LA ESTABILIDAD DE LOS RESULTADOS EN EL ANÁLISIS DE CORRESPONDENCIAS. APLICACIÓN AL ESTUDIO DEL MERCADO DE TRABAJO

Ramón ÁLVAREZ ESTEBAN
Facultad CC.Económicas y Empresariales. Universidad de León

1.- INTRODUCCIÓN

En el proceso de modelización del análisis económico regional debiéramos plantearnos algunos aspectos a la hora de la elaboración del modelo y la interpretación de los resultados, como puede ser la validez interna, la validez externa y la estabilidad.

Frecuentemente nos cuestionamos si la selección de otra técnica estadística, modelo o algoritmo confirmarán o refutarán nuestras conclusiones. El objetivo estaría, entonces, encaminado a buscar una técnica óptima. Este planteamiento de la búsqueda del modelo verdadero ha estado ampliamente tratado en la literatura al intentar explicar la preferencia de una nueva técnica frente a las existentes, olvidando en muchos casos que también la propia naturaleza y estructura de los datos condicionan el tratamiento estadístico utilizado.

La investigación desarrollada bajo esta concepción tiene el grave peligro de ignorar los datos, el origen de la información que deseamos procesar. Sin rechazar este enfoque, debiéramos partir inicialmente del estudio de la adecuación de la técnica de análisis de datos elegida para la consecución de los objetivos planteados, teniendo en cuenta el peligro que supone desplazar la importancia que poseen los datos a la técnica en sí.

Con la aparición de ordenadores cada vez más potentes las técnicas de análisis de datos parecen estar incompletas sin el profundo estudio y análisis de la estabilidad de los resultados obtenidos, comprobando si pequeñas modificaciones de los datos pueden afectar a un modelo hasta el punto de llevarnos a reconsiderar, e incluso rechazar, la interpretación dada y, por lo tanto, las conclusiones. El problema es similar a la construcción de un edificio, si no resiste pequeños temblores de tierra no se considera apto para vivir.

2.- OBJETIVOS Y METODOLOGÍA

Este proceso de validación, estrechamente vinculado a la sensibilidad y a los cálculos de estabilidad, ha estado generalmente dirigido hacia la construcción de modelos

bajo la hipótesis de determinados supuestos como la normalidad, igualdad de varianzas, linealidad o independencia.

La falta del cumplimiento de estas hipótesis o la dificultad de contrastarlas, así como la construcción de modelos cada vez más complejos requiere la necesidad de considerar técnicas más flexibles que integren este problema dentro del análisis de datos tradicional. La elección de variables, la ponderación que se da a cada una, el peso de cada individuo, la determinación de valores extremos y la estabilidad de las dimensiones y agrupaciones son algunos ejemplos de los factores que pueden afectar y condicionar los resultados.

Por otro lado, la complejidad de los análisis de validación ha supuesto en muchos casos que los estudios dentro de una determinada técnica de análisis de datos hayan sido parciales o fragmentados, dirigidos hacia algún aspecto en concreto, como puede ser la estabilidad de los valores propios para elegir el número adecuado de ellos frente a los criterios tradicionales.

Las técnicas de simulación (Validación Cruzada, Jackknife y Bootstrap, entre otras), han mostrado ser una ayuda en el proceso de cuantificación de la sensibilidad y estabilidad al no estar sujetas a hipótesis previas sobre la naturaleza de los datos.

Nuestro estudio parte del análisis de una tabla de contingencia con la estructura de los trabajadores por ramas de actividad y Comunidades Autónomas. Sobre esta tabla se ha realizado un Análisis de Correspondencias. Mediante un programa informático construido al efecto, se han simulado mil tablas mediante remuestreo, extracción con reemplazamiento y sin suponer las limitaciones de independencia de otros trabajos, proceso que Efron y Tibshirani denominan “bootstrap no paramétrico” al no existir una distribución empírica que deba suponerse o contrastarse. El proceso de bootstrap de Efron puede resumirse de la siguiente forma:

1. Partiendo de la tabla inicial se obtienen los estimadores.
2. Se construyen m muestras de tamaño n mediante extracción aleatoria con reemplazamiento de la tabla inicial, recomendándose que m sea al menos de 250, aunque en muchos análisis estadísticos suele situarse alrededor de 30.
3. Se utilizan las muestras secundarias generadas para construir los estimadores simulados. Habrá, por lo tanto, tantos valores simulados de un estimador como muestras generadas.
4. Se calcula un estimador de los m estimadores creados por simulación.
5. Se calcula el intervalo de confianza del estimador obtenido en el paso anterior, generalmente mediante el cálculo de percentiles.

Para cada una de estas tablas se obtienen los valores propios, vectores propios, coordenadas, contribuciones absolutas y relativas. De la comparación de los resultados

simulados y los resultados de la tabla original se establecen los intervalos de confianza, las zonas de confianza de las representaciones gráficas, la estabilidad de los factores y el número de ellos que deben ser retenidos.

2.1.- Análisis de los valores propios

A partir de transformaciones de la tabla de contingencia se obtiene la matriz de varianzas covarianzas, simétrica semidefinida positiva de orden igual al menor número entre filas (r) y columnas (s), siendo las frecuencias relativas p_{ij}, y las marginales $p_{i.}$ y $p_{.j}$:

$$S_{jj'} = \sum_{i=1}^{r} \frac{p_{ij}p_{ij'}}{p_{i.}\sqrt{p_{.j}}\sqrt{p_{.j'}}} - \sqrt{p_{.j}}\sqrt{p_{.j'}}$$

si número de filas es mayor que el de columnas

$$S_{ii'} = \sum_{j=1}^{s} \frac{p_{ij}p_{i'j}}{p_{.j}\sqrt{p_{i.}}\sqrt{p_{i.'}}} - \sqrt{p_{i.}}\sqrt{p_{i.'}}$$

si número de filas es menor que el de columnas.

La reducción del espacio vectorial sobre estas matrices proporciona los valores y vectores propios.

El problema de la elección del número de ejes o factores que deben conservarse puede resumirse en cuatro procedimientos: reglas empíricas (Cattell se basa en las variaciones o saltos entre un valor propio y el siguiente, y Kaiser quien retiene los valores propios por encima de su media), procedimientos externos (utilizando variables externas para confirmar que los factores son reales y no aleatorios), estudios asintóticos (basados en la suposición de que los valores propios siguen una determinada distribución) y estudios de estabilidad mediante simulación.

La hipótesis de independencia entre filas y columnas de la tabla de contingencia es la base de la mayor parte de los estudios asintóticos y se ha utilizado para definir los intervalos de confianza o niveles de significación de los valores propios, demostrando que para el análisis de correspondencias éstos siguen leyes no paramétricas, obteniendo la normalidad como resultado de la convergencia de la ley multinomial hacia la ley normal, siempre que las frecuencias de las celdas sean suficientemente grandes, pudiendo ser aproximados por una matriz de Wishart. Esta suposición de independencia generalmente es irreal y provoca una fuerte relación entre los valores propios, correlación positiva entre valores propios cercanos y negativa entre los valores propios alejados.

Este enfoque suele encontrarse relacionado con los modelos logarítmicos lineales y la reconstrucción de la tabla original a partir de los valores y vectores propios, teniendo

el grave inconveniente de que deben ser utilizados con muchas precauciones, ya que la ley de los valores propios sólo permite establecer la significación del primer eje, debido a que el resto de los ejes están condicionados a él y por lo tanto sus leyes, también condicionadas, son desconocidas.

Actualmente los programas de análisis de correspondencias están incorporando el "método delta", un método clásico de estadística asintótica que proporciona la dispersión de valores propios y modalidades.

Los estudios de los valores propios mediante simulación abarcan varias técnicas, como la validación cruzada que determina el porcentaje de clasificación correcta, el Jackknife y el bootstrap, quedando el método de Monte Carlo generalmente reservado para estudios sobre la hipótesis de independencia.

Este análisis de los valores propios obtenidos mediante simulación nos plantea una duda inicial ¿es posible comparar valores propios entre sí cuando corresponden a factores diferentes?. En ocasiones se supone que al simular una tabla el primer valor propio simulado puede ser representado en forma de histograma, comprobando la normalidad y obteniendo el intervalo de confianza a partir de los percentiles. Esto, como trataremos posteriormente sólo será cierto si el primer factor se mantiene a lo largo de todas las observaciones. Por otro lado, parece demostrado que la mayor parte de la variación de los valores propios no depende de la variación de los vectores propios, por lo que no parece prudente definir un subespacio exclusivamente a partir de los valores propios.

2.2.- Análisis de la estabilidad de las coordenadas

La representación de las diferentes modalidades simuladas dentro del plano factorial no plantea ningún problema al poder utilizar elementos suplementarios, bien filas o columnas, una vez conocidos los valores y vectores propios de la tabla original. Teniendo en cuenta que las coordenadas de los puntos fila y de los puntos columna son respectivamente

$$\hat{\Psi}_{\alpha i} = \sum_{j=1}^{q} \left(\frac{p_{ij}}{p_{i.}\sqrt{p_{.j}}} \right) u_{\alpha j} \qquad \hat{\Phi}_{\alpha j} = \sum_{i=1}^{n} \left(\frac{p_{ij}}{p_{.j}\sqrt{p_{i.}}} \right) v_{\alpha i}$$

siendo u_{aj} y v_{ai} los vectores propios, es posible expresar las coordenadas fila en función de las coordenadas columna:

$$\hat{\Psi}_{ai} = \sum_{j=1}^{q} \left(\frac{p_{ij}}{p_{i.}\sqrt{p_{.j}}} \right) \frac{1}{\sqrt{l}_a} \hat{\Phi}_{aj} \sqrt{p_{.j}} = \frac{1}{\sqrt{l}_a} \sum_{j=1}^{q} \frac{p_{ij}}{p_{i.}} \hat{\Phi}_{aj}$$

Y las coordenadas columna en función de las filas:

$$\hat{\Phi}_{aj} = \sum_{i=1}^{n} \left(\frac{p_{ij}}{p_{.j}\sqrt{p_{i.}}} \right) \frac{1}{\sqrt{l}_a} \hat{\Psi}_{ai} \sqrt{p_{i.}} = \frac{1}{\sqrt{l}_a} \sum_{i=1}^{n} \frac{p_{ij}}{p_{.j}} \hat{\Psi}_{ai}$$

Las nuevas coordenadas de una fila considerada como ilustrativa se determinan a partir de las coordenadas columna según la siguiente expresión:

$$\hat{\Psi}_{ai} = \frac{1}{\sqrt{l}_a} \sum_{j=1}^{q} \frac{p_{i*j}}{p_{i*.}} \hat{\Phi}_{aj}$$

Siendo p_{i*j} las frecuencias relativas suplementarias y $p_{i*.}$ la marginal.

Para el caso de nuevas columnas utilizaríamos: $\hat{\Phi}_{aj} = \frac{1}{\sqrt{l}_a} \sum_{i=1}^{n} \frac{p_{ij*}}{p_{.j*}} \hat{\Psi}_{ai}$

Los elipsoides así obtenidos mediante simulación son un indicador de la estabilidad de la modalidad en el factor y en el plano. Grandes elipsoides son un signo de pequeña estabilidad, lo que nos debiera llevar a reconsiderar la interpretación de la modalidad en el plano. Por otro lado, al representar ejes con menores valores propios es fácilmente constatable que las dispersiones aumentan, lo que hace que sean inestables y deban no ser considerados en el estudio.

2.3.- Análisis de la estabilidad de los factores

La duda sobre si es posible comparar valores propios entre sí cuando corresponden a factores diferentes no queda resuelta con la proyección de las filas y columnas simuladas sobre los factores iniciales.

Al simular la tabla inicial pueden cambian los valores propios, pero también pueden hacerlo los vectores propios. Si un factor desaparece o aparece intermitentemente en las simulaciones es un claro signo de inestabilidad, pero puede ocurrir que, por

ejemplo, el factor uno pase en ocasiones a intercambiarse con el dos debido a que tienen valores propios similares.

Este problema no tiene, hasta la fecha, una solución claramente definida y los enfoques pueden dirigirse en varias direcciones, como son el estudio de las variaciones en los resultados por simulación en las coordenadas, las contribuciones absolutas y las relativas.

En relación a las coordenadas, la comparación de coordenadas entre un factor i original y los m factores i simulados deben plantear varias consideraciones. En primer lugar la orientación del eje, ya que si cambiamos el signo de todas las coordenadas de un eje éste sigue siendo el mismo. En segundo lugar, la dilatación del eje, ya que hay que tener en cuenta que en el análisis de correspondencias utilizando el método habitual de normalización simétrica se produce un efecto de dilatación o reescalado para poder representar conjuntamente filas y columnas en dos ejes, aplicando la inversa de la raíz cuadrada del valor propio. Esto implica que dos vectores propios idénticos puedan no tener la misma representación gráfica si los valores propios son diferentes, por ejemplo, al tener diferente inercia total la tabla inicial y la simulada.

Por otro lado, incluso en el caso de que poseamos factores con coordenadas iguales, ello no garantiza que los ejes sean iguales, ya que habrá que analizar también las contribuciones absolutas de cada uno.

Definimos la contribución absoluta de la fila i como el porcentaje de la varianza del factor explicada por la modalidad i:

$$Ca_a(i) = \frac{p_{i.}\hat{\Psi}_{ai}^2}{\sum_{i=1}^{n} p_{i.}\hat{\Psi}_{ai}^2} = \frac{p_{i.}\hat{\Psi}_{ai}^2}{l_a}$$,deduciéndose fácilmente que $$\sum_{i=1}^{n} Ca_a(i) = 1$$

Y la contribución absoluta de la columna j como:

$$Ca_a(j) = \frac{p_{.j}\hat{\Phi}_{aj}^2}{\sum_{j=1}^{q} p_{.j}\hat{\Phi}_{aj}^2} = \frac{p_{.j}\hat{\Phi}_{aj}^2}{l_a}$$

Este análisis de las contribuciones absolutas de los ejes puede darnos una idea más clara de la naturaleza de los mismos. No obstante, hay que considerar que al indicar el porcentaje de la varianza que se explica del eje no se tiene en cuenta si se hace sobre la

parte positiva o sobre la negativa al tener la coordenada elevada al cuadrado, como puede observarse en la fórmula. Por ello, hemos añadido el signo que tiene la modalidad en las coordenadas para el tratamiento posterior de las contribuciones absolutas. Si las contribuciones absolutas de las modalidades no variaran al realizar la simulación ésto sería un indicador de que el eje es estable. También es posible encontrarnos que las modalidades fila no cambiaran y por el contrario sí lo hicieran las modalidades columna, interpretándose entonces como que son las modalidades fila las que definen la estructura del factor y no las modalidades columna.

Otra información útil es la proporcionada por las contribuciones relativas al indicar el porcentaje de la varianza de la modalidad explicada por el factor:

$$Cr_a(i) = \cos^2(q\ i) = \frac{\hat{\Psi}_{ai}^2}{d^2(i, \text{centro gravedad})} \quad \text{,y}$$

$$Cr_a(j) = \cos^2(q\ j) = \frac{\hat{\Phi}_{ai}^2}{d^2(j, \text{centro gravedad})}$$

De la misma forma, contribuciones relativas similares para un factor determinado son un indicador de la estabilidad del mismo.

Surge a continuación el problema de evaluar la similitud de resultados entre el análisis de correspondencias inicial y los obtenidos mediante simulación. Las soluciones son variadas, pudiendo considerar cada una de las simulaciones como una variable más, como si fueran tablas bidimensionales en el tiempo y realizando un análisis de K-tablas, similar al utilizado en STATIS y AFMULT sobre la base de construcción de una tabla media y posicionamiento sobre ésta de las restantes, aplicación del método INDSCAL o técnicas “Procustes”, entre otras.

La solución que presentamos es más simple, tanto en desarrollo como en interpretación, e implica obtener las correlaciones lineales entre el estadístico del análisis de correspondencias obtenido de la tabla inicial y el simulado. Los límites de los intervalos de confianza pueden ser estimados directamente una vez conocidos los percentiles de las distribuciones simuladas.

3.- ANÁLISIS EMPÍRICO

La información sobre la que se ha realizado el análisis procede de la Encuesta de Población Activa del segundo trimestre del año 1999 realizada por el I.N.E. a 196532 personas, de las que 68963 estaban activas. Para este trabajo se ha escogido el nivel de sección (1 dígito y 17 secciones en total), eliminando la última de ellas correspondiente al epígrafe Organismos Extraterritoriales (código 99 CNAE-93) que contenía tan sólo a cuatro personas. Se ha mantenido el nivel de Pesca separado de Agricultura y Ganadería para observar su estabilidad. Las actividades son pues, 1:Agricultura, ganadería, caza y silvicultura; 2: Pesca; 3: Industrias extractivas; 4: Industrias manufactureras; 5: Producción y distribución de energía, electricidad, gas y agua; 6: Construcción; 7: Comercio, reparación de vehículos de motor, motocicletas, ciclomotores y artículos personales de uso doméstico; 8: Hostelería; 9: Transporte, almacenamiento y comunicaciones; 10: Intermediación Financiera; 11: Actividades inmobiliarias y de alquiler, servicios empresariales; 12: Administración Pública, Defensa y Seguridad Social obligatoria; 13: Educación; 14: Actividades sanitarias y veterinarias, servicios sociales; 15: Otras actividades sociales y de servicios prestados a la comunidad, servicios personales; 16: Hogares que emplean personal doméstico.

Esta información se ha cruzado con la Comunidad Autónoma, incluyendo las ciudades de Ceuta y Melilla con el fin ya indicado de confrontar la estabilidad de modalidades que a priori no parecen adecuadas para su inclusión dentro de un análisis de correspondencias. Los códigos asignados por el I.N.E. son los siguientes: 1:ANDALUCIA; 2:ARAGÓN; 3:ASTURIAS; 4:BALEARS; 5:CANARIAS; 6:CANTABRIA; 7:CASTILLA Y LEON; 8:CASTILLA-LA MANCHA; 9:CATALUÑA; 10:COM.VALENCIANA; 11:EXTREMADURA; 12:GALICIA; 13:MADRID; 14:MURCIA; 15:NAVARRA; 16:PAIS VASCO; 17:RIOJA LA; 18:CEUTA Y MELILLA.

Los datos se muestran en el cuadro 1, las actividades en fila y las Comunidades Autónomas en columna:

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18
A1	1121	244	161	35	219	91	673	538	370	292	341	661	36	246	98	76	88	1
A2	32	0	8	5	14	17	0	1	13	6	0	147	0	6	0	12	0	1
A3	36	22	65	4	6	10	72	14	23	12	32	37	11	5	3	6	0	4
A4	1320	706	276	212	252	304	1166	957	1938	1507	194	793	637	342	454	1139	312	18
A5	73	31	11	19	31	12	45	27	49	38	29	18	34	9	7	21	3	5
A6	1198	271	185	235	484	174	728	694	851	609	349	512	350	219	135	351	88	23
A7	1854	525	295	353	834	221	1059	814	1296	1156	452	741	568	442	197	544	144	93
A8	672	165	102	350	503	87	387	277	500	323	163	221	204	82	71	217	44	30
A9	515	159	80	161	248	73	356	239	439	292	85	228	328	90	73	208	21	33
A10	245	98	33	43	55	22	165	120	200	140	41	105	195	44	41	91	30	8
A11	570	193	88	113	249	61	352	183	612	364	105	203	463	91	97	275	54	26
A12	747	234	128	122	355	75	609	362	335	272	209	314	388	106	87	184	50	166
A13	745	188	115	68	254	64	452	312	459	300	162	256	235	119	111	245	41	40
A14	582	235	91	97	172	67	447	290	451	290	166	236	221	112	117	229	55	23
A15	349	105	66	82	170	63	238	160	353	225	94	157	184	44	55	150	22	24
A16	344	62	46	32	93	34	165	117	240	182	58	117	133	46	37	157	11	5

CUADRO 1. Activos por actividad y CCAA. Fuente: E.P.A.

3.1.- Análisis de los valores propios

Realizadas las mil simulaciones de la tabla de contingencia, los resultados del cuadro 2 muestran que no hay una gran diferencia entre los valores propios originales y simulados, a excepción de los tres últimos factores que se encuentran fuera de los intervalos de confianza.

También es posible constatar cómo la dispersión relativa medida a través del coeficiente de variación aumenta a medida que el número del factor es mayor, hecho que ya apuntábamos y que indica la inestabilidad creciente. Realizado el análisis de normalidad sobre los valores propios simulados se obtuvo que tan sólo se rechazaba para los factores 13, 14 y 15.

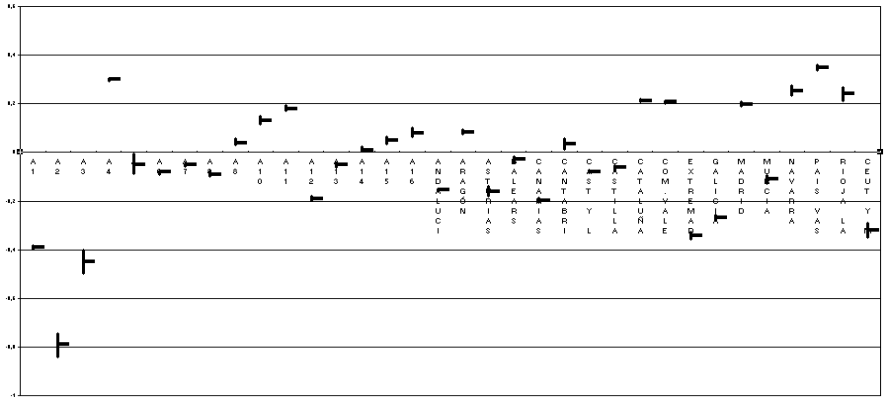
Por todo ello, el estudio de la estabilidad de los valores propios no puede constituir la exclusiva fuente de información para el análisis de correspondencias.

	Valor	Valor Propio 95%		CV		PCT 95%		CV	
	Propio	Lim Inf	Lim Sup	Val.Prop.	PCT	Lim Inf	Lim Sup	PCT	PCT AC
FACTOR 1	0,03847	0,0378	0,0394	0,0127	35,818	34,8643	36,3938	0,0128	35,818
FACTOR 2	0,02826	0,0276	0,0292	0,0178	26,313	25,6221	26,8733	0,0147	62,131
FACTOR 3	0,01313	0,0126	0,0141	0,0332	12,225	11,7247	12,8740	0,0283	74,356
FACTOR 4	0,01164	0,0110	0,0123	0,0338	10,836	10,2572	11,3131	0,0295	85,193
FACTOR 5	0,00575	0,0051	0,0067	0,0862	5,349	4,7217	6,2159	0,0826	90,542
FACTOR 6	0,00447	0,0041	0,0049	0,0591	4,165	3,7684	4,5545	0,0562	94,706
FACTOR 7	0,00218	0,0020	0,0024	0,0614	2,027	1,8541	2,2488	0,0609	96,733
FACTOR 8	0,00141	0,0013	0,0017	0,0764	1,316	1,1915	1,5261	0,0765	98,049
FACTOR 9	0,00096	0,0009	0,0011	0,0842	0,890	0,7938	1,0490	0,0835	98,939
FACTOR 10	0,00051	0,0004	0,0007	0,1265	0,473	0,4052	0,6187	0,1246	99,412
FACTOR 11	0,00033	0,0003	0,0005	0,1561	0,303	0,2521	0,4173	0,1551	99,715
FACTOR 12	0,00020	0,0001	0,0003	0,2004	0,186	0,1336	0,2624	0,2011	99,901
FACTOR 13	0,00006	0,0001	0,0001	0,2626	0,060	0,0482	0,1176	0,2624	99,961
FACTOR 14	0,00004	1,9E-05	7,0E-05	0,3692	0,034	0,0177	0,0651	0,3699	99,995
FACTOR 15	0,00001	1,2E-06	2,5E-05	0,7295	0,005	0,0011	0,0234	0,7304	100,000

CUADRO 2. Valores propios originales y simulados

3.2.- Análisis de las coordenadas

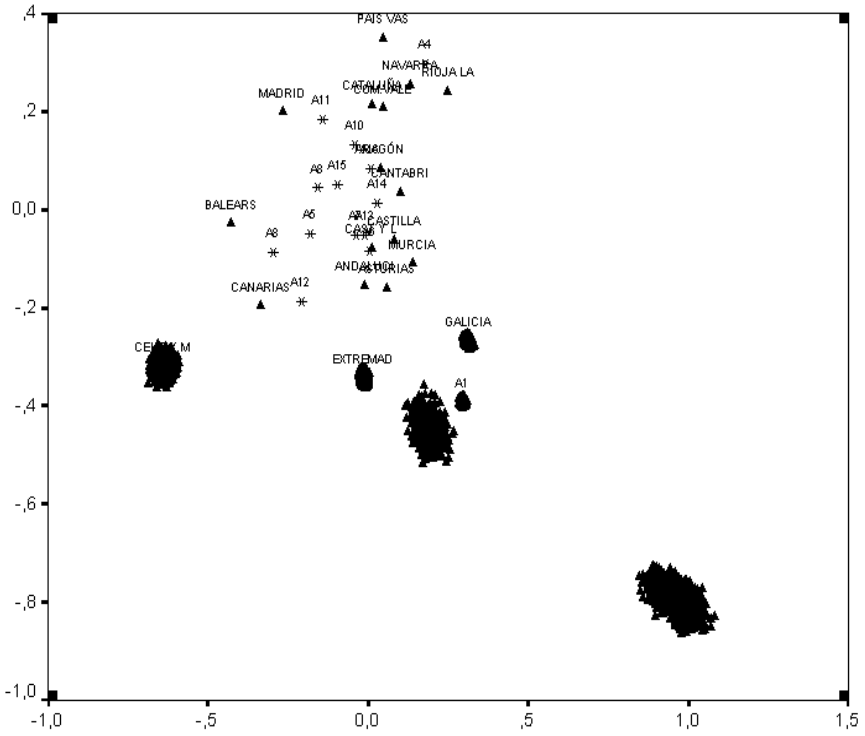
Realizado el análisis de correspondencias simulado, las coordenadas de los resultados se muestran en el cuadro 3. Una vez generada la primera tabla de contingencia simulada calculamos las nuevas coordenadas de las actividades y de las CCAA mediante posicionamiento y se proyectan. Como indicamos anteriormente, grandes elipsoides pueden interpretarse como un signo de pequeña estabilidad. No obstante, esta dispersión medida en términos absolutos puede conducir a conclusiones erróneas. Lógicamente, a medida que nos separamos del baricentro del eje la dispersión aumenta, por lo que sólo debiéramos comparar dispersiones de distancias al baricentro similares. Así, por ejemplo, la dispersión en el primer eje es mayor para la actividad A2 (pesca), situada en la parte inferior izquierda de la gráfica 1, y en menor medida A3 (industrias extractivas).



GRÁFICA 1. Dispersión de las coordenadas simuladas para el primer eje

En la gráfica 2 se muestran las coordenadas originales del análisis de correspondencias posicionando como ilustrativos de izquierda a derecha Ceuta y Melilla, Extremadura, A3, A1, Galicia y A2.

Como puede comprobarse en la gráfica 2, la mayor dispersión corresponde a la actividad A2, contrastada con los resultados del cuadro 3 al tener las mayores desviaciones típicas en los factores 1 y 2 (0,24 y 0,38 respectivamente), seguido de A3. No obstante, al tomar estas variaciones en términos relativos a través del coeficiente de variación de Pearson no son significativamente elevadas en comparación al resto de las modalidades. Representando las dispersiones de forma similar a la gráfica 1, a medida que se aumenta el número de factores la aleatoriedad también se incrementa hasta llegar a la representación del eje número 15 (gráfica 3), pudiendo observar una gran dispersión en todas las modalidades, así como que las coordenadas iniciales se sitúan en valores cercanos a cero y en ocasiones quedan en los extremos de los intervalos simulados, e incluso fuera de ellos.

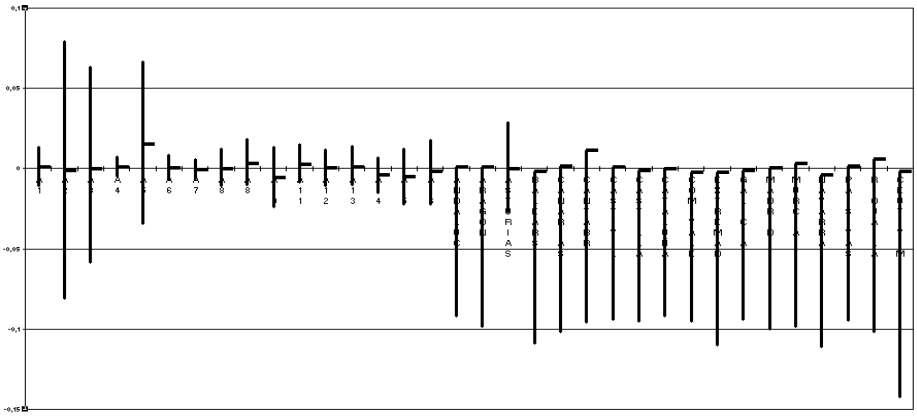


GRÁFICA 2. Planos factoriales 1 y 2

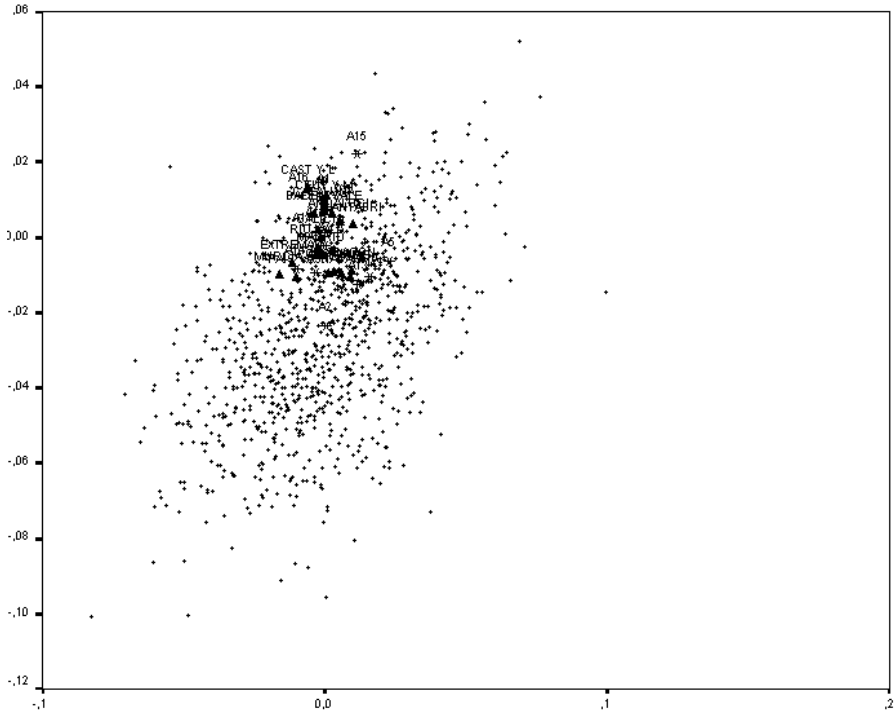
	COORDENADAS				DESV.TIP.				COEF.VAR.			
	F1	F2	F3	F4	F1	F2	F3	F4	F1	F2	F3	F4
A1	-0,390	0,300	-0,070	0,080	0,004	0,004	0,004	0,004	-0,010	0,013	-0,061	0,050
A2	-0,790	0,960	1,260	-0,960	0,024	0,038	0,033	0,034	-0,031	0,040	0,026	-0,036
A3	-0,450	0,190	-0,350	-0,230	0,024	0,022	0,025	0,027	-0,053	0,117	-0,072	-0,120
A4	0,300	0,180	0,000	0,000	0,003	0,002	0,002	0,003	0,009	0,013	-0,829	0,909
A5	-0,050	-0,180	-0,050	0,020	0,021	0,022	0,024	0,027	-0,437	-0,120	-0,440	1,141
A6	-0,080	0,000	0,020	0,090	0,004	0,003	0,004	0,003	-0,045	0,870	0,156	0,039
A7	-0,050	-0,040	0,020	0,060	0,003	0,003	0,003	0,003	-0,051	-0,081	0,145	0,053
A8	-0,090	-0,300	0,210	0,140	0,005	0,007	0,006	0,005	-0,063	-0,023	0,028	0,036
A9	0,040	-0,160	0,070	-0,070	0,006	0,008	0,006	0,006	0,132	-0,048	0,097	-0,091
A10	0,130	-0,040	-0,060	-0,130	0,009	0,010	0,010	0,011	0,068	-0,238	-0,149	-0,089
A11	0,180	-0,140	0,010	-0,110	0,005	0,005	0,006	0,006	0,029	-0,036	0,557	-0,053
A12	-0,190	-0,210	-0,160	-0,220	0,005	0,006	0,005	0,007	-0,024	-0,027	-0,032	-0,031
A13	-0,050	-0,010	-0,080	-0,010	0,005	0,005	0,006	0,007	-0,098	-0,484	-0,078	-0,810
A14	0,010	0,030	-0,080	0,010	0,005	0,006	0,005	0,006	0,338	0,214	-0,065	0,589
A15	0,050	-0,090	0,040	-0,040	0,008	0,009	0,007	0,009	0,153	-0,093	0,175	-0,210
A16	0,080	0,010	0,010	-0,020	0,009	0,009	0,008	0,008	0,111	1,108	0,642	-0,524
CCAA1	-0,154	-0,009	-0,031	0,045	0,003	0,003	0,004	0,003	-0,018	-0,296	-0,122	0,074
CCAA2	0,084	0,039	-0,101	-0,016	0,005	0,005	0,005	0,005	0,065	0,118	-0,050	-0,315
CCAA3	-0,159	0,060	-0,116	-0,062	0,010	0,008	0,013	0,012	-0,062	0,136	-0,111	-0,190
CCAA4	-0,029	-0,428	0,307	0,144	0,007	0,008	0,010	0,009	-0,236	-0,018	0,034	0,064
CCAA5	-0,197	-0,335	0,143	0,068	0,005	0,005	0,007	0,006	-0,023	-0,014	0,047	0,090
CCAA6	0,034	0,102	0,137	-0,026	0,012	0,010	0,020	0,017	0,337	0,100	0,147	-0,653
CCAA7	-0,079	0,011	-0,128	-0,010	0,003	0,003	0,003	0,003	-0,042	0,309	-0,020	-0,305
CCAA8	-0,063	0,080	-0,075	0,091	0,005	0,004	0,004	0,004	-0,072	0,048	-0,047	0,041
CCAA9	0,213	0,013	0,043	0,028	0,003	0,003	0,004	0,003	0,014	0,250	0,082	0,121
CCAA10	0,207	0,049	0,026	0,056	0,003	0,003	0,004	0,004	0,017	0,070	0,141	0,064
CCAA11	-0,343	-0,013	-0,132	0,106	0,007	0,005	0,006	0,006	-0,021	-0,417	-0,049	0,061
CCAA12	-0,268	0,312	0,227	-0,202	0,005	0,006	0,011	0,009	-0,020	0,019	0,047	-0,044
CCAA13	0,198	-0,264	-0,051	-0,226	0,004	0,004	0,004	0,004	0,021	-0,013	-0,073	-0,019
CCAA14	-0,110	0,140	-0,048	0,105	0,009	0,007	0,009	0,008	-0,078	0,052	-0,192	0,077
CCAA15	0,254	0,131	-0,077	-0,004	0,010	0,007	0,006	0,006	0,040	0,054	-0,072	-1,176
CCAA16	0,348	0,048	0,045	-0,036	0,005	0,005	0,006	0,005	0,015	0,097	0,139	-0,130
CCAA17	0,240	0,250	-0,070	0,070	0,014	0,010	0,008	0,009	0,060	0,041	-0,119	0,125
CCAA18	-0,320	-0,639	-0,367	-0,617	0,015	0,016	0,026	0,024	-0,047	-0,025	-0,070	-0,038

CUADRO 3. Coordenadas y dispersión de modalidades simuladas

La gráfica 4 muestra la representación de las simulaciones de la modalidad A2 proyectada sobre los ejes 13 (abscisa) y 14 (ordenada). Como puede observarse, los puntos proyectados sobre los últimos ejes pueden alcanzar tal grado de dispersión que engloben a todas las modalidades, tanto filas como columnas, lo que nos lleva a pensar que la explicación de ejes con valor propio pequeño es altamente inestable debido al gran componente de aleatoriedad que contienen.



GRÁFICA 3. Dispersión de las coordenadas simuladas para el último eje



GRÁFICA 4. Posicionamiento sobre los ejes 13 y 14 de la modalidad simulada A2.

A continuación se calculó el coeficiente de correlación lineal de Pearson entre las coordenadas iniciales de las modalidades de actividad del primer factor y las coordenadas de las modalidades de actividad del primer factor simulado. Se repitió el proceso las mil veces mencionadas, obteniendo una media de correlación de 0,998 y se construyó el intervalo de confianza al 95% mediante la utilización de percentiles, que en este caso es de [0,994; 1,000]. También se realizó la prueba de normalidad de Kolmogorov-Smirnov de la que se muestra el valor p en el cuadro 4. Similar procedimiento se sigue para las CCAA y los catorce factores restantes.

Se puede comprobar cómo las correlaciones entre coordenadas iniciales y simuladas son muy altas, así como que se rechaza la hipótesis de normalidad de la distribución generada con las correlaciones.

	CORRELACIONES DE COORDENADAS							
	ACTIVIDAD				CCAA			
FACTOR	Lim Inf	Media	Lim Sup	NORM	Lim Inf	Media	Lim Sup	NORM
1	0,994	0,998	1,000	0,000	0,993	0,998	1,000	0,000
2	0,997	0,999	1,000	0,000	0,993	0,998	1,000	0,000
3	0,822	0,964	0,998	0,000	0,948	0,990	1,000	0,000
4	0,921	0,983	0,999	0,000	0,849	0,975	0,999	0,000
5	0,690	0,949	0,997	0,000	0,965	0,991	0,999	0,000
6	0,912	0,983	0,999	0,000	0,524	0,921	0,996	0,000
7	0,954	0,983	0,996	0,000	0,894	0,961	0,992	0,000
8	0,906	0,969	0,993	0,000	0,897	0,965	0,992	0,000
9	0,875	0,954	0,989	0,000	0,729	0,905	0,980	0,000
10	0,519	0,884	0,982	0,000	0,549	0,928	0,993	0,000
11	0,376	0,879	0,987	0,000	0,245	0,878	0,991	0,000
12	0,292	0,865	0,982	0,000	0,391	0,857	0,983	0,000
13	0,061	0,649	0,933	0,000	0,073	0,712	0,974	0,000
14	0,057	0,607	0,938	0,000	0,070	0,668	0,974	0,000
15	0,038	0,584	0,953	0,000	0,373	0,888	0,992	0,000

CUADRO 4. Correlación entre coordenadas iniciales y simuladas

3.3.- Análisis de los factores

Como indicamos anteriormente, es posible que coordenadas similares definan factores diferentes si la contribución que las modalidades aportan a la explicación del eje es distinta. Por ello se partió del análisis conjunto de las tablas de contribuciones absolutas y relativas para las modalidades de actividades y CCAA (cuadro 5).

	COORDENADAS				CONTRIB. ABSOLUTAS				CONTRIB. RELATIVAS			
	F1	F2	F3	F4	F1	F2	F3	F4	F1	F2	F3	F4
A1	-0,390	0,300	-0,070	0,080	30,82	23,82	3,15	4,23	0,60	0,34	0,02	0,03
A2	-0,790	0,960	1,260	-0,960	6,21	12,39	45,95	30,23	0,16	0,23	0,39	0,23
A3	-0,450	0,190	-0,350	-0,230	2,74	0,67	4,80	2,33	0,15	0,03	0,09	0,04
A4	0,300	0,180	0,000	0,000	41,75	20,91	0,01	0,01	0,71	0,26	0,00	0,00
A5	-0,050	-0,180	-0,050	0,020	0,04	0,77	0,15	0,03	0,03	0,43	0,04	0,01
A6	-0,080	0,000	0,020	0,090	1,96	0,01	0,49	7,06	0,33	0,00	0,03	0,36
A7	-0,050	-0,040	0,020	0,060	1,21	0,81	0,59	4,67	0,17	0,09	0,03	0,20
A8	-0,090	-0,300	0,210	0,140	1,27	19,73	21,94	11,45	0,04	0,50	0,26	0,12
A9	0,040	-0,160	0,070	-0,070	0,27	4,60	1,70	2,29	0,05	0,56	0,10	0,12
A10	0,130	-0,040	-0,060	-0,130	1,10	0,14	0,78	3,44	0,21	0,02	0,05	0,20
A11	0,180	-0,140	0,010	-0,110	5,19	4,33	0,05	6,12	0,37	0,23	0,00	0,13
A12	-0,190	-0,210	-0,160	-0,220	6,33	10,60	14,24	27,57	0,21	0,25	0,16	0,27
A13	-0,050	-0,010	-0,080	-0,010	0,42	0,02	2,94	0,04	0,14	0,01	0,33	0,00
A14	0,010	0,030	-0,080	0,010	0,03	0,17	2,78	0,05	0,01	0,05	0,39	0,01
A15	0,050	-0,090	0,040	-0,040	0,24	1,15	0,48	0,55	0,11	0,38	0,07	0,07
A16	0,080	0,010	0,010	-0,020	0,49	0,01	0,03	0,06	0,14	0,00	0,00	0,01
CCAA1	-0,154	-0,009	-0,031	0,045	9,32	0,05	1,08	2,64	0,68	0,00	0,03	0,06
CCAA2	0,084	0,039	-0,101	-0,016	0,86	0,26	3,61	0,11	0,23	0,05	0,33	0,01
CCAA3	-0,159	0,060	-0,116	-0,062	1,68	0,33	2,60	0,84	0,12	0,02	0,07	0,02
CCAA4	-0,029	-0,428	0,307	0,144	0,06	18,12	20,09	5,02	0,00	0,55	0,28	0,06
CCAA5	-0,197	-0,335	0,143	0,068	5,75	22,68	8,95	2,27	0,22	0,62	0,11	0,03
CCAA6	0,034	0,102	0,137	-0,026	0,06	0,74	2,83	0,12	0,02	0,21	0,37	0,01
CCAA7	-0,079	0,011	-0,128	-0,010	1,61	0,04	12,44	0,08	0,23	0,00	0,61	0,00
CCAA8	-0,063	0,080	-0,075	0,091	0,75	1,68	3,15	5,26	0,11	0,18	0,15	0,23
CCAA9	0,213	0,013	0,043	0,028	13,94	0,07	1,68	0,77	0,90	0,00	0,04	0,02
CCAA10	0,207	0,049	0,026	0,056	9,66	0,73	0,45	2,32	0,74	0,04	0,01	0,05
CCAA11	-0,343	-0,013	-0,132	0,106	10,98	0,02	4,74	3,49	0,75	0,00	0,11	0,07
CCAA12	-0,268	0,312	0,227	-0,202	12,83	23,72	26,97	24,21	0,27	0,37	0,20	0,16
CCAA13	0,198	-0,264	-0,051	-0,226	5,92	14,28	1,14	25,35	0,21	0,37	0,01	0,27
CCAA14	-0,110	0,140	-0,048	0,105	0,91	2,02	0,50	2,74	0,17	0,28	0,03	0,16
CCAA15	0,254	0,131	-0,077	-0,004	3,84	1,39	1,02	0,00	0,62	0,17	0,06	0,00
CCAA16	0,348	0,048	0,045	-0,036	17,86	0,46	0,89	0,64	0,89	0,02	0,02	0,01
CCAA17	0,240	0,250	-0,070	0,070	2,09	3,08	0,52	0,58	0,32	0,35	0,03	0,03
CCAA18	-0,320	-0,639	-0,367	-0,617	1,93	10,47	7,44	23,70	0,08	0,32	0,11	0,30

CUADRO 5. Coordenadas y contribuciones absolutas y relativas de la tabla original

A continuación se realizó el mismo análisis para cada una de las mil tablas simuladas, obteniendo las correlaciones entre las contribuciones originales y simuladas, construyendo una nueva distribución de correlaciones. También se analizó la normalidad de esa distribución. Los resultados se muestran en el cuadro 6.

En él se observa una alta estabilidad de las contribuciones absolutas de las actividades que decrece a medida que los factores aumentan, es decir, a medida que se incorporan elementos de aleatoriedad.

	CORRELACIONES CONTRIBUCIONES ABSOLUTAS							
	ACTIVIDAD				CCAA			
FACTOR	Lim Inf	Media	Lim Sup	NORM	Lim Inf	Media	Lim Sup	NORM
1	0,991	0,998	1,000	0,000	0,200	0,303	0,409	0,930
2	0,993	0,998	1,000	0,000	0,333	0,383	0,434	0,486
3	0,764	0,948	0,998	0,000	0,251	0,765	0,902	0,000
4	0,808	0,949	0,998	0,000	0,368	0,706	0,828	0,000
5	0,819	0,976	0,999	0,000	0,875	0,978	0,996	0,000
6	0,693	0,961	0,998	0,000	0,007	0,121	0,274	0,024
7	0,926	0,975	0,994	0,000	0,530	0,720	0,819	0,000
8	0,861	0,958	0,992	0,000	0,012	0,275	0,553	0,008
9	0,859	0,959	0,994	0,000	0,567	0,790	0,897	0,000
10	0,470	0,897	0,988	0,000	0,014	0,278	0,620	0,000
11	0,261	0,825	0,987	0,000	0,028	0,391	0,720	0,002
12	0,348	0,802	0,962	0,000	0,015	0,283	0,638	0,000
13	0,056	0,665	0,950	0,000	0,014	0,303	0,619	0,014
14	0,084	0,581	0,919	0,000	0,021	0,369	0,700	0,001
15	0,021	0,515	0,970	0,000	0,009	0,259	0,565	0,001

CUADRO 6. Correlación entre contribuciones absolutas iniciales y simuladas

Hay que señalar en este estudio que son las actividades las que constituyen los factores y no las CCAA como demuestra el hecho de que la correlación entre las contribuciones absolutas de las CCAA simuladas e iniciales es muy baja. Una excepción es el eje número cinco, con una correlación media de 0,978 e intervalos [0,875; 0,978]. Ello es debido a que este eje está caracterizado por la presencia de Asturias altamente relacionada con el A3 (Asturias: 63% de contribución absoluta y 69,5% de contribución relativa; A3 Industrias extractivas: 79,3% de contribución absoluta y 64,7% de contribución relativa).

A excepción de los dos primeros factores de las CCAA en los que la distribución de correlaciones puede considerarse normal, la hipótesis nula de normalidad es rechazada en el resto. Estos resultados fueron contrastados al realizar los cálculos de los coeficientes de asimetría y curtosis, observando cómo las distribuciones de correlaciones resultaron ser, en general, fuertemente asimétricas y leptocúrticas. Estas consideraciones nos llevan a reflexionar sobre la idoneidad de la utilización, al menos en este trabajo, de técnicas asintóticas como instrumento del estudio de la estabilidad de los resultados.

Lo atípico de la inclusión en esta aplicación de la pesca (A2) de forma separada a la agricultura, ganadería, caza y silvicultura, así como la consideración de Ceuta y Melilla,

nos ha servido para la contrastación de la estabilidad de los resultados, a pesar de perder capacidad comparativa con otros estudios que siguen los criterios de agrupación tradicionales.

Ello ha planteado el inconveniente de que a pesar de que la mayor parte de la varianza se encuentra explicada por los primeros ejes (1:35,8%; 2:26,3%; 3:12,2% y 4:10,8%, sumando los cuatro primeros el 85,2% del total), nos encontramos que parte de la información se encuentra bastante repartida en un buen número de ejes, como lo demuestra el cuadro de las contribuciones relativas (cuadro 5). En él se puede observar que hay un considerable número de modalidades que no son suficientemente explicadas por los cuatro primeros factores, quedando pendiente de explicar por orden decreciente de importancia las modalidades A16(84,6%), CCAA3(77,3%), A3(69,4%), A14(53,8%), A10(52,4%), A13(51,8%), A7(51,3%), A5(48,7%), CCAA6(38,7%), CCAA2(38,2%), A15(36,5%), CCAA14(36%) y CCAA8(33,6%), entre otras.

Esta dispersión de la información se plasma también en el estudio de las correlaciones entre las contribuciones relativas originales y simuladas recogido en el cuadro 7. Los ejes tienen una buena estabilidad analizando las medias de las correlaciones y los límites de los intervalos, especialmente los tres primeros. El cuarto disminuye el grado de correlación, especialmente en cuanto al límite inferior, ya que es altamente asimétrico. El quinto eje ya mencionado por estar caracterizado por las Industrias extractivas en Asturias tiene un incremento de la correlación media, pero fundamentalmente en el límite inferior. Por el contrario, el siguiente factor, el número seis, muestra una reducción drástica, que se recupera en el siete para volver a caer a partir del eje décimo.

La falta de normalidad también en las contribuciones relativas confirma los problemas de utilización de técnicas asintóticas en este caso.

	CORRELACIONES CONTRIBUCIONES RELATIVAS							
	ACTIVIDAD				CCAA			
FACTOR	Lim Inf	Media	Lim Sup	NORM	Lim Inf	Media	Lim Sup	NORM
1	0,993	0,997	0,999	0,000	0,892	0,921	0,940	0,000
2	0,981	0,993	0,998	0,000	0,834	0,877	0,913	0,000
3	0,809	0,953	0,996	0,000	0,672	0,858	0,945	0,000
4	0,626	0,907	0,994	0,000	0,329	0,762	0,901	0,000
5	0,859	0,977	0,998	0,000	0,754	0,968	0,999	0,000
6	0,381	0,879	0,987	0,000	0,311	0,830	0,928	0,000
7	0,723	0,913	0,982	0,000	0,887	0,969	0,997	0,000
8	0,801	0,942	0,993	0,000	0,780	0,949	0,994	0,000
9	0,829	0,948	0,996	0,000	0,772	0,932	0,993	0,000
10	0,182	0,749	0,960	0,000	0,354	0,857	0,991	0,000
11	0,217	0,828	0,990	0,000	0,060	0,823	0,996	0,000
12	0,079	0,727	0,973	0,000	0,244	0,726	0,964	0,000
13	0,020	0,543	0,950	0,000	0,057	0,654	0,977	0,000
14	0,154	0,553	0,908	0,003	0,387	0,668	0,929	0,003
15	0,017	0,635	0,996	0,000	0,152	0,797	0,993	0,000

CUADRO 7. Correlación entre contribuciones relativas iniciales y simuladas

4.- CONCLUSIONES

El esfuerzo en la búsqueda del modelo verdadero, así como la utilización de técnicas de análisis de datos cada vez más potentes y sofisticadas puede que nos hagan disminuir la atención sobre los datos.

Este estudio nos ha permitido comprobar que es necesario plantearnos la pregunta sobre la estabilidad de nuestros resultados, si pequeñas modificaciones pueden afectar sustancialmente a un modelo o a los resultados de una determinada técnica.

Las técnicas de remuestreo o bootstrap están demostrando que pueden ser instrumentos válidos para el análisis de la estabilidad de los resultados, especialmente aplicando el “bootstrap no paramétrico” que no implica independencia de las variables ni distribuciones empíricas que deban suponerse o contrastarse.

Una vez construida mediante simulación la distribución de un estimador es posible trabajar con él utilizando sencillas técnicas descriptivas y construir los intervalos de confianza deseados mediante percentiles.

Uno de los inconvenientes de las técnicas de bootstrap frente a otras como la de Jackknife es la gran cantidad de cálculos y tiempo de procesamiento requerido. Mientras hace veinte años se podía necesitar más de quince minutos para efectuar un análisis de correspondencias, hoy en día ambos conceptos son relativos, ya que supone unos

segundos. El mayor consumo de tiempo se produce al generar la tabla mediante remuestreo al contener 68959 individuos. El proceso completo de generación de la tabla, realización del análisis de correspondencias sobre la tabla simulada, cálculo de las correlaciones e impresión es aproximadamente de seis segundos por simulación. Hay que reconocer que número de simulaciones elegido (1000) es excesivamente elevado en relación con otros análisis de simulación aplicados a diferentes técnicas, pero evidentemente proporciona resultados más fiables. También hay que destacar que las matrices de correlaciones han mostrado ser en diferentes estudios más estables que las medias.

Quedaría, por último, el problema de evaluar la similaridad de resultados entre el análisis de correspondencias inicial y los obtenidos mediante simulación. Como ya indicamos, hay diferentes aproximaciones.

La solución que presentamos en este trabajo es la más simple, pero a su vez es intuitiva y fácilmente interpretable. Se obtienen las correlaciones lineales entre un estadístico calculado del análisis de correspondencias de la tabla inicial y el estadístico simulado. Los límites de los intervalos de confianza se estiman directamente a partir de la determinación de los percentiles de las distribuciones simuladas.

Hemos comprobado que en la mayor parte de los casos las distribuciones de estos estimadores no siguen una ley normal, lo que implica un riesgo si se desean aplicar técnicas asintóticas para la construcción de los intervalos de confianza, como puede ser el método delta.

Otras soluciones que serán objeto de posteriores trabajos es la consideración de las tablas simuladas como una variable más en la realización de un análisis de K-tablas o técnicas “Procustes”, por ejemplo.

5.- BIBLIOGRAFÍA

ÁLVAREZ, R.; GÓMEZ, M.; HUERGA, C.; MURES, M.J. Estudio de la población ocupada en Castilla y León mediante el Análisis de Correspondencias utilizando tablas de frecuencia en diferentes momentos de tiempo. Actas del IV Congreso de Economía Regional de Castilla y León, Burgos, Vol.2, pp.1146-1155.

GIFI, A. Nonlinear multivariate analysis, John Wiley & Sons Ltd., Chichester, 1990, pp.391-424.

GREENACRE, M.J. Theory and applications of correspondence analysis, Academic Press, London, 1984.

LEBART, L.; MORINEAU, A.; PIRON, M. Statistique exploratoire multidimensionnelle, Dunod, Paris, 1995, pp.357-404.

MARKUS, M.TH. Bootstrap confidence regions for homogeneity analysis; the influence of rotation coverage percentages. Proceedings in Computational Statistics. COMPSTAT. 1994. pp.337- 342.

MICHAILIDIS, G.; DE LEEW, J. Multilevel homogeneity analysis with differential weighting. Computational Statistics & Data Analysis. Vol 32, January 2000, pp.411-442.

MURES QUINTANA, M.J.; ÁLVAREZ ESTEBAN, R. Tipología de la población ocupada en España (1987-1992). Análisis de correspondencias múltiples mediantes tablas de Burt. VII Reunión de ASEPELT, Palma, Junio 1994. Estudios de Economía Aplicada, Vol.I, pp.407-414.

REICZIGEL, J. Bootstrap tests in correspondence analysis. Applied Stochastic Models and Data Analysis, Vol.12, 1996, pp.107-117.