Análisis de Correspondencias: Más Allá de un Mapa...

Raúl Vargas y Norma Escobedo
Auditoría y Estrategia Empresarial (Delta Research)

Como muchos sabemos, el Análisis de Correspondencias es un técnica muy utilizada en investigación de mercados, cuando es necesario implementar modelos de análisis estadístico multivariado, con la finalidad de reducir dimensiones o en busca de las estructuras o relaciones entre ciertas variables de tipo categórico. Aunque esta técnica no es reciente, el problema que ha surgido para la misma es que muchas veces las relaciones obtenidas están basadas solamente en cómo se ven los puntos en un mapa y del criterio de quien lo esta analizando. Otro problema es que las relaciones entre las variables no siempre son tan claras en un gráfico como nos gustaría que lo fueran. Por lo anterior es importante que las personas que realizan este tipo de análisis tengan en mente algunos detalles de lo que involucra, que no es solamente la parte gráfica obtenida del mismo.

Cuando la relación entre las variables no es lineal o éstas han sido medidas en una escala de tipo ordinal o nominal; es decir, carecen de propiedades que nos indiquen el significado de la distancia entre un valor y otro de dichas variables, es posible aplicar ciertos procedimientos conocidos como procedimientos de escalamiento óptimo, para trabajar con variables medidas de esta manera y, si es necesario, asignarles un nuevo tipo de escala, de tal forma que el modelo utilizado se ajuste mejor a los datos.

El análisis de correspondencias es un procedimiento de escalamiento óptimo y se puede implementar a partir de una tabla en la que los datos no sean negativos, que puede ser de alguno de las siguientes tipos:

- √ **Tablas de Contingencia que agrupen a los individuos en ciertas categorías.**
- √ **Tablas de Frecuencias.**
- √ **Tablas de Valoración con puntuaciones como medias, sumas, índices, etcétera.**
- √ **Tablas de 0 y 1 con preferencias por ciertas marcas por ejemplo.**
- √ **Tablas con otras medidas de correspondencias entre filas y columnas.**
- √ **Tablas múltiples con marcas, atributos, estilos de vida, etcétera.**

Entonces, entre las ventajas de este procedimiento, encontramos que no es necesario que la tabla de datos utilizada sea una tabla de contingencia, o en otras palabras, que cada uno de los individuos a partir de los cuales se creó la tabla sea clasificado solamente en una celda y sólo en una, de hecho en muchas ocasiones la tabla de datos no cumple esta característica ya que, si se permite a un individuo asociar un atributo a dos marcas distintas o se le pide al mismo individuo asociar varios atributos a ciertas marcas, se rompe con la característica de que cada individuo debe estar clasificado en una única celda de la tabla.

Este análisis nos aporta información que no es posible obtener con el estadístico X^2 o con los coeficientes de correlación y aunque no deja de ser una técnica de análisis factorial, no se centra exclusivamente en la reducción de dimensiones, sino que trata de descubrir afinidades entre las variables.

El Análisis

Como se dijo anteriormente, este análisis puede partir de una tabla de frecuencias, que resulte del cruce de dos variables, de tipo categórico.

Como ejemplo supongamos que tenemos que analizar los datos de la tabla 1 que es una tabla de asociación entre atributos y marcas.

Lo primero que se obtiene…

Con base en esta tabla se obtienen los ***perfiles por filas*** (tabla 2**),** que nos indican *la presencia de una*

Tabla 1

	MARCA 1	MARCA 2	MARCA 3	MARCA 4	Total
Atrib A	14	38	25	18	95
Atrib B	14	28	25	25	92
Atrib C	42	22	11	13	88
Atrib D	10	10	32	26	78
Atrib E	6	33	5	27	71
Atrib F	54	33	8	2	97
Atrib G	24	21	16	34	95
Atrib H	24	3	20	3	50
Atrib I	5	3	3	31	42
Total	193	191	145	179	708

Tabla 2

PERFILES POR FILAS					
	MARCA 1	MARCA 2	MARCA 3	MARCA 4	Total
Atrib A	0.147	0.400	0.263	0.189	1
Atrib B	0.152	0.304	0.272	0.272	1
Atrib C	0.477	0.250	0.125	0.148	1
Atrib D	0.128	0.128	0.410	0.333	1
Atrib E	0.085	0.465	0.070	0.380	1
Atrib F	0.557	0.340	0.082	0.021	1
Atrib G	0.253	0.221	0.168	0.358	1
Atrib H	0.480	0.060	0.400	0.060	1
Atrib I	0.119	0.071	0.071	0.738	1
Total	0.273	0.270	0.205	0.253	1

marca en cada uno de los atributos y se calculan por medio del cociente entre la frecuencia de cada celda en una fila y el total para esa fila. También los ***perfiles por columnas***, que nos indican *la presencia de un atributo en cada una de las marcas*, calculados por medio del cociente entre la frecuencia de cada celda en una columna y el total para esa columna. (Tabla 3).

Qué se busca…

A partir de estos perfiles, lo que se busca es: obtener una representación geométrica (mapa en dos o tres dimensiones) de los atributos, tomando en cuenta la forma en que se distribuyen las frecuencias relativas de las marcas y viceversa, para esto se utiliza una distancia denominada ***Distancia*** X^2 (ji-cuadrada). Para encontrar la distancia X^2 entre dos atributos, digamos el Atributo A y el Atributo H, se transforman los valores de la tabla de perfiles por fila (tabla 2) dividiendo cada renglón de ésta entre la raíz cuadrada del total para cada marca en la tabla 1 (valores en azul de dicha tabla). Estos valores se muestran en la tabla 4.

Tabla 3

PERFILES POR COLUMNA					
	MARCA 1	MARCA 2	MARCA 3	MARCA 4	Total
Atrib A	0.073	0.199	0.172	0.101	0.134
Atrib B	0.073	0.147	0.172	0.140	0.130
Atrib C	0.218	0.115	0.076	0.073	0.124
Atrib D	0.052	0.052	0.221	0.145	0.110
Atrib E	0.031	0.173	0.034	0.151	0.100
Atrib F	0.280	0.173	0.055	0.011	0.137
Atrib G	0.124	0.110	0.110	0.190	0.134
Atrib H	0.124	0.016	0.138	0.017	0.071
Atrib I	0.026	0.016	0.021	0.173	0.059
Total	1.000	1.000	1.000	1.000	1.000

Tabla 4

	MARCA 1	MARCA 2	MARCA 3	MARCA 4
Atrib A	0.011	0.029	0.022	0.014
Atrib B	0.011	0.022	0.023	0.020
Atrib C	0.034	0.018	0.010	0.011
Atrib D	0.009	0.009	0.034	0.025
Atrib E	0.006	0.034	0.006	0.028
Atrib F	0.040	0.025	0.007	0.002
Atrib G	0.018	0.016	0.014	0.027
Atrib H	0.035	0.004	0.033	0.004
Atrib I	0.009	0.005	0.006	0.055

A continuación utilizamos la fórmula de la ***distancia euclídea*** para los atributos A y H (valores sombreados en azul en la tabla 4), la fórmula es:

$$\mathbf{D^2(AtribA,AtribH)} = (A_1-H_1)^2+\ldots+(A_4-H_4)^2. \mathbf{(1)}$$

Donde:

A_i = Celda correspondiente al Atributo A y a la Marca **i.**

H_i = Celda correspondiente al Atributo H y a la Marca **i**.

Así cada atributo queda representado por nuevos valores, que toman en cuenta la forma en que se distribuyen las frecuencias de las marcas y la reducción de dimensiones para los atributos se realiza a partir de estos valores transformados.

Cada uno de los I Atributos está representado en un espacio de dimensión 4, ya que tenemos 4 marcas, y separado de los otros por la distancia euclídea ordinaria (fórmula **(1)**).

Después de esto se realiza un nuevo ajuste a los datos de la tabla 4 multiplicando cada celda por la raíz del *Total General* de la tabla 1, en este caso por √780, estos datos se muestran en la tabla 5.

Tabla 5

	MARCA 1	MARCA 2	MARCA 3	MARCA 4
Atrib A	0.282	0.770	0.581	0.377
Atrib B	0.291	0.586	0.600	0.540
Atrib C	0.914	0.481	0.276	0.294
Atrib D	0.246	0.247	0.907	0.663
Atrib E	0.162	0.895	0.156	0.756
Atrib F	1.066	0.655	0.182	0.041
Atrib G	0.484	0.426	0.372	0.712
Atrib H	0.919	0.116	0.884	0.119
Atrib I	0.228	0.138	0.158	1.468

Con los nuevos datos obtenidos en la tabla 5 se puede calcular una matriz de datos que contiene las ***covarianzas*** entre las marcas tomando en cuenta la ***masa*** de los atributos, la masa se *define como la influencia que ejerce un atributo* y es igual a su frecuencia marginal (valores sombreados en azul de la tabla 3), y calculando su desviación con respecto a las ***medias de las marcas*** definidas como la raíz cuadrada de las masa para cada marca (valores sombreados en azul en la tabla 2), los valores en esta matriz deben ser los mismos por encima y por debajo de la diagonal.

Como ejemplo, para calcular la covarianza entre la Marca 1 y la Marca 4, tomamos la columna de valores de dichas marcas de la tabla 5 y la masa de cada atributo de la columna de **Total** para cada atributo de la tabla 3, multiplicamos estas tres columnas y sumamos los valores obtenidos; por otra parte, multiplicamos la raíz cuadrada de la masa de la marca 1 y la raíz cuadrada de la masa de la marca 4 y a la suma anterior le restamos esta multiplicación, con lo que obtenemos la covarianza como se muestra en la tabla 6.

Tabla 6

	Marca 1	Marca 4	Masa Atrib	MasaAtribX Marca1XMarca4
Atrib A	0.28	0.38	0.13	0.014
Atrib B	0.29	0.54	0.13	0.020
Atrib C	0.91	0.29	0.12	0.033
Atrib D	0.25	0.66	0.11	0.018
Atrib E	0.16	0.76	0.10	0.012
Atrib F	1.07	0.04	0.14	0.006
Atrib G	0.48	0.71	0.13	0.046
Atrib H	0.92	0.12	0.07	0.008
Atrib I	0.23	1.47	0.06	0.020
MEDIAS	**0.522**	**0.503**	**Suma**	**0.178**
Media1 X Media4 =				**0.26**
Covarianza entre Marca 1 y 4				**-0.084**

Las covarianzas obtenidas entre las marcas 1 a la 4 son:

	MARCA 1	MARCA 2	MARCA 3	MARCA 4
MARCA 1	**0.11**	-0.01	-0.02	**-0.08**
MARCA 2	-0.01	**0.06**	-0.03	-0.02
MARCA 3	-0.02	-0.03	**0.07**	-0.01
MARCA 4	**-0.08**	-0.02	-0.01	**0.12**

Matriz de Covarianzas

Los datos que resultan a partir de la matriz de covarianzas son…

A partir de esta matriz de covarianzas se obtienen los siguientes resultados, con sus correspondientes tablas:

Los valores propios que nos indican el porcentaje de la variabilidad total que representa cada dimensión y por medio de los cuales, podemos darnos cuenta del porcentaje de variabilidad que conservamos con la representación de los datos en 2 dimensiones, por ejemplo.

La inercia total que nos indica la variabilidad total de los datos, se define como el promedio de las distancias de los distintos puntos al origen, donde a cada punto se le asigna un peso de acuerdo a su masa, si la tabla con la que trabajamos es una tabla de frecuencias la inercia total es igual al Estadístico X^2 dividido entre el Total General y es igual a la su-

ma de los elementos en la diagonal principal de la matriz de covarianzas (valores en azul de dicha matriz) o a la suma de los cuadrados de los valores propios. En este ejemplo la inercia total explicada es de 0.355 aproximadamente.

El Estadístico X^2 que se utiliza para tablas de contingencia y nos indica si dos variables cruzadas en una tabla son independientes o no, si lo son los valores en cada celda de la tabla serían muy parecidos, entre más grande sea este estadístico existe mayor asociación entre las categorías de la tabla, está definido como la suma ponderada de todas las distancias al cuadrado entre los perfiles fila y la media de los mismos y los perfiles columna y la media de los mismos, para este caso es de 251.327 aproximadamente, si lo dividimos entre el total general de la tabla 1 que es de 708 nos da la inercia total explicada 0.355, como se había señalado. Tabla 7.

Tabla 7

Resumen				
	Valor	Inercia	Chi-cuadrado	Sig.
	Propio			
Dimensión				
1.00	0.45	0.20		
2.00	0.30	0.09		
3.00	0.25	0.06		
Total		0.35	251.33	0.00
a	24 grados de libertad			

La Proporción de Inercia explicada por cada dimensión debe ser distinta, porque si cada dimensión explica la misma proporción de inercia no existe asociación entre marcas y atributos; es decir, la proporción de inercia explicada por cada dimensión debe ser mayor a 33% en nuestro caso porque a lo más tenemos 3 dimensiones.

Es importante que el porcentaje de inercia acumulada para las dimensiones elegidas sea lo mayor posible, este porcentaje ***nos indica la calidad de nuestra solución***, lo que se pretende con este análisis es representar todos los atributos y las marcas en menos dimensiones que las originales pero perdiendo la mínima cantidad de información, en este caso es de 82.1% con 56.5% para la primera dimensión y de 25.6% para la segunda dimensión.

La proporción de inercia explicada por cada dimensión es igual al cuadrado del valor propio correspondiente a dicha dimensión, al proyectar los puntos sobre la primera dimensión (las coordenadas del primer eje), la deformación de las distancias entre los mismos es menor que al proyectarlos sobre cualquier otra dimensión, al proyectarlos sobre la segunda dimensión (coordenadas del segundo eje) la deformación de las distancias entre los puntos será la segunda menor y así sucesivamente, generalmente utilizamos dos dimensiones (dos ejes) o tres.

El máximo número de dimensiones posibles es el mínimo entre el número de filas en la tabla menos uno y el número de columnas en la tabla menos 1(en nuestro ejemplo es de 3).

La proporción de inercia para cada dimensión se encuentra en la tabla 8.

Tabla 8

Proporción de inercia		Confianza para el Valor propio	
		Desviación	Correlación
Explicada	Acumulada	típica	2
0.56	0.56	0.03	-0.04
0.26	0.82	0.03	
0.18	1.00		
1.00	1.00		

Las nuevas coordenadas de los atributos para las dimensiones que hemos elegido, en este caso 2, se

Tabla 9

Examen de los puntos de fila				
		Puntuación		
	Masa	en la dimensión		Inercia
		1	2	
ATRIB				
A	0.13	-0.11	-0.02	0.02
B	0.13	-0.29	0.15	0.01
C	0.12	0.65	-0.18	0.03
D	0.11	-0.54	0.91	0.04
E	0.10	-0.57	-0.88	0.04
F	0.14	1.10	-0.43	0.08
G	0.13	-0.26	-0.08	0.01
H	0.07	0.73	1.15	0.05
I	0.06	-1.33	-0.31	0.07
Total activo	1.00			0.35
a		Normalización Simétrica		

encuentran en la columna de puntuación en la dimensión en la tabla de examen de los puntos fila. Tabla 9.

Todo el procedimiento anterior se puede seguir para las marcas a partir de la tabla de perfiles por columna y obtenemos los mismos resultados junto con ***las nuevas coordenadas de las marcas*** para las dimensiones elegidas, que se encuentran en la columna de examen de puntos columna. Tabla 10

Tabla 10

		Puntuación		
	Masa	en la dimensión		Inercia
		1	2	
Marca				
M1	0.27	0.89	-0.01	0.11
M2	0.27	0.08	-0.59	0.06
M3	0.20	-0.12	0.99	0.07
M4	0.25	-0.95	-0.16	0.12
Total activo	1.00			0.35
a	Normalización Simétrica			

Las coordenadas obtenidas para los atributos y las marcas se unen en un mismo plano, esto es posible debido a que existe una correspondencia entre las coordenadas de los atributos y las marcas; es decir, es posible expresar cada coordenada de un atributo en función de la suma de las coordenadas de las marcas y viceversa, donde las contribuciones de las coordenadas de cada marca a las coordenadas de los atributos en esta suma, toman en cuenta la presencia de la marca j en el atributo (renglón correspondiente al atributo en la tabla de perfiles por renglón), de acuerdo con esto *mientras más asociados estén un atributo y una marca*, mayor será la presencia de esta marca en el atributo, mayor será su aportación a las coordenadas del mismo y *ambos estarán más cercanos en el plano*.

El origen en el gráfico está determinado por las medias ponderadas de los atributos y de las marcas, para la ponderación se utiliza la masa (valores en azul en las tablas 3 y 4) con la finalidad de que las categorías con mayor frecuencia tengan mayor influencia en la dirección de los ejes, la razón de esto es que al representar los atributos y las marcas en un espacio de dimensión pequeña los puntos que los representan sufren cierta deformación y al darles un peso de acuerdo a su masa los puntos más importantes se ven menos afectados por dicha deformación.

Los atributos y las marcas que menos discriminan cada una de las dimensiones se encuentran más cercanos al origen, es decir los atributos y las marcas que menos contribuyen a la variabilidad explicada por cada una de las dimensiones estarán próximos al origen, entre más cercanos estén los puntos al origen su desviación con respecto a su media es menor. Sin embargo, ya que el origen esta determinado por la media de cada uno de los atributos y de las marcas y la media de los atributos y de las marcas es una media ponderada por su masa, los atributos o marcas con una mayor masa tendrán mayor influencia en el origen, por lo tanto, las marcas o atributos con una masa pequeña influyen en la variabilidad o inercia solamente cuando estén alejadas del origen y las que tienen masa muy alta influyen en la variabilidad aún cuando no estén tan alejadas del origen.

Tabla 11

TABLA DE EXAMEN DE LOS PUNTOS FILA			
	De la dimensión		
	a la inercia del punto		
ATRIB	1	2	Total
A	0.03	0.00	0.03
B	0.47	0.08	0.55
C	0.81	0.04	0.85
D	0.34	0.66	0.99
E	0.35	0.56	0.91
F	0.90	0.09	0.99
G	0.51	0.03	0.54
H	0.37	0.61	0.97
I	0.64	0.02	0.66

La calidad de la representación de cada uno de los atributos y de las marcas en dos dimensiones se encuentra en la tabla de examen de los puntos de fila y de examen de los puntos de columna en la parte de la contribución de la dimensión a la inercia del punto y debe ser lo más cercana posible a 1.00 o al 100%, en nuestro ejemplo los atributos con la calidad más baja son el atributo A, el atributo B y el atributo G con 3%, 55% y 54% respectivamente, la marca con menor calidad de representación es la marca 2 con 52% . Tablas 11 y 12.

Tabla 12

TABLA DE EXAMEN DE LOS PUNTOS COLUMNA		
De la dimensión		
a la inercia del punto		
1	2	Total
0.86	0.00	0.86
0.01	0.51	0.52
0.02	0.87	0.89
0.87	0.02	0.88

La parte de variabilidad de cada dimensión que es debida a cada uno de los atributos o a cada una de las marcas se encuentra en la columna de contribución de los puntos a la inercia de la dimensión en las tablas de examen de los puntos fila para los atributos y examen de los puntos columna para las marcas.

En nuestros resultados los atributos que más contribuyen en la variabilidad de la primera dimensión son el Atributo F con 37%, el Atributo I con 24%, así como el Atributo C con 12%, para el caso de la segunda dimensión, los atributos que más contribuyen son el Atributo D con 31%, el Atributo E con 31% y el Atributo H con 26%.

Tabla 13

TABLA DE EXAMEN DE LOS PUNTOS FILA	
Contribución De los puntos a la inercia de la dimensión	
1	2
0.00	0.00
0.02	0.01
0.12	0.01
0.07	0.31
0.07	0.26
0.37	0.08
0.02	0.00
0.08	0.31
0.24	0.02
1.00	1.00

En el caso de las marcas las que más contribuyen a la inercia o variabilidad de la primera dimensión son la Marca 4 con 51% y la Marca 1 con 48%, para la segunda dimensión las que más contribuyen son la Marca 3 con 67% y la Marca 2 con 31%.

MEMORIAS

Aún tenemos ejemplares de nuestros seminarios y talleres:

IX Seminario de Actualización Profesional
Septiembre 2002 Ciudad de México

IV Talleres de Investigación de Mercados
Mayo 2002 Ciudad de México

VIII Seminario de Actualización Profesional
Agosto 2002 Ciudad de México

Nuestros teléfonos:
5250-2107 5250-8936 5545-1465
Tel. / Fax: 5254-4210

Tabla 14

TABLA DE EXAMEN DE LOS PUNTOS COLUMNA	
Contribución De los puntos a la inercia de la dimensión	
1	2
0.48	0.00
0.00	0.31
0.01	0.67
0.51	0.02
1.00	1.00

Por último se muestra el gráfico correspondiente a este análisis donde los atributos y las marcas del mismo color son los que están asociados, es importante notar que la distancia euclídea entre los puntos del gráfico (atributos y marcas) aproxima distancias X^2 de la tabla. Ver Gráfica 1 en la página siguiente.

Bibliografía

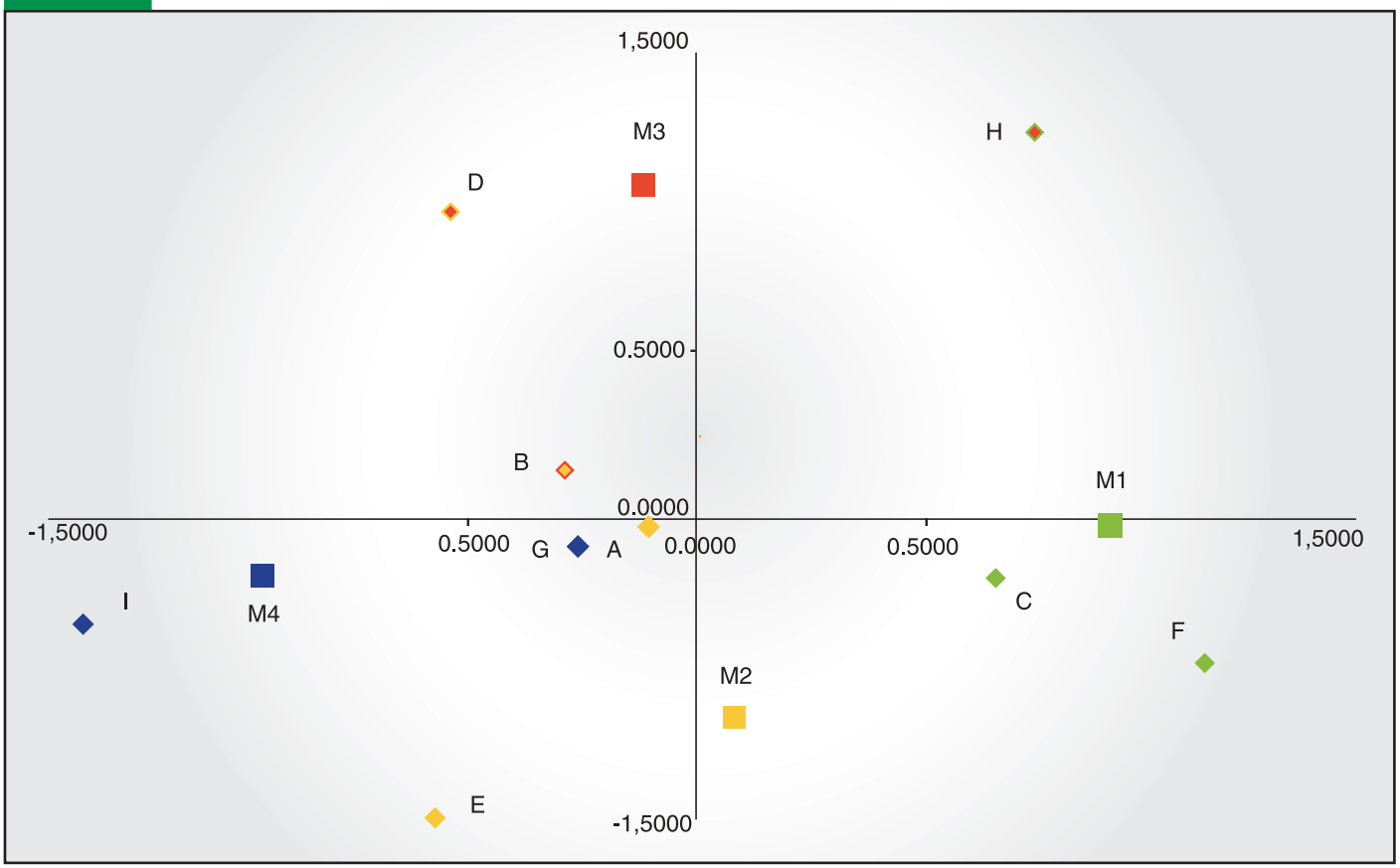
Cuadras Avellana, Carlos. *Métodos de Análisis Multivariante*, Barcelona 1996,644p.

Bendixen, Mike. *A practical guide to the Use of the Correspondence Análisis in Marketing Research*. University of the Witwatersrand. South Africa.

Visauta Vinacua, B. *Análisis Estadístico con SPSS para Windows*, McGraw-Hill Interamericana, España 1998.

Ferrán Aranaz Magdalena. SPSS para Windows Análisis Estadístico. McGraw-Hill Interamericana, España 2001.

Gráfica 1



¡Anúnciese!
en
Datos
Diagnósticos
Tendencias
un medio dirigido a gente como USTED
Ventas: 5545-1465