Metodología de Encuestas
Volumen 9, 2007, 93-103
ISSN: 1575-7803

LA SIMULACIÓN EN EL ANÁLISIS DE DATOS. UNA HERRAMIENTA DE VALIDACIÓN DE RESULTADOS

Olga Valencia García
Departamento de Economía Aplicada

RESUMEN. El objetivo de este trabajo es verificar la validez de los resultados del Análisis Factorial de Correspondencias (AFC). Para comprobar que la estructura detectada por el AFC es real y que sus resultados son estables respecto a las fluctuaciones del muestreo, se propone un método basado en la simulación de muestras: el Bootstrap no paramétrico. La aplicación de un Bootstrap Parcial permite evaluar la estabilidad de las configuraciones de filas y columnas y el uso de un Bootstrap Total sirve para verificar la estabilidad de los ejes factoriales. En concreto, el AFC se ha aplicado a una tabla con datos textuales que cruza las unidades léxicas más repetidas en la respuesta a una pregunta abierta (en filas) por varios grupos de encuestados (en columnas). Se han obtenido regiones de confianza para los elementos representados en los planos factoriales, así como medidas de estabilidad de los factores.

PALABRAS CLAVE. Análisis de datos, Simulación, Bootstrap, Estabilidad, Tabla léxica.

ABSTRACT. The aim of this paper is to validate the results of Correspondence Analysis (CA). To check that the structure revealed by CA is real and their results are stable with regard to sampling fluctuations, a method based on simulated samples, the non-parametric Bootstrap, is suggested. The application of a Partial Bootstrap allows evaluating the stability of row and column configurations and the use of a Total Bootstrap measures the stability of the factorial axes. CA has been applied to a particular data set, a cross-tabulation of textual units in the response of an open question (rows) by several groups of respondents (columns). Confidence regions for the elements represented in factorial planes and a measure of the stability of the factors have been obtained.

KEYWORDS. Data analysis, Simulation, Bootstrap, Stability, Textual data.

Introducción

El Análisis Factorial de Correspondencias (AFC) es un método exploratorio: detecta la estructura básica de la matriz de datos sin necesidad de verificar condiciones matemáticas sobre su procedencia y no trata de generalizar ni confirmar patrones.

Sin embargo, es inevitable plantearse si los resultados que proporciona el AFC reflejan una estructura de asociaciones real o producto del azar y por tanto, hasta qué punto son válidos. El carácter exploratorio de una técnica no significa que sus resultados no deban someterse a algún tipo de "control de calidad". Consideramos que los resultados son válidos en la medida en que sean estables.

¿Qué entendemos por estables? El concepto general de estabilidad que utilizamos es el propuesto por Gifi (1990): la estabilidad tiene lugar cuando pequeños cambios en el input (matriz de datos) generan modificaciones escasas y poco importantes en el output (factores y configuraciones). Ahora bien, siguiendo la distinción efectuada por Greenacre (1984), la estabilidad puede estudiarse a dos niveles: a nivel de la matriz de datos (estabilidad interna) y a nivel de la población de la que proceden los datos (estabilidad externa). En la primera se verifica "qué ocurre si cambia algún elemento de la matriz". Su estudio es una fase implícita en el análisis de datos exploratorio. En la estabilidad externa, se trata de comprobar "qué ocurre al efectuar el análisis sobre otras muestras extraídas de la misma población" y, en consecuencia, de observar la variabilidad de los resultados respecto a las fluctuaciones del muestreo.

El objeto de este trabajo es el estudio de la estabilidad externa. Para ello es necesario algún mecanismo de generación de muestras ya sea real, abstracto o simulado. La extracción de muestras reales no suele ser viable en la investigación social. Por otro lado, la replicación abstracta supone sustituir la extracción de muestras por modelos que expliquen el comportamiento en el muestreo de los estadísticos de interés. Ésta es la base de la inferencia clásica. Pero a veces estos modelos no existen o no pueden aplicarse. La alternativa es acudir a la simulación de muestras. Las muestras simuladas se obtienen por mecanismos aleatorios que perturban los datos observados en un sentido probabilístico sin asumir modelos de probabilidad previos.

Como método de simulación, se sugiere el Bootstrap no paramétrico (Efron y Tibshirani, 1993). Las muestras simuladas o muestras Bootstrap se obtienen a partir de un muestreo con reemplazamiento de la muestra original. En el caso del AFC, Greenacre (1984) establece el modo de generación de las muestras Bootstrap: extracción de muestras del mismo tamaño que la muestra original a partir de una distribución multinomial definida por las celdas de la matriz original, con probabilidades estimadas por las frecuencias relativas.

Estabilidad y simulación Bootstrap

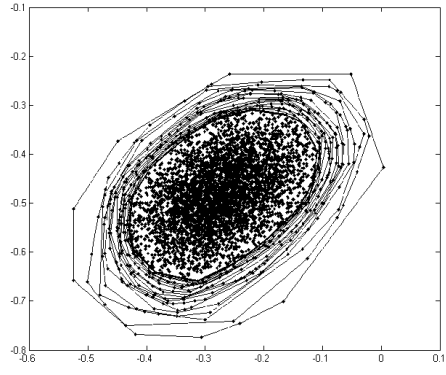
En la literatura, las aplicaciones del Bootstrap a los métodos factoriales se pueden agrupar en dos tipos: las que realizan un Bootstrap Parcial (BP) y las que

realizan un Bootstrap Total (BT). El BP es un procedimiento sencillo: proyecta los elementos de las tablas simuladas sobre el subespacio de referencia generado a partir del análisis factorial original como elementos suplementarios, obteniendo así coordenadas simuladas, por lo que es adecuado para estudiar la estabilidad de las configuraciones. El BT es un método más complejo puesto que efectúa un análisis factorial de cada muestra Bootstrap, generando todos los estadísticos del análisis, lo que permite analizar la estabilidad de la estructura factorial completa.

En AFC, el objetivo del BP es estudiar la estabilidad de la posición de una categoría *a* (fila o columna) en un subespacio factorial, generalmente un plano. El AFC de la tabla original proporciona su coordenada real. Una vez generadas las *B* tablas simuladas, se proyecta la categoría *a* de cada una de ellas como elemento suplementario sobre el plano factorial original. El resultado es una nube de *B* coordenadas Bootstrap.

A partir de esa nube se puede construir una región de confianza bidimensional de nivel (1-δ)% eliminando el δ% de los puntos más extremos mediante un algoritmo de "pelado". El resultado es un *convex hull* que contiene exactamente el porcentaje de puntos deseado (polígono interior o de menor área de la Figura 1), el cual constituye una región de confianza Bootstrap no paramétrica al (1-δ)% para la posición en el plano factorial de la categoría (fila o columna) analizada. Regiones de confianza reducidas indican posiciones estables en los planos factoriales.

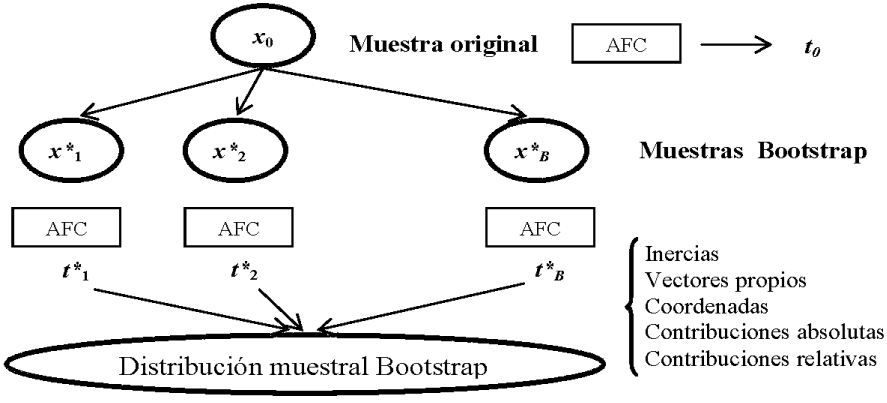
Figura 1.
Bootstrap Parcial. "Pelado" del convex hull y determinación de una Región Bootstrap



En cuanto al BT, además del AFC de la tabla original, se realiza un AFC de cada una de las tablas Bootstrap, generándose todos los estadísticos propios del análisis: valores propios, vectores propios, coordenadas y contribuciones. De esta forma tendremos estadísticos originales y estadísticos Bootstrap. La Figura 2 designa por t_0 al conjunto de estadísticos obtenidos del AFC de la muestra original

x_0. Cada muestra Bootstrap x^*_b da lugar a los estadísticos Bootstrap correspondientes, denotados por t^*_b, para b: 1,..., B.

Figura 2.
Bootstrap Total: Generación de estadísticos Bootstrap



Analizar la estabilidad de la estructura factorial implica un estudio de la variabilidad de todos los estadísticos que conforman los factores. Vamos a centrarnos en uno de estos estadísticos: los vectores propios. Su estabilidad se verifica comparando cada vector original con cada uno de los vectores simulados. Esa comparación se ha efectuado computando ángulos entre cada vector propio original $u\alpha$ y cada vector propio simulado mediante Bootstrap, $u^* \alpha b$, tanto para vectores del mismo rango (por ejemplo, u_1 y u^*_{1b}) como de distinto rango (por ejemplo, u_1 y u^*_{2b}). Si los vectores propios son estables, los menores ángulos se obtienen entre vectores originales y simulados del mismo rango en un elevado porcentaje de las B simulaciones y estos ángulos estarán próximos a 0º. Cuanto más se alejen de 0º, mayor inestabilidad existe.

Ahora bien, la aplicación directa del Bootstrap a métodos factoriales provoca algunos problemas que han de resolverse. En el BP, el aumento de inercia de las muestras Bootstrap puede dilatar las coordenadas simuladas. Por ello, hemos aplicado una corrección de dichas coordenadas tomando como referencia la inercia original, tal y como se recoge en Álvarez, Bécue y Valencia (2006).

En cuanto al BT, como subrayan Milan y Whittaker (1995), la comparación de estadísticos de diferentes subespacios da lugar a una variabilidad excesiva, en forma de reflejos, intercambios y rotaciones de factores. Por ello, las distribuciones muestrales Bootstrap de los valores y vectores propios directamente obtenidos del BT no son buenas aproximaciones para evaluar la estabilidad de los vectores propios. En lo que se refiere al caso particular del AFC, según la bibliografía consultada, las soluciones ofrecidas en la literatura no son satisfactorias. Algunos autores como Greenacre (1984), Ringrose (1992) y Lebart (2004) sugieren el uso del

BP, más enfocado a la estabilidad de las configuraciones y cuando se aplica un BT, como en Reiczigel (1996) o en Beck-Nielsen et al. (2004), no se implementan ajustes para abordar los problemas mencionados. Esto conduce a una evaluación inadecuada del grado de estabilidad.

Por ello, se ha aplicado una corrección basada en una rotación Procustes de las coordenadas simuladas a las originales, con reconstrucción posterior de los ejes simulados. Esta corrección se ha desarrollado extensamente en Valencia (2006). El resultado de su aplicación es una evaluación más apropiada del grado de estabilidad de los ejes, basada en su variabilidad real y no en su variabilidad aparente.

3. Aplicación a una tabla con datos textuales

La metodología propuesta se ha aplicado al AFC de una tabla léxica agregada. Los datos proceden de una encuesta a jueces españoles para conocer sus opiniones sobre la carrera judicial: "Jueces en su primer destino, 2002", realizada en el marco del proyecto "Observatorio de cultura judicial" (SEC 2001-2581-C02). El método de recogida de la información es la entrevista personal mediante un cuestionario estructurado que contiene preguntas abiertas y cerradas. La información detallada sobre las características técnicas de la encuesta puede encontrarse en Álvarez et al. (2003).

La tabla léxica analizada surge de la información cruzada de una pregunta cerrada y una pregunta abierta del cuestionario. La pregunta cerrada es "¿Qué valoración le merece la formación obtenida en la Facultad de Derecho?", con 5 categorías de respuesta, desde "Muy Negativa" hasta "Muy Buena". La pregunta abierta es la cuestión complementaria "¿Por qué?". Este tipo de preguntas en que se pide una valoración y una justificación de esa valoración es muy habitual en los cuestionarios.

De este modo se obtiene una tabla léxica agregada, con más de 2000 ocurrencias, que posee 114 filas, relativas a las unidades léxicas cuya frecuencia es superior a 5, y 5 columnas con las categorías de encuestados consideradas. Es una matriz con muchos ceros, con frecuencias conjuntas muy bajas y con frecuencias marginales escasas, especialmente en el caso de las unidades léxicas, que incluyen tanto formas gráficas (palabras) como lemas. Estos últimos proceden de una agrupación de las palabras que corresponden con una misma entrada de diccionario (lematización) operación que, en un texto en castellano, requiere principalmente convertir las formas verbales al infinitivo, los sustantivos al singular y los adjetivos al masculino singular. En los planos factoriales mostrados a continuación, los lemas terminan con el símbolo %. Cuando éste no aparece, las unidades léxicas son las formas gráficas originales del texto.

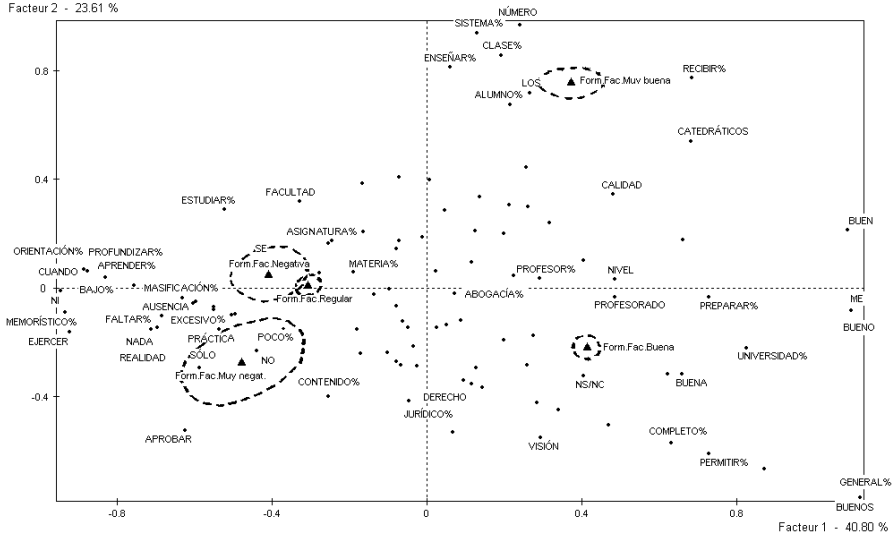
Presentamos a continuación algunos resultados proporcionados por la aplicación del Bootstrap Parcial y del Bootstrap Total, ambos corregidos por los procedimientos mencionados. En todos los casos se ha generado un número B=5000 muestras simuladas.

El BP permite evaluar la estabilidad de las configuraciones. Los planos factoriales originales han sido obtenidos a partir del programa SPAD 5.5. Las coordenadas de las categorías simuladas, categorías proyectadas como elementos

suplementarios en el plano principal del AFC original, se han computado mediante un software específico y han sido superpuestas a dichos planos.

La Figura 3 muestra las regiones del BP de las columnas en el plano principal (con coordenadas corregidas). Su reducido tamaño indica que las categorías de respuesta son muy estables en el primer plano factorial, especialmente "Buena" y "Regular". Las categorías minoritarias, sobre todo, "Muy negativa", son algo menos estables. El solapamiento entre las regiones de "Negativa" y "Regular" señala que los dos primeros factores no diferencian los perfiles léxicos de estos dos grupos de respuesta. Por el contrario, las valoraciones favorables tienen perfiles léxicos bien diferenciados.

Figura 3.
Bootstrap Parcial. Regiones Bootstrap al 90%. Categorías de respuesta

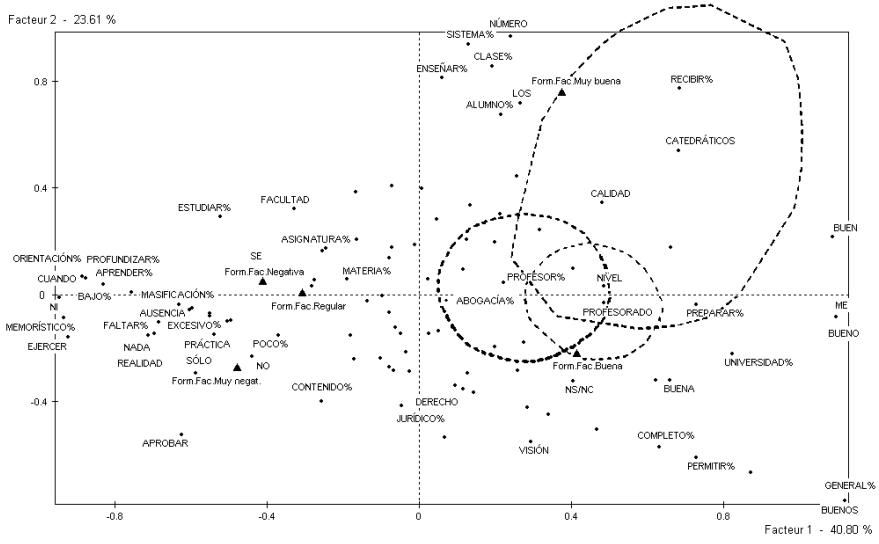


Mostramos también regiones de confianza para algunas unidades léxicas no gramaticales (formas/lemas semánticamente llenos). Hemos representado conjuntamente regiones Bootstrap de unidades léxicas que poseen una conexión semántica, con el fin de que la interpretación comparada de su estabilidad tenga sentido. Como ejemplo, la Figura 4 presenta las regiones Bootstrap de las unidades "profesorado", "profesor%" y "catedráticos".

Como se puede apreciar, estas tres unidades léxicas no son tan estables como las columnas de la tabla (categorías de encuestados). "Profesorado" y "profesor%" tienen un grado de estabilidad comparable y "catedráticos", con menor frecuencia, es menos estable. El solapamiento parcial revela perfiles de respuesta similares entre "profesorado" y "profesor%" y un perfil de respuesta menos delimitado, pero en parte compartido, del término "catedráticos". La mención del "profesorado" está ligada a la opinión "Buena" mientras que "catedráticos" está más

vinculada a la valoración "Muy buena". En todo caso, la localización de las tres regiones dentro del plano factorial refleja que se trata de unidades léxicas asociadas a las valoraciones favorables.

Figura 4.
Bootstrap Parcial. Regiones Bootstrap al 90%. Unidades léxicas



La estabilidad de los ejes factoriales se analiza a partir de los resultados del BT. A modo de ejemplo, se ha analizado la variabilidad de los vectores columna, denotados por $u\alpha$. Como se ha indicado anteriormente, si los ejes son estables, los vectores originales deben presentar ángulos muy pequeños con vectores simulados del mismo rango.

La Tabla 1 muestra el porcentaje de las 5000 muestras simuladas en las que se produce esta circunstancia para cada uno de los vectores columna α: 1,..., 4. En el caso del primer vector, todas las simulaciones cumplen esta característica. En cambio, en los tres restantes, buena parte de las simulaciones no la cumplen. Por ejemplo, el tercer vector original presenta el menor ángulo con el tercer vector simulado en un 51,4% de las simulaciones (ángulo medio de 29º), mientras que en buena parte de las simulaciones (un 26,9%) el tercer vector original presenta ángulos más pequeños con el cuarto vector simulado y en un 21,7% de las simulaciones, con el segundo vector simulado. Esto refleja cierta inestabilidad. Algo similar podría afirmarse para los vectores originales segundo y cuarto. Ahora bien, después de la corrección se demuestra que esta inestabilidad tiene un carácter aparente en todos los vectores.

Tabla 1.
Porcentaje de las 5000 simulaciones en las que los menores ángulos se computan entre vectores originales y simulados del mismo rango

	% Simulaciones con ángulos mínimos			
	u_1	u_2	u_3	u_4
Antes de la corrección	100%	72,5%	51,4%	70,8%
Después de la corrección	100%	100%	100%	100%

Esta conclusión se confirma al completar el estudio calculando los ángulos medios entre vectores originales y simulados del mismo rango (Tabla 2). A partir de los resultados "brutos" del BT (es decir, antes de la corrección) se podría deducir que, a excepción del primero, los vectores presentan una considerable inestabilidad. Sin embargo, una vez efectuada la corrección, se observa que los cuatro ejes poseen un grado de estabilidad muy elevado (la inestabilidad de los vectores 2, 3 y 4 era aparente y no real). Así, el ángulo medio entre el segundo vector original y el segundo vector simulado es tan solo de 1,41º y en las restantes dimensiones se han computado ángulos medios similares.

Tabla 2.
Ángulos medios entre vectores originales y simulados del mismo rango en 5000 simulaciones

	Ángulos medios			
	u_1	u_2	u_3	u_4
Antes de la corrección	9,68º	23,00º	29,42º	25,66º
Después de la corrección	1,96º	1,41º	1,82º	1,61º

En definitiva, frente a un primer vector estable y tres aparentemente inestables, la utilización de un Bootstrap Total corregido demuestra que la estabilidad de los cuatro ejes es muy elevada.

Conclusiones

Hemos sugerido que la simulación de muestras es un medio adecuado de validación de resultados en análisis de naturaleza exploratoria como el AFC. El método de remuestreo Bootstrap, aplicado con algunas correcciones, ofrece una forma de valorar el grado de estabilidad de las posiciones de las categorías en los planos (Bootstrap Parcial) y de cuantificar el grado de estabilidad de los ejes factoriales (Bootstrap Total).

Las regiones de confianza obtenidas por Bootstrap Parcial permiten comparar la estabilidad de filas y columnas, en nuestro caso unidades léxicas y grupos de respuesta. Interesa analizar tanto el tamaño de las regiones como los solapamientos entre ellas. Regiones sin solapar indican perfiles diferenciados. Regiones parcialmente solapadas señalan perfiles compartidos y un solapamiento elevado muestra perfiles similares.

Las regiones Bootstrap de los grupos de encuestados permiten comparar sus perfiles léxicos, es decir, determinar si utilizan vocablos similares en sus respuestas. Por otro lado, el estudio de las regiones de las unidades léxicas permite comparar perfiles de respuesta, esto es, delimitar si determinados vocablos se pueden vincular a ciertos grupos de encuestados.

La aplicación del Bootstrap Total proporciona una medida del nivel de estabilidad de los factores. La comparación de los resultados simulados con los resultados originales solamente puede realizarse cuando se elimina la variabilidad aparente de los factores y se considera únicamente su variabilidad real. Por lo tanto, desde el punto de vista de la estabilidad podemos calificar un factor como estable, como aparentemente inestable pero realmente estable o como un factor inestable.

Referencias

Álvarez, R.; Ayuso, M.; Bécue, M; Guillén, M.; Hernández, M. L.; Santolino, M. y Valencia, O. (2003). *Jueces en su primer destino 2002*. Análisis Estadístico de las Encuestas realizadas a Jueces en sus primeros destinos (Promociones 48/49 y 50) y Análisis Comparativo con Jueces de mayor experiencia. Barcelona: Informe de resultados nº 2, elaborado para la Escuela Judicial de Barcelona.

Álvarez, R.; Bécue, M.; Valencia, O. (2006). Partial Bootstrap in Correspondence Analysis. Correction of the coordinates. *Proceedings of the 8th International Conference on Textual Data Statistical Analysis*, 43-53. Viprey, J. M.; Condé, C.; Lelu, A. y Silberztein, M. (eds.). Besançon: Presses Universitaires de Franche-Comté.

Beck-Nielsen, H.; Brusgaard, K.; Gaster, M.; Hansen, L.; Hemmings, B.; Kruse, T.A. ; Oakeley, E. y Tan, Q. (2004). Correspondence analysis of microarray time-course data in case-control design. *Journal of Biomedical Informatics*, 37, 358-365.

Efron, B. y Tibshirani, R. (1993). *An Introduction to the Bootstrap.* New York: Chapman & Hall.

Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester: J. Wiley & Sons Ltd.

Greenacre, M. (1984). *Theory and Applications of Correspondence Analysis*. London: Academic Press.

Lebart, L. (2004). Validité des visualisations de données textuelles. *Le poids des mots. Proceedings of the 7th International Conference on Textual Data Statistical Analysis*, 708-715. Purnelle, G.; Fairon, C. y Dister, A. (eds). Louvain: Presses Universitaires de Louvain.

Milan, L. y Whittaker, J. (1995). Application of the parametric Bootstrap to models that incorporate a singular value decomposition. *Applied Statistics*, 44, 31-49.

Reiczigel, J. (1996). Bootstrap tests in Correspondence Analysis. *Applied Stochastic Models and Data Analysis*, 12, 107-117.

Ringrose, T. J. (1992). Bootstrapping and Correspondence Analysis in Archaeology. *Journal of Archaeological Science,* 19, 615-629.

Valencia, O. (2006). *Estabilidad de los métodos factoriales mediante procedimi*entos de remuestreo. Aplicación al Análisis de Correspondencias de tablas léxicas. León: Servicio de Publicaciones de la Universidad de León.