

**EL HJ-BILOT COMO BASE PARA LA BUSQUEDA DE
CLUSTERS EN FUNCION DE LA DISTRIBUCION DE PARADOS
SEGUN PROFESIONES
EN LA COMUNIDAD DE CASTILLA Y LEON**

**Santiago VICENTE TAVERA Guillermo RAMIREZ
Purificación GALINDO VILLARDON**
*Depto. de Estadística y Matemática Aplicadas.
Universidad de Salamanca*

1.- INTRODUCCION:

En análisis de datos es interesante, por lo general, considerar un modelo que refleje las analogías y diferencias existentes entre los elementos de un conjunto, las cuales son debidas a la observación de varios caracteres. Así, en análisis multivariante se estudian métodos que tienen por objetivo la búsqueda, representación e interpretación de los datos obtenidos a través de varias variables.

Según el espacio geométrico modelo, podemos dividir las representaciones en dos grupos: modelos espaciales o continuos y modelos de redes o discretos. Este trabajo, cuya metodología puede encuadrarse dentro de los modelos discretos, pone de manifiesto que la información resultante de la representación geométrica de una matriz de datos multivariante puede ser tomada como base para la obtención de una Clasificación Ascendente cuyas clases, si bien pueden no verificar las propiedades teóricas óptimas de compacidad y/o separabilidad, proporcionan clases fácilmente interpretables, para las cuales podemos conocer las variables responsables de esa configuración.

Utilizaremos como base la información suministrada por la representación simultánea **HJ-BILOT** en la línea propuesta por GALINDO (1986) y por GALINDO y CUADRAS (1987) por ser una técnica mas general que el Análisis Factorial de Correspondencias BENZECRI (1973).

El criterio utilizado es una extensión del Criterio de la Inercia propuesto por Benzècri (1985), el cual puede ser consultado con detalle en VICENTE TAVERA (1992)

2.- MATERIAL Y METODOLOGIA:

El objetivo del trabajo consiste en analizar las posibles agrupaciones (Clusters), de las provincias de Castilla León en función de la distribución de parados que han trabajado antes y atendiendo a las diferentes profesiones que han tenido.

Mediante la aplicación de este criterio podremos determinar cuales son las variables responsables de las diferentes agrupaciones.

2.1.- Criterio De Clasificación Ascendente Jerárquica Basado En La Representación HJ-BIPLLOT:

El criterio se sintetiza en los siguientes pasos:

Sea B el conjunto de puntos representados según la HJ-bigeometría a clasificar.

Definimos:

INERCIA DENTRO $I_D(P_k)$ de una partición P_k sobre B a la suma de las **INERCIAS INTERNAS** de las k clases que forman dicha partición.

La **INERCIA INTERNA** de una clase S_i^k formada por n_i elementos es la suma de las distancias al cuadrado de cada elemento S_{ij}^k al dentro de gravedad $S_{i\bullet}^k$ de la clase.

$$I(S_i^k) = \sum_{j=1}^{n_i} \|s_{ij}^k - s_{i\bullet}^k\|^2 = \sum_{j=1}^{n_i} \sum_{\alpha} [F_{\alpha}(s_{ij}^k) - F_{\alpha}(s_{i\bullet}^k)]^2$$

donde $F_{\alpha}(S_{i\bullet}^k)$ y $F_{\alpha}(S_{ij}^k)$ son las coordenadas del centro de gravedad de la clase S_i^k y de los elementos S_{ij}^k de la misma respectivamente.

luego:

$$I_D(P_k) = \sum_{i=1}^k I(S_i^k)$$

La INERCIA ENTRE $I_E(P_k)$ las k clases definidas en la partición P_k sobre el conjunto B es la inercia de los k centros de gravedad de las clases ponderadas por sus tamaños.

$$I_E(P_k) = \sum_{i=1}^k n_i \|s_{i\bullet}^k - O\|^2 = \sum_{i=1}^k n_i \sum_{\alpha} [F_{\alpha}(s_{i\bullet}^k) - F_{\alpha}(O)]^2$$

donde $F_{\alpha}(s_{i\bullet}^k)$ y $F_{\alpha}(O)$ son las coordenadas del centro de gravedad de la clase S_{ij}^k y del centro de gravedad O de B .

n_i es el número de elementos de la clase S_i^k

Existe una perfecta complementariedad entre estos tres tipos de inercia de tal manera que:

$$I_T(P_k) = I_D(P_k) + I_E(P_k)$$

2.1.1.- Criterio De Clasificación:

El paso de una partición P_k formada por k clases a otra P_{k-1} en la que hemos agrupado las clases S_t^k y S_v^k y formado la clase $(S_t^k \cup S_v^k)$ se hace tal que:

$$P_{k-1} = [P_k - \{S_t^k, S_v^k\}] \cup \{S_t^k \cup S_v^k\}$$

Basándonos en las definiciones anteriores se deduce:

$$I_D(P_{k-1}) = I_D(P_k) + I_E(\{S_t^k, S_v^k\})$$

La **Inercia Dentro** en la partición P_{k-1} depende solamente de la **Inercia Entre** las clases que hemos añadido y no del resto de las clases, de aquí que el criterio a seguir sea **MINIMIZAR ÉSTA INERCIA ENTRE LAS CLASES QUE AGRUPAMOS**.

Es fácil demostrar que minimizar la inercia entre las clases S_t^k y S_v^k es equivalente a minimizar la siguiente relación:

$$I_E(\{S_t^k, S_v^k\}) = n_t \left\{ \sum_{\alpha=1}^q [F_{\alpha}(S_{t\bullet}^k) - F_{\alpha}(O_{tv})]^2 \right\} + n_v \left\{ \sum_{\alpha=1}^q [F_{\alpha}(S_{v\bullet}^k) - F_{\alpha}(O_{tv})]^2 \right\}$$

donde $F_{\alpha}(O_{tv})$ son las coordenadas del centro de gravedad de la clase agrupada, n_t y n_v el número de elementos de las respectivas clases y q el número de factores retenidos para la clasificación en la representación **HJ-BIPLLOT**.

Este criterio de inercia lo podemos analizar desde un punto de vista local ya que nos mide la dispersión entre dos clase S_t^k y S_v^k ; pero también tiene un análisis global si nos fijamos en las sucesivas particiones que se hacen en el conjunto B .

El criterio tiene la ventaja de propiciar particiones que tienen clases homogéneas y bien separadas unas de otras. Homogéneas, puesto que la agregación se hace de manera que la inercia alrededor del centro de la clase es lo más pequeña posible y esto es así para todas las clases de la partición, por tanto la "inercia dentro" de la partición es baja. Bien separadas, puesto que las nubes de los centros de gravedad de las clases de una partición está dispersa ya que la "inercia entre" es grande.

Cada nudo del árbol, que se forma con éste criterio, incorpora un nivel equivalente a la inercia acumulada hasta el mismo. Efectuando la suma de los niveles de los distintos nudos de un árbol se obtiene la inercia total de la nube.

Si denotamos con $D(n)$ el nivel del nudo n que agrega las clases S_t^k y S_v^k ; obviamente $D(n)$ representa la contribución del nudo n a la inercia total de la nube $N(I)$.

$D(n) / \text{Inercia total de la nube}$, representará la contribución relativa del nudo a la inercia total de la nube $N(I)$.

De la misma manera se calculan las correspondientes contribuciones para los diversos nudos del árbol.

Sea S_i^k una clase definida en un árbol de clasificación ascendente, y sea S_i^k el centro de gravedad de dicha clase. Es posible añadir este nuevo elemento al análisis **BILOT** pero sin ser considerado como elemento de la batería activa. Este elemento suplementario puede ser representado en el espacio de los ejes factoriales y obtener así sus coordenadas sobre los ejes factoriales. Así es posible cuantificar la **contribución (absoluta o relativa) de cada eje factorial al nudo** representante de cada clase y **de cada nudo a los ejes**, de la misma forma en que se hace para cualquier otro punto proyectado sobre el subespacio.

Cuando se lleva a cabo la representación gráfica de una clasificación en forma de árbol se evidencian los Cluster latentes en la estructura de los datos y es posible una definición de los mismos en términos de extensión o enumeración de los elementos que lo componen. Cuando el árbol ha sido construido por el procedimiento de la inercia sobre un análisis **HJ-BILOT** es posible descubrir una definición en términos de comprensión, es decir, enumerando las cualidades propias de los individuos que pertenecen a la clase.

En el caso de tablas pequeñas no resultaría difícil reconocer las características que definen los diversos nudos del árbol, pero para grandes matrices de datos, es imprescindible, para conocer que caracteres están en juego, llevar a cabo un exhaustivo estudio de las listas de contribuciones que proporciona el análisis BIPLLOT.

2.1.2.- Algoritmo De Clasificación:

Tomando como base las coordenadas **HJ-Biplot** de los individuos sobre los primeros **q** factores, se procede de la siguiente manera:

1.- Se construye una matriz inicial de distancias entre individuos:

$$d^2(i_1, i_2) = \sum_{\alpha=1}^q \left[F_{\alpha}(S_{i_1}) - F_{\alpha}(O_{i_1, i_2}) \right]^2 + \sum_{\alpha=1}^q \left[F_{\alpha}(S_{i_2}) - F_{\alpha}(O_{i_1, i_2}) \right]^2$$

2.- Se agrupan aquellos individuos entre los cuales haya una distancia mínima.

3.- Se recalculan, en cada etapa, las distancias entre el grupo formado y el resto, por medio de:

$$d^2(S_t, S_v) = I_E(\{S_t, S_v\})$$

4.- Se repite el procedimiento desde 2 hasta que se hayan agrupado todos los individuos.

2.2.- Descripción De Los Datos Analizados:

Los datos hacen referencia a la distribución de parados en Castilla León habiendo trabajado antes según provincia y profesión. Los datos han sido obtenidos del Censo de Población y Viviendas 1991. Muestra Avance. Principales Resultados. INE. Madrid. 1992.

Tabla 4, pág 110.

Tabla 19, pág. 169

Descripción de los códigos de profesión en Anexo 2, pág. 45.

Se han establecido los siguientes códigos para las provincias y para las profesiones:

Provincia	Código	Profesión	Código
1.- Avila	AV	1.- Prof. Tit. Ciencias	TC
2.- Burgos	BU	2.- Prof. Aux. Ciencias	AC
3.- León	LE	3.- Artistas.Deportistas	AD
4.- Palencia	PA	4.- Personal Docente	PD
5.- Salamanca	SA	5.- Directivos	DI
6.- Segovia	SG	6.- Personal Admintvo	PA
7.- Soria	SO	7.- Comerciantes	CO
8.- Valladolid	VA	8.- Servicios Hostelerías	SH
9.- Zamora	ZA	9.- Servicios Prot. y Seg.	SG
		10.-Otros Servicios	OS
		11.-Agric., Ganad, Pescad.	AG
		12.-Jefes de Taller	JT
		13.-Construcción	CT
		14.-Metalurgia	MT
		15.-Text., Alim. y Artes Graf.	TA
		16.-Operadores de.Máquinas	OM
		17.-Peones	PN
		18.-Fuerzas Armadas	FA
		19.-No Consta	NC

Tabla 1: Asignación de códigos para las provincias y profesiones para el análisis multivariante posterior

2.3 Análisis General:

Cada valor de la matriz de datos de 9 filas y 19 columnas representa la tasa por mil habitantes de los parados en Castilla León habiendo trabajado antes según provincia y profesión. Sobre esta matriz se llevó a cabo la técnica de representación simultánea **HJ-BIPLLOT** sin dividir por las medias y haciendo doble centrado, obteniéndose los siguientes resultados en las tablas 2,3 y 4.

Valores Propios	% de Inercia explicada	% de Inercia acumulada
50.9343	57.46	57.46
10.7821	12.16	69.63
10.3697	11.70	81.33
7.4674	8.42	89.75
6.3151	7.12	96.88
2.0257	2.29	99.16
0.4587	0.52	99.68
0.2816	0.32	100.00
0.0017	0.00	100.00

TABLA 2: Valores propios, Tasas de inercia explicada y acumuladas en los 9 Factores

Como podemos apreciar en la Tabla 2, existe un eje dominante que absorbe el 57.46% de la inercia total del sistema. La tendencia de los valores propios se trunca en el quinto eje (codo) alcanzando estos, una inercia acumulada del 96.88%. El resto de factores apenas

disponen de información por lo que decidimos retener los cinco primeros factores para realizar nuestra clasificación.

Al analizar las Contribuciones Relativas del Factor al Elemento, sobre un máximo de 1000, tanto para las diferentes profesiones como para las provincias, Tablas 3 y 4, vemos como efectivamente los cinco ejes aportan suficiente información para explicar tanto las variables como las poblaciones.

CONTRIBUCIONES RELATIVAS DEL FACTOR AL ELEMENTO

	EJE1	EJE2	EJE3	EJE4	EJE5		EJE1	EJE2	EJE3	EJE4	EJE5
TC	950	10	2	1	27	AC	820	27	16	63	7
AD	930	16	14	5	1	PD	680	4	28	10	35
DI	960	7	0	2	17	PA	3	150	22	28	338
CO	540	370	63	1	0	SH	400	19	97	27	372
SG	790	8	60	50	85	OS	99	260	420	14	20
AG	270	690	18	0	17	JT	860	4	37	8	29
CT	320	110	0	135	325	MT	220	30	47	604	70
TA	1	110	390	452	43	OM	45	22	37	164	389
PN	930	17	40	2	15	FA	920	7	38	14	0
NC	180	160	480	156	29						

TABLA 3: Contribuciones Relativas del Factor al Elemento para las diferentes profesiones en los 5 primeros ejes factoriales.

CONTRIBUCIONES RELATIVAS DEL FACTOR AL ELEMENTO

	EJE1	EJE2	EJE3	EJE4	EJE5		EJE1	EJE2	EJE3	EJE4	EJE5
AV	340	4	590	42	0	BU	220	71	220	423	39
LE	2	16	1	891	6	PA	560	2	4	63	242
SA	570	64	170	65	122	SG	880	41	0	14	51
SO	670	110	150	32	31	VA	240	380	29	44	127
ZA	500	400	16	13	62						

TABLA 4: Contribuciones Relativas del Factor al Elemento para las diferentes provincias en los 5 primeros ejes factoriales

Al analizar las contribuciones anteriores, desglosadas por los diferentes ejes, apreciamos como el primer eje viene explicado por la mayoría de las diferentes profesiones, nueve de las mismas (DI, TC, AD, PN, FA, JT, AC, SG y PD) poseen Contribuciones Relativas del Factor al Elemento comprendidas entre 960 (DI) y 680 (PD). El segundo eje factorial está casi exclusivamente dominado por la variable AG la cual tiene una contribución de 690, las dos siguientes variables con mayor contribución (CO y OS) tienen mayor contribución en los ejes 1 y 3 respectivamente. El tercer eje está explicado por las variables NC, OS y TA aunque esta última tiene mayor importancia en el eje 4 donde la variable MT con una contribución de 604 es fundamental en el mismo. El quinto eje tiene menor importancia explicativa pero en él aparecen el resto de variables que no han sido

explicadas por los factores anteriores como son OM, SH, PA y CT con contribuciones que oscilan entre 389 y 325 respectivamente.

Al analizar la provincias, vemos como sus contribuciones están mucho mas dispersas entre los cinco ejes, de tal manera que vemos como Avila tiene mayor contribución en el eje 3 (590) y en el primero (340) ; Burgos reparte sus contribuciones entre los ejes 4 (423),1 (220) y 3 (220) respectivamente; León está perfectamente determinada en el eje 4 donde tiene una contribución de 981; Palencia tiene su mayor representatividad en los ejes 1 (560) y 5 (242); Salamanca queda claramente definida en el primer eje (570) teniendo relativa importancia en los ejes 3 y 5; Segovia es claramente una provincia de eje 1 al tener una contribución al mismo de 880; Soria es una provincia que viene bien explicada en el primer eje (670) y en menor medida en los ejes 3 y 2 respectivamente; Valladolid es la provincia que mas diversificación tiene en las contribuciones estando repartidas estas entre los ejes 2, 1 y 5 respectivamente y por último Zamora viene explicada por los dos primeros ejes factoriales con contribuciones de 500 y 400 respectivamente.

Tomando como referencia las coordenadas de las provincias en los cinco factores retenidos se aplicó la técnica de Clasificación Ascendente Jerárquica mediante el Criterio de Inercia (VICENTE-TAVERA, S. 1992) obteniéndose la siguiente clasificación (ver Figura 1). En la clasificación podemos apreciar 4 Clusters claramente diferenciados, el primero formado por las provincias de León, Valladolid y Zamora, el segundo por Burgos y Avila, el tercero por Segovia y Soria y el último por Salamanca y Zamora.

METODO DE INERCIA CON REPRESENTACION HJ-BIPLLOT

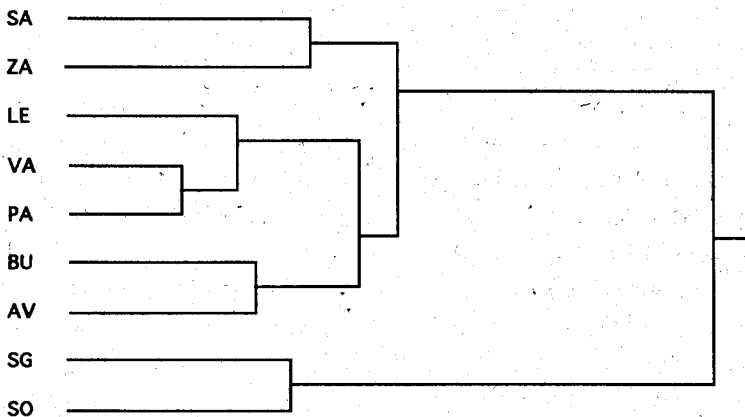


Figura 1: Dendograma de las provincias de Castilla León mediante el Criterio de Inercia reteniendo 5 Factores de la representación HJ-BIPLLOT.

Una de las principales ventajas de este método consiste en determinar cuales son las variables responsables de la configuración de los diferentes Clusters. En tal sentido, intentamos determinar las diferentes configuraciones que los Clusters tienen en los diferentes planos factoriales, analizando especialmente las Contribuciones Relativas del Factor al Elemento de las variables que mayor incidencia tengan en la formación de los Clusters.

2.4 Análisis Del Primer Cluster:

			PROFESIONES			
			PLANOS			
PROVINCIAS	EJES	NIVEL	1-2	1-3	1-4	1-5
León	4		CO		MT	
Valladolid	2, 1, 5		(910)		(804)	
Palencia	1, 5		AG			
			(960)			
			OS			
			(359)			

TABLA 5: *Provincias que forman el primer Cluster de clasificación junto con las principales variables que intervienen en la configuración en los diferentes planos factoriales.*

El primer Cluster, (ver Tabla 5) está formado por las provincias de León, Valladolid y Palencia, se caracteriza por tener una alta tasa de paro en la variable CO (Comerciantes) especialmente en Valladolid siendo la segunda mas alta de todas las provincias, por contra en la variable AG (Agricultura) la tasa es especialmente pequeña en Valladolid, ligeramente superior en León y normal en Palencia. La tercera variable en importancia en el primer plano factorial, pero con menor calidad de representación (359), es OS (Otros Servicios) que efectivamente se caracteriza por tomar los valores mas altos, en las tres provincias, después de Salamanca. El siguiente plano, en el que encontramos información sobre este Cluster, es el 1-4 donde la variable MT (Metalurgia) tiene una relevancia especial puesto que junto con la alta calidad de representación (804) podemos apreciar que la tasa de paro, en esta variable, es la mas alta de todas las provincias. Como podemos apreciar la variables que mas importancia tienen en la configuración del Cluster es la variable MT en el plano 1-4. En la Figura 2 podemos apreciar como efectivamente dicha variable tiene una especial relevancia; las otras dos variables CT y OM tienen menor calidad de representación.

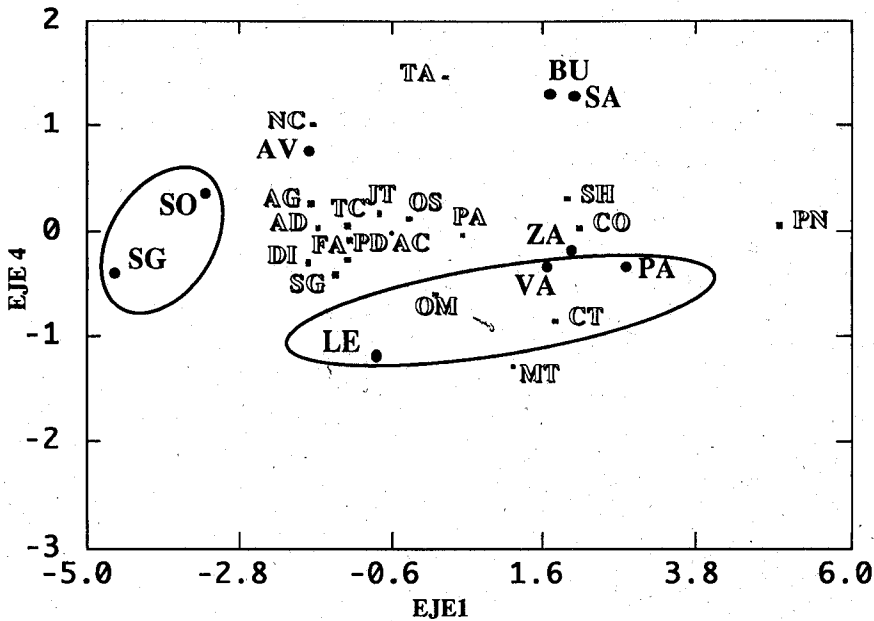


Figura 2: Representación HJ-BIPLLOT en el plano 1-4 de la tasa de desempleados según profesiones habiendo trabajado antes.

2.5 Análisis Del Segundo Cluster:

			PROFESIONES			
			PLANOS			
PROVINCIAS	EJES	NIVEL	1-2	1-3	1-4	1-5
Avila	3,1			NC	NC	
Burgos	4,1,3			(660)	(644)	

TABLA 6: Provincias que forman el segundo Cluster de clasificación junto con las principales variables que intervienen en la configuración en los diferentes planos factoriales.

El segundo Cluster compuesto por las provincias de Avila y de Burgos se forma exclusivamente por la importancia que toman estas provincias en la variable NC (No Consta) donde efectivamente las dos tienen las tasas mas altas, claramente diferenciadas del resto. En la Figura 3 podemos apreciar como efectivamente esta variables es la de mayor relevancia en dicho Cluster.

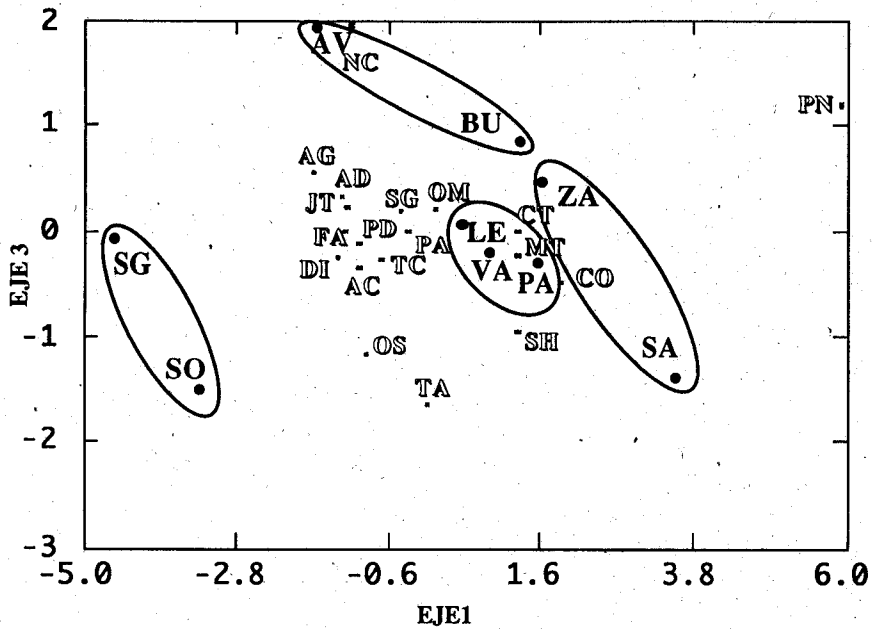


Figura 3: Representación HJ-BIPLLOT en el plano 1-3 de la tasa de desempleados según profesiones habiendo trabajado antes.

2.6 Análisis Del Tercer Cluster:

			PROFESIONES			
			PLANOS			
PROVINCIAS	EJES	NIVEL	1-2	1-3	1-4	1-5
Segovia	1,2		PN			
Soria	1,2,3		(947)			
			FA			
			(938)			
			JT			
			(864)			
			SG			
			(798)			
			TC			
			(952)			
			AC			
			(847)			

TABLA 7: Provincias que forman el tercer Cluster de clasificación junto con las principales variables que intervienen en la configuración en los diferentes planos factoriales.

El tercer Cluster formado por las provincias de Segovia y de Soria queda perfectamente determinado por el primer eje factorial. En la Tabla 7 se pueden apreciar cuales son las variables diferenciadoras de las dos provincias y cuya calidad de representación es muy alta. En las variables PN (Peones), FA (Fuerzas Armadas), JT (Jefes de Taller), y SG (Servicios Prot. y Seg.) las tasa de paro, en ambas provincias son las mas pequeñas de Castilla León; por contra en las variables TC (Prof. Tit. Ciencias) y AC (Prof. Aux. Ciencias) se caracterizan por tener las tasas mas altas en Segovia y relativamente altas en Soria. Otro tipo de variables como pudieran ser DI (Directivos) y AD (Artistas.Deportistas) con altas calidades de representación no tienen excesiva relevancia en el Cluster. En la Figura 4 podemos apreciar el Cluster con las variables que mas intervienen en su configuración.

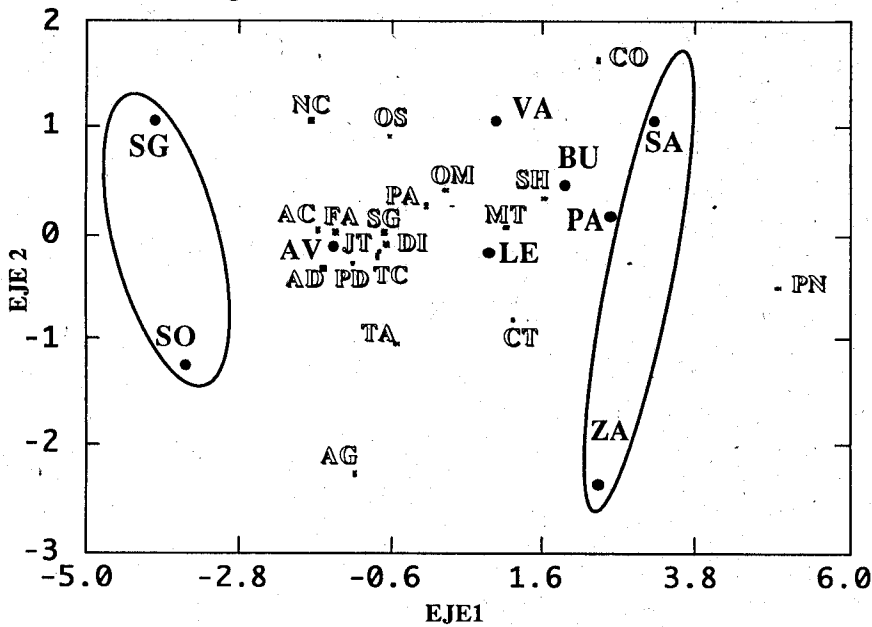


Figura 4: Representación HJ-BIPLLOT en el plano 1-2 de la tasa de desempleados según profesiones habiendo trabajado antes.

2.7 Análisis Del Cuarto Cluster:

			PROFESIONES			
			PLANOS			
PROVINCIAS	EJES	NIVEL	1-2	1-3	1-4	1-5
Salamanca	1,3,5		PN	PN		SH
Zamora	1,2		(947)	(970)		(772)
				OS		CT
				(519)		(655)
				TA		
				(391)		

TABLA 8: Provincias que forman el cuarto Cluster de clasificación junto con las principales variables que intervienen en la configuración en los diferentes planos factoriales.

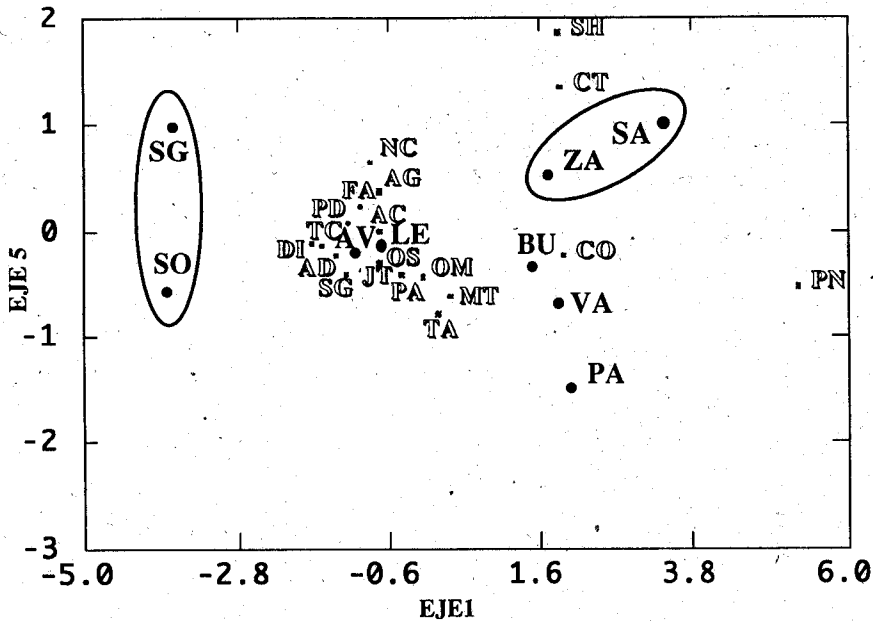


Figura 5: Representación HJ-BIPLLOT en el plano 1-5 de la tasa de desempleados según profesiones habiendo trabajado antes.

El último Cluster por analizar es el formado por Salamanca y Zamora el cual es el de mayor nivel y junto con el tercero es el que mejor representado queda en los diferentes planos. Este Cluster tiene una variable PN (Peones) dominante sobre el resto, la cual es claramente de eje 1; las dos provincias tienen las tasas más altas en contraposición al Cluster anterior (Segovia y Soria) que tenían las tasas más bajas en dicha variable. En el

plano 1-3 conviene citar las variables OS (Otros Servicios) y TA (Text., Alim. y Artes Graf.) especialmente en Salamanca puesto que esta provincia tiene las tasas mas altas en las dos variables. En el plano 1-5 es donde realmente encontramos las variables diferenciadoras del resto de los demás Clusters; las variables SH (Servicios Hostelerías) y CT (Construcción) tienen tasas claramente mas elevadas que el resto de las provincias. Estas dos variables junto con la PN (Peones) son las variables que determinan la configuración del Cluster. En las figuras 3 y 5 podemos apreciar este Cluster junto con las variables anteriormente citadas.

BIBLIOGRAFIA CONSULTADA

ARCAS, A. y CUADRAS, C.M.(1987): 'Métodos geométricos de representación mediante modelos en arbol'. Public. de Bioest. y Biomatem. nº 20. Univ. Barcelona.

BENZECRI, J.P. (1973/82): L'Analyse des données. Tomo II. L'Analyse des correspondences. Dunod, Paris. 4ª Ed.

EVERITT, B.S. (1993): Cluster Analysis. Edward arnold. London.

GABRIEL K.R. (1971): ' The Biplot-graphic display of matrices with application to Principal Component Analysis'. Biometrika 58: 453-467

GABRIEL K.R. (1981): 'Biplot display of multivariate for inspection of data and diagnosis'. In V. Barnett Ed. , Interpreting Multivariate Data : 147-174

GALINDO, M.P. (1986): 'Una alternativa de representación simultanea: HJ-Biplot. Qüestiió . nº1:13-23

GALINDO, M.P. y CUADRAS, C.M. (1986): 'Una Extensión del Método Biplot y su Relación con Otras Técnicas'. Public. de Bioestadística y Biomatemática. Nº 17. Univ. de Barcelona.

GOWER, J.C. (1988): 'Classification, geometry and data analysis' (in Classification and related method of data analysis.). H.H. Bock Ed. North Holland.

KAUFMAN, L. and ROUSSEEUW, P.J. (1989): ' Finding groups in data. An introduction to Cluster Analysis'. John Wiley and Sons, Inc. New York.

VICENTE TAVERA, S. (1992): 'Las técnicas de representación de datos multidimensionales en el estudio del índice de producción industrial en la C.E.E'. Tesis Doctoral. Salamanca.