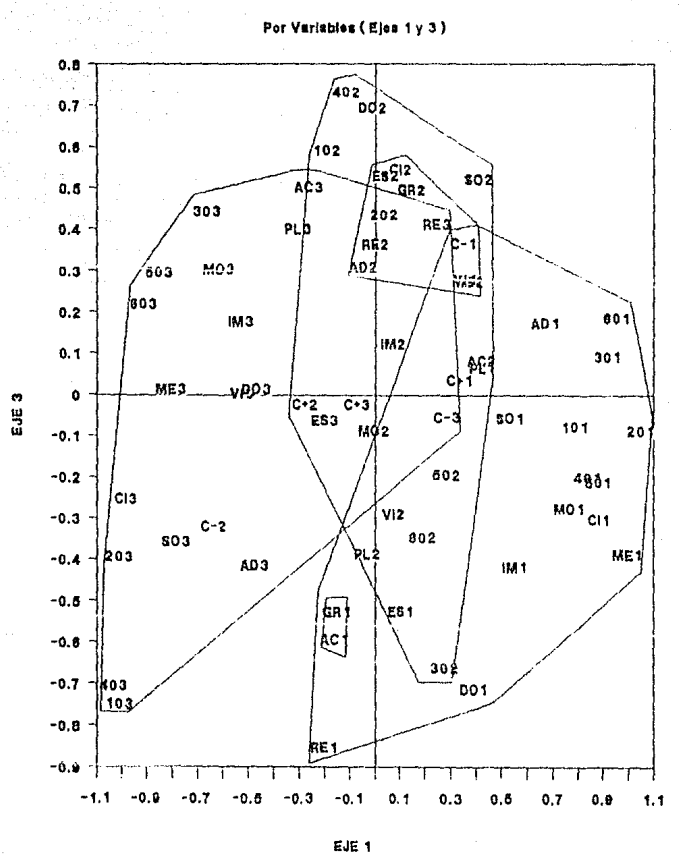
GRAFICA Nº 3

A. C. M. POR RENGLON



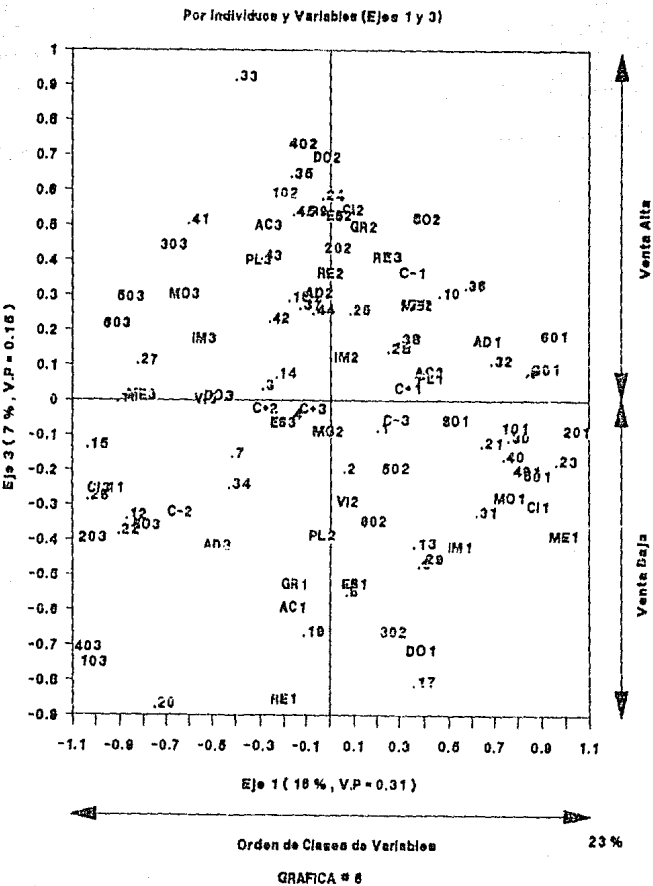
GRAFICA # 4

A. C. M. POR COLUMNA



GRAFICA # 6

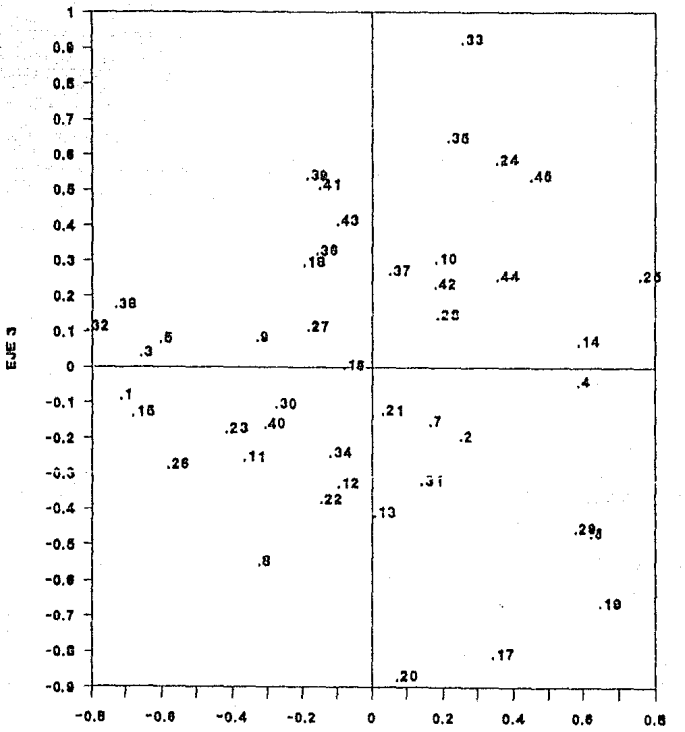
A. C. M. CONJUNTA



GRAFICA # 6

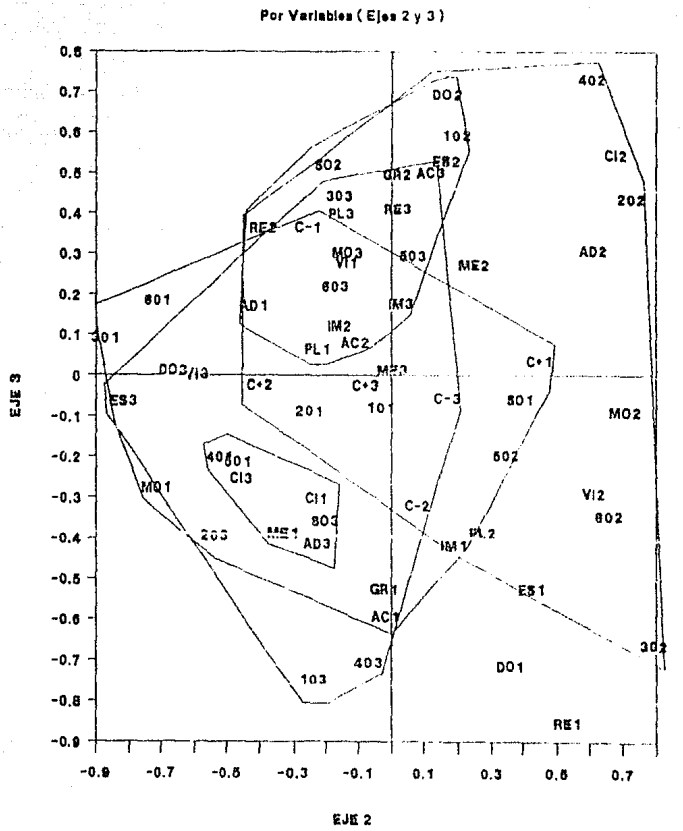
A. C. M. POR RENGLON

Por individuos (Ejes 2 y 3)



GRAFICA # 7

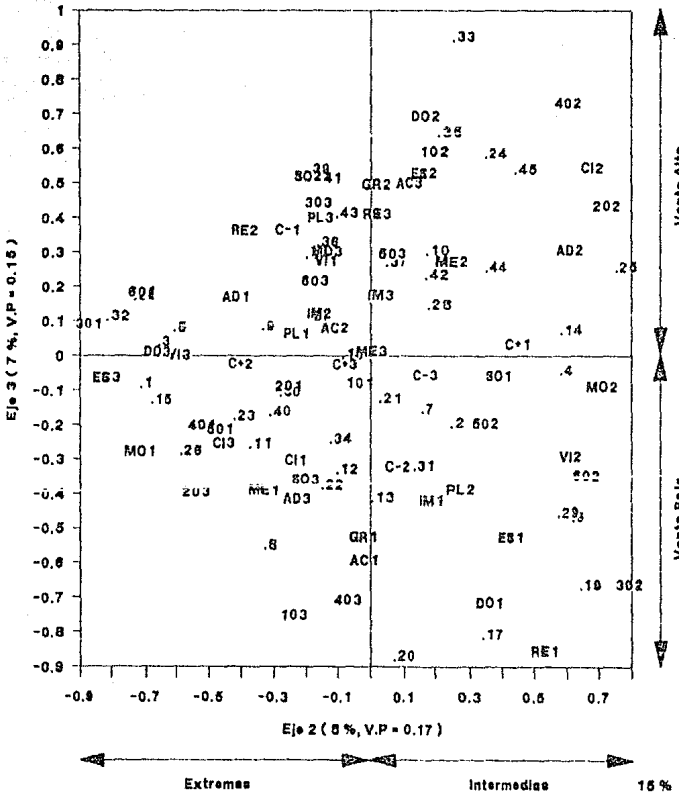
A. C. M. POR COLUMNA



GRAFICA # 8

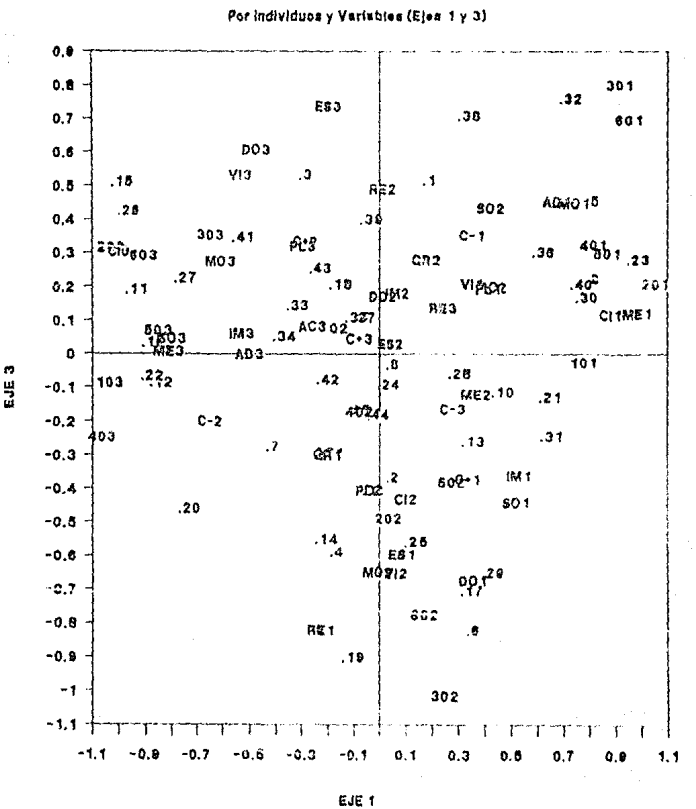
A. C. M. CONJUNTA

Por Individuos y Variables (Ejes 2 y 3)



GRAFICA # 9

A. C. M. CONJUNTA VARIABLE GRU ACTIVA



GRAFICA # 10

CONCLUSIONES

Una vez analizado como se relacionan las variables a traves de una técnica descriptiva como lo es el Análisis de Correspondencias, es importante hacer notar las ventajas y desventajas que ofrece dicha técnica al aplicarlo de manera simple o múltiple

En general, para una Tabla Contingencia de dos variables se observó que al bajar de dimensión, la contribución de la inercia explicada por los ejes mantiene una representación gráfica " buena " en el sentido de que la información " perdida " es relativamente pequeña A pesar de esto, existen ejes que explican muy bien a las variables y que algunos de ellos resultan ser relativamente despreciables o de poca importancia, es decir, en el Capítulo II, el eje 1 contribuía a explicar a las variables en un 95.1 % de la inercia total mientras que el eje 2 solamente aporta el 4.9 %. Estas diferencias entre ejes casi siempre son muy grandes por lo que resulta poco gratificante, pues siempre se espera que el hecho de tener la participación de más ejes, contribuya a explicar mejor a los resultados requeridos.

Sin embargo existen Tablas de Contingencia de dos variables que tienen gran dimensionalidad y que el poder bajarlas de dimensión J a un subespacio de dimensión K, el cual pueda ser interpretado siempre es satisfactorio a pesar de las grandes diferencias entres los ejes.

Al continuar analizando el Análisis de Correspondencias Simples, se puede observar que si se requiere de un individuo que cumpla las características observadas en dicho análisis, no es posible encontrarlo, pues en una Tabla de Contingencia solo interaccionan las categorías de las variables y no los individuos, por lo que de estos no se puede decir nada.

Con lo que respecta al Análisis de Correspondencias Múltiples se tienen las siguientes observaciones:

Siempre el hecho de explicar resultados con más de dos variables es tentativo, más aún, si para cada variables se toma en cuenta todas las categorías posibles y tambien una cantidad considerable de individuos para describir el resultado ó resultados a observar.

Sin embargo el involucrar varias variables implica mayor dimensión en el espacio generado y por lo tanto mayor " perdida " de información, pues para el análisis se requiere un subespacio de menor de dimensión de tal manera que pueda ser representado gráficamente para analizarlo.

Es importante hacer notar que en el Análisis de Correspondencia Múltiples la contribución de la inercia explicada por los ejes es mucho menor que en el Análisis de Correspondencias Simples, lo cual resulta poco atractivo, pero que las diferencias entre los ejes no son tan drásticas, es decir en el Capítulo IV se observó que los 3 ejes explicaban los resultados en un 31 %, cada eje contribuyó en 16 %, 8 % y 7 % respectivamente que a comparación del ejemplo del Capítulo II, el involucrar más ejes implicaba tener mayor contribución a

explicar a las variables y resulta significativa.

Además de observar como interacciona las variables, se puede ver en que nivel se da dicha interacción, lo cual resulta de gran interés.

Con lo que respecta a los individuos, mediante este análisis, se puede observar cual es el más representativo con respecto al conjunto de variables que se forman en cada gráfica.

Siempre el hecho de trabajar con promedios, en este caso para las coordenadas de individuos y variables resulta contraproducente, pues en algunos casos existen observaciones aberrantes que no pueden detectarse y que siempre contribuyen a afectar los resultados del análisis a pesar de tratarse de variables cualitativas.

En general el Análisis de Correspondencias es una técnica que uno de sus objetivos principales es describir a los individuos y variables sin tener que hacer suposiciones de ningun tipo, que a diferencia de otras técnicas resulta de gran interés puesto que muchas veces esos supuestos no se cumplen.

Otra de las grandes aportaciones que da el Análisis de Correspondencias es que una vez obtenida la representación gráfica de variables - e individuos - se observa la estructura de asociación que existe entre ellas y por tanto por medio de otras técnicas esta asociación puede ser modelada.

BIBLIOGRAFIA

CASTAÑO, Eduardo. (1986). " Primer Foro de Estadística: Una Aplicación del Análisis de Correspondencias en el Análisis Sensorial de Alimentos "UACPyP (IIMAS), UNAM.

CASTAÑO, Eduardo. (1987). " Inferencia Estadística en el Análisis de Correspondencias ". Tesis de Maestría, UACPyP (IIMAS), UNAM.

ENRIQUEZ, Jaqueline. (1989). " El Uso de Paquetes Estadísticos para Modelos Loglineales " . Tesis Profesional, Facultad de Ciencias, UNAM.

GREEACRE, Michael J. (1984). " Theory and Aplication of Correspondence Analysis ". Edit. Academic Press.

HOFFMAN, Kenneth. (1982). " Algebra Lineal " . Edit. Prentice/Hall Internacional.

JOHNSTON, J. (1985). " Econometric Methods "Tercera Edición. Edit. McGraw Hill.

LARA, Mario. (1989). " Zonación y Caracterización de los Escleractinios en el Arrecife Anegada de Afuera, Veracruz, México ". Tesis Profesional, Facultad de Ciencias, UNAM.

LEBART, Ludovic. (1984). " Multivariate Descriptive

Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices ". Edit. Wiley.

PADILLA, Claudia. (1989). " Estructura Comunitaria de Escleractineos del Arrecife el Cabezo, Veracurz ". Tesis Profesional, Facultad de Ciencias, UNAM.

PANIAGUA, Mónica. (1986). "El Análisis de Correspondencias: un Método Multivariado Descriptivo ". Tesis Profesional. ITAM.

STRANG. (1982). " Algebra Lineal y sus Aplicaciones ". Edit. Fondo Educativo Interamericano.

VAN RIJCKEVORSEL, Jan. (1988). " Component and Correspondence Analysis : Dimension Reduction by Functional Approximation ". Edit. Wiley.