LA CONSTRUCCIÓN DE UN ÍNDICE CUANTITATIVO SOBRE EDUCACIÓN SUPERIOR UTILIZANDO LA TÉCNICA DE ANÁLISIS DE COMPONENTES PRINCIPALES

Jesús Francisco Estévez García*

Introducción: los posibles usos del ACP en el análisis de diferentes características sobre la educación superior

El ACP puede ser empleado para la consecución de diferentes objetivos, entre los que (sin pretensión alguna de ser exhaustivos) podemos señalar los siguientes[1]:

Exploración de relaciones hipotéticas entre variables (análisis exploratorio). El análisis exploratorio a través de ACP, que se ilustra en las siguientes páginas de este artículo, supone un procedimiento inductivo, que permite establecer hipótesis probabilísticas a partir de los datos empíricos. El investigador dispone, por ejemplo, de una matriz de datos con un conjunto de variables[2]que, postula, se encuentran relacionadas con la eficacia observada (a la vista del número de graduaciones, su competitividad y otros indicadores) en una serie de instituciones educativas, pero no está seguro de la forma en que dichas variables se relacionan con el concepto teórico que emplea (la eficacia de tales instituciones para cumplir su cometido). El ACP exploratorio puede construir un puente entre la teoría y los datos, si el investigador exige que el análisis muestre libremente las dimensiones en la que se agruparían las diferentes variables a partir de la matriz de correlaciones original, lo que a su vez puede descubrirle influencias insospechadas que le permitan plantear hipótesis sobre el motivo por el que los componentes subyacentes generan tales relaciones.

Descripción y clasificación. Por otra parte, puede realizarse un uso más sencillo de la técnica si el investigador únicamente necesita simplificar y reducir los datos para hacerlos comprensibles. El ACP, en la misma línea exploratoria, puede ser empleado para obtener una tipología empírica, agrupando variables interdepen-

* Universidad Autónoma Metropolitana (Azcapotzalco). Correo-e: jfeg@correo.azc.uam.mx

[1] Para una descripción más amplia y detallada de los argumentos que aquí exponemos puede verse Schwartzman (1977: 39 y ss).

[2] Existen, según nuestra experiencia investigadora, tres formas de entender la relación entre las variables y los indicadores: a) los indicadores, que proceden de las variables en la operacionalización, especifican en mayor grado los conceptos y posibilitan la medición precisa; b) los indicadores son los valores resultantes de la medición; c) los indicadores son idénticos a las variables, pero el último término se utiliza en sentido abstracto en el desarrollo de disciplinas como la matemática o la estadística, mientras el primero se emplea cuando efectivamente se mide un fenómeno, por ejemplo social, que realmente *está siendo representado* por una vinculación entre los indicadores y la teoría que los sustenta. En el presente artículo optamos por identificar ambos términos, es decir nos adherimos a la última versión de las aquí señaladas.

dientes de forma que los componentes puedan ser interpretados como categorías descriptivas que permitan, por ejemplo, clasificar instituciones educativas por sus similares perfiles en cuanto a la dedicación lectiva del alumnado, la calidad de las titulaciones que los potenciales usuarios perciben subjetivamente o la disponibilidad de recursos materiales o humanos que poseen. Esto capacitaría al investigador para describir logros o problemas considerados de forma específica, y para establecer comparaciones partiendo de dimensiones concretas del quehacer institucional que no habían sido observadas teóricamente.

Comprobación de hipótesis multivariables (análisis confirmatorio). Complementariamente a lo argumentado en los puntos anteriores, el investigador puede tener alguna hipótesis previa sobre las relaciones existentes entre las diferentes variables, postulando una vinculación entre diferentes indicadores, como la proporción de profesores con el título de doctor, el número de horas lectivas destinadas a la ejecución de prácticas de campo, la disponibilidad de laboratorios o la extracción socioeconómica del alumnado, y algunas dimensiones teóricas previstas, como "eficacia institucional", "adecuación del procedimiento educativo" o cualquier otro concepto que considere pertinente. Un ACP confirmatorio descubrirá si en realidad surgen componentes que corroboren inicialmente la hipótesis (esto es, que puedan ser etiquetados como "eficacia", "adecuación"), o si realmente tal vinculación no existe o es poco clara.

Identificación de estructuras latentes. El ACP, o más adecuadamente el Análisis Factorial (AF), puede ser empleado para identificar la estructura básica subyacente a los datos empíricos recopilados sobre un campo temático determinado, a partir de la covarianza entre variables interdependientes[3]. La vinculación que el análisis efectúa entre grupos de variables descubriría al investigador, en ese sentido, que existen dimensiones empíricas recopiladas junto a sus datos, pero no directamente observables, susceptibles de estar provocando los fenómenos que desea describir, más allá de la observación directa de los indicadores inicialmente medidos (Schwartzman, 1977: 44; Tinsley & Brown, 2000: 268 y ss.). De esa manera, es posible encontrar que la eficacia de los centros educativos para producir una alta tasa de graduación se relaciona fundamentalmente con dos o tres rasgos, derivados de los datos empíricos pero no directamente observables, que articulan, y subyacen a, una multitud de características cuyo número y complejidad hubiera impedido al investigador obtener conclusiones precisas sobre sus relaciones concretas con dicha tasa[4].

Transformación funcional de los datos. En ocasiones se utiliza el ACP, muy certeramente, como una técnica auxiliar para emplear los datos del investigador en otros análisis, al efectuar una transformación que asegura el cumplimiento de determina-

[3] No estableceremos explícitamente aquí las diferencias existentes entre el ACP y el Análisis Factorial (AF), en la medida en que si bien el AF "supone que son los factores los que causan la correlación entre las variables originales [tratando de] determinar la estructura de los datos", mientras el modelo ACP no efectúa supuesto alguno al respecto, en la práctica, sin embargo, "estos dos análisis se pueden usar indistintamente, ya que dan resultados similares" (Camacho, 1995: 13).

[4] Nótese que un análisis multivariable de dependencia, como el de regresión múltiple, exige al investigador identificar previamente una serie de variables independientes; lo que el ACP ofrecería como resultado, interpretado en la línea que aquí hablamos, es un conjunto de variables independientes inobservadas que *se ocultan* en el conjunto de nuestros datos y que pueden ser introducidas posteriormente, junto a otras, en un análisis de regresión.

dos supuestos restrictivos. El análisis de regresión múltiple, por ejemplo, supone que no existe multicolinealidad entre las variables independientes o regresoras, esto es: que dichas variables no se encuentran altamente correlacionadas, lo que generaría una gran indeterminación sobre los efectos que, en realidad, cada una de ellas provoca sobre la variable dependiente (Greene, 1997: 418 y ss.; Pindyck & Rubinfeld, 1984: 87 y ss.), produciendo coeficientes mínimo cuadráticos no definidos cuyos estimadores pueden incluso aparecer con el signo contrario al que cabría esperar (Castro y Galindo, 2000: 281 y ss.). El ACP permite superar este problema, al generar componentes ortogonales, incorrelacionados, de forma que al introducir en un modelo de regresión tales componentes, y no las variables originales, nos aseguramos de que se cumpla el supuesto prescrito en el caso de esta última técnica. Nótese, por ejemplo, que si deseamos explicar (y/o predecir) el ingreso que los sujetos perciben por su trabajo a partir de variables como los años cursados de educación superior y el prestigio de la ocupación que poseen, cabe una elevada probabilidad de que la primera variable y la segunda se encuentren positiva y altamente correlacionadas, problema subsanable al emplear un ACP y obtener un componente único, o sea, una nueva variable que integra a las dos anteriores.

Formación de escalas. El investigador puede desear construir una escala para evaluar y/o comparar instituciones educativas a partir de rasgos cuantitativos referentes a, por ejemplo, los recursos humanos disponibles, la infraestructura técnica y los procedimientos empleados para impartir materias concretas. Al respecto, una de las decisiones más complejas para la construcción de tal escala es la ponderación que deben recibir las diferentes características que la integran. El ACP ofrece una solución al dividir estas características en fuentes independientes de variación; esto es: los componentes, que podemos considerar, de esta forma, índices en sí (Cortés y Rubalcava, 1993: 238), aditivos y construidos sobre una base objetiva, empírica (Hair *et al*., 1999: 11).

En el presente artículo se revisan los usos que el ACP puede tener para explorar relaciones hipotéticas entre siete indicadores sobre educación superior, identificar estructuras latentes en los mismos y construir escalas que resuman parcialmente la información contenida en todos ellos, pero esta revisión es un medio, y no un fin en sí misma. La elección del ACP –en un primer momento de la construcción de un índice sobre el desempeño institucional en educación superior– se debe a las ventajas que ofrece esta técnica de análisis para reducir un amplio número de variables a un conjunto menor de factores o componentes no correlacionados entre sí que den cuenta, de manera óptima, de los diferentes porcentajes de varianza común existente entre las variables inicialmente introducidas al análisis[5], pero lo que pretendemos, en última instancia, es elaborar un índice sintético que aglutine todas las escalas parciales, es decir, que integre los componentes obtenidos

[5] En un momento posterior del cálculo del índice, algunas técnicas como el Análisis de Conglomerados (que no introduciremos aquí por limitaciones de espacio) permitirían la identificación de intervalos de calidad, desempeño, eficacia, o lo que se desee medir, con el fin de clasificar el conjunto de instituciones educativas consideradas, garantizando a un mismo tiempo dos condiciones: i) la mínima varianza dentro de un mismo intervalo de casos, y ii) la máxima varianza entre los distintos estratos construidos, de tal forma que se logre una óptima discriminación de los grupos y de la mayoría de los casos incluidos en éstos.

a través del análisis, de forma legítima. Obtendríamos así un solo índice final que permita la clasificación, el ordenamiento y la estimación de las diferencias entre hipotéticas instituciones a través de un único valor que represente, convendremos más adelante, al concepto "eficacia institucional". El procedimiento para llegar a este resultado es relativamente sencillo; lo costoso, inicialmente, es explicar la razón por la que resulta válido de la manera en la que lo planteamos, y no de otra[6].

Algunos requisitos antes de efectuar un análisis de componentes principales

Antes de comenzar, recordaremos que, previamente a la aplicación del análisis que aquí planteamos y tal como ocurre con nuestro quehacer investigador habitual, requerimos un marco teórico que oriente nuestra labor técnica, así como la construcción de indicadores fácilmente determinables que se encuentren lógicamente relacionados con los conceptos que empleamos, y que sean válidos y fiables. Al hablar de indicadores fácilmente determinables nos referimos a que resulte más sencillo detectar su presencia o ausencia (si es dicotómico) y su ubicación, su valor (si es métrico) que el concepto original en su totalidad, medido integralmente (Mora y Araujo *et al.*, 1971: 127). Por otra parte, dichos indicadores también deben corresponder razonablemente con el universo de características más amplio que empleamos cuando tenemos en cuenta el concepto original, lo que supone a su vez un cierto criterio de validez: deben orientarse a medir lo que afirmamos estar midiendo. Unido a esto, es necesario atender a la fiabilidad de los mismos: difícilmente nos servirán los indicadores para algo si cada vez que los empleamos en un conjunto de casos con similares características arrojan resultados dispares.

Una vez que disponemos de variables que cumplen como mínimo las características señaladas existen pocos requisitos para realizar un ACP. Uno de ellos es que los indicadores introducidos en el análisis sean métricos, aunque es posible emplear algunas variables ficticias[7]. Esto es lógico, en la medida en que la técnica trabaja con las matrices de covarianzas y de correlaciones, lo que exige valores significativos, y no códigos arbitrarios. Se requiere, por otra parte, que el número de casos en la matriz de datos no sea inferior a 50, y que supere los 100 preferentemente; en todo caso, el analista debe asegurarse de que posee, como mínimo, un número de observaciones cinco veces mayor que el número de variables empleadas.

Por otra parte, existen varias ventajas para llevar a cabo el ACP, a saber: a) podemos obviar los supuestos de normalidad, homocedasticidad y linealidad, siempre que tengamos en cuenta que "su incumplimiento llevará a una disminución en las correlaciones observadas" (Hair *et al.*, 1999: 88). Esto es relevante, a la vista

[6] Tanto en las obras que citamos aquí como en aquéllas sobre las que, aún habiendo sido revisadas, no hemos hecho referencia, se omite este procedimiento. En varias de ellas se admite la interpretación de los diferentes componentes extraídos del ACP como índices en sí mismos, pero en ninguna se expone procedimiento alguno para la combinación de tales índices parciales en un único índice final, comprehensivo, que dé cuenta del conjunto de componentes, lo que supone un paso adicional más allá del procesamiento con las variables originales.

[7] Si todas las variables son ficticias o dicotómicas, esto es, únicamente poseen los valores 0 y 1, es necesario efectuar otro tipo de análisis factorial, denominado Boolean. En todo caso, SPSS 10.0 dispone de la opción, aunque no es utilizada con frecuencia.

del problema que generalmente causa a los investigadores en educación, y más generalmente en la investigación social, el cumplimiento de tales requisitos (por ejemplo, en análisis como el discriminante o el de regresión), dada la naturaleza de las variables empleadas; b) puede obviarse el problema de multicolinealidad: de hecho, es deseable que exista en un cierto grado de colinealidad para que sea eficaz el uso de este tipo de análisis (Hair *et al.*, 1999: 88).

Finalmente, lo que requerimos al finalizar el ACP es que los resultados sean parsimónicos e interpretables. Esto es, el número de componentes retenidos debe perder la menor información sobre las variables originales, pero también debe ser lo más reducido posible dentro del conjunto de soluciones potenciales, y los componentes deben ser susceptibles de interpretación sustantiva.

Construyendo un índice de desempeño institucional

La selección de las variables[8]

Una vez que damos por supuesto que cumplimos los requisitos señalados, comenzaremos indicando que la primera de las cuestiones al plantear un ACP es la determinación de las variables que van a ser introducidas en el análisis. Disponemos de una matriz que contiene 150 casos de instituciones educativas del país, y en la que se ha introducido la información sobre 20 características de tales instituciones[9]. Nuestra pretensión es obtener un índice de eficacia en la consecución de los objetivos de educación propuestos para el común de organizaciones educativas del país, según los documentos al respecto revisados con anterioridad. Aunque no tenemos claro la forma en que se relacionan, creemos que entre nuestro listado de indicadores se encuentran 7 variables susceptibles de ser empleadas para construir tal índice. No obstante, aunque así fuera, no tenemos forma de efectuar ponderaciones fundamentadas, legítimas, para otorgar el peso específico que cada indicador debería poseer en el índice final, de manera que optamos por efectuar un ACP.

Para esta ilustración de la técnica se han seleccionado, a modo de ejemplo, los siguientes indicadores[10]:

TR: Tasa de rendimiento escolar. Esta tasa se calcula para cada curso académico específico, como:

$$TR=\frac{CS_{ucp}}{CM_{ucp}}$$

siendo CS_{ucp} el número de créditos superados en la universidad (u) durante un curso (c) en el periodo (p). Cuando el valor se aproxime a 1 (o a 100, si lo expresamos en porcentajes) supondrá un mayor grado de eficacia del alumnado

[8] Agradecemos ampliamente la invaluable colaboración de María Jesús Pérez García para la inicial elaboración del presente apartado.
[9] A partir de aquí, el procedimiento ha sido efectuado con el paquete estadístico SPSS® 10.0.
[10] Estos indicadores, algunos de ellos reelaborados para su adaptación al sistema educativo mexicano, han sido extraídos del Catálogo propuesto para el sistema universitario español por el Ministerio de Educación, Cultura y Deporte: . Debemos recordar aquí que estamos ilustrando una técnica para llegar a un índice estadístico: el lector podría seleccionar los indicadores que considere convenientes para obtener cualquier índice particular sobre el tema.

y de la institución docente.

TE: Tasa de éxito escolar. Suponemos inicialmente que esta tasa complementa el indicador de rendimiento. Su cálculo se efectúa para cada curso académico, de forma que:

$$TE=\frac{CS_{ucp}}{CE_{ucp}}$$

siendo CE_{ucp} el número de créditos presentados a examen en la universidad (u) durante un curso (c) en el periodo (p).

TG: Tasa de graduación. Esta tasa, que teóricamente podemos considerar de eficacia productiva, expresa la proporción o porcentaje de alumnos egresados por titulación universitaria en el curso académico, en relación con el total de alumnos matriculados en el primer curso por primera vez (nuevo ingreso) en dicha titulación cuando esa cohorte comenzó sus estudios superiores. Su cálculo sería como sigue:

$$TE=\frac{AG_{ucp}}{APM_{u1(p-n)}} \qquad (3)$$

siendo AG_{ucp} el número de alumnos graduados en la universidad (u) durante el curso (c) en el periodo (p); $APM_{uc(p-n)}$ sería el número de alumnos matriculados por primera vez en la universidad (u) en el primer curso (1) en el periodo en que inició la promoción de alumnos que se encuentra en el denominador. Es decir, si "p" representa un determinado trimestre del año 2001 y la carrera dura 5 años (duración que denominamos "n"), el denominador se constituye por los matriculados en ese trimestre en el año 2001-5 = 1996.

PPD: Porcentaje de profesores investigadores con el grado de doctor. Es el cociente entre el número de doctores y el total de profesores de la carrera en una universidad determinada. Este indicador ofrece información sobre el potencial investigador de la plantilla docente, y puede complementar otras mediciones para determinar el perfil investigador del centro educativo, por carreras o en su conjunto.

TPPI: Tasa de participación en proyectos de investigación. Indica la proporción o porcentaje de profesores que participan en proyectos científicos competitivos que posean procesos rigurosos de evaluación (por ejemplo, los proyectos financiados por CONACyT). Simplemente es el cociente entre profesores que participan respecto al total de profesores en un periodo determinado por universidad y carrera.

PEP: Horas de práctica en empresas, según los planes de estudio. Puede indicar la orientación práctica del centro, o de determinadas carreras. Evidentemente, señala la vinculación entre la institución educativa, que representa la oferta de potenciales empleados cualificados y el mercado laboral, que representa la demanda de tales trabajadores. Se calcula como el cociente entre las horas prácticas del alumnado en empresas y las horas totales de prácticas (o bien las horas lectivas totales de la carrera).

PROD_D: Producción de doctores. Este indicador puede sugerir el nivel de implicación del profesorado en la docencia de posgrado y en la investigación. Se calcula como el cociente entre el número de tesis defendidas con éxito por periodo respecto al número total de doctores que laboran en la institución en ese periodo.

La matriz de comunalidades, la varianza explicada y la retención de los factores

Al programar y efectuar el análisis obtenemos la siguiente matriz de comunalidades –*Communalities*– que expresa *la parte de la varianza de cada una de las variables que es explicada por los componentes en su conjunto*, incluye dos valores para cada una de las variables: el primero de ellos, 1.000 en todos los casos, indica que *si fueran retenidos todos los factores posibles* –es decir, tantos como variables: en este caso 7– la varianza de cada una de las variables sería totalmente explicada por los factores en su conjunto[11,12]. El segundo valor, en la tercera columna de la tabla, hace referencia a la parte de la varianza de cada variable que es explicada por los factores que son *finalmente* retenidos.

Cuadro 1
Comunalidades

Como puede observarse en la matriz anterior, TPPI y PEP son las variables mejor explicadas por el conjunto de los factores retenidos tras el análisis: en ambos casos, la varianza explicada para cada una de las variables supera el 83%. En sentido inverso, TR y PROD_D ofrecen las menores proporciones de varianza explicada por los factores: .343 en el primer caso y .212 en el segundo.

Para un análisis más detallado de lo que apuntamos, consideremos la tabla que insertamos a continuación –Total Variance Explained–, en la que pueden observarse seis columnas distribuidas en tres grandes grupos: *Initial Eigenvalues*, *Extraction Sums of Squared Loadings* y *Rotation Sums of Squared Loading*.

Cuadro 2
Varianza total explicada

Communalities	Initial	Extraction
TR	1	0.343211
TE	1	0.519374
TG	1	0.626860
PPD	1	0.606157
TPPI	1	0.834457
PEP	1	0.841634
PROD_D	1	0.212287

Extraction Method: Principal Component Analysis.

[11] El programa a ejecutar en SPSS es:
FACTOR
/VARIABLES tr te tg ppd tppi pep prod_d /MISSING LISTWISE /ANALYSIS tr te tg ppd tppi pep prod_d
/PRINT INITIAL EXTRACTION ROTATION FSCORE /PLOT EIGEN ROTATION
/CRITERIA MINEIGEN(1) ITERATE(25) /EXTRACTION PC
/CRITERIA ITERATE(25) /ROTATION VARIMAX /SAVE REG(ALL)
/METHOD=CORRELATION .

[12] Lógicamente, el valor es 1.000 debido a que las variables son previamente estandarizadas para ejecutar el ACP, de manera que todas ellas adoptan una media de 0.000 y una varianza igual a 1.000.

Analizando el grupo de "Valores propios iniciales" –*Initial Eigenvalues*– puede apreciarse:

a. El *valor propio*, o parte de la varianza de las variables inicialmente consideradas, que explica cada uno de los componentes que arroja el ACP. Este valor propio es superior a 1.000 en los dos primeros casos, lo que significa, laxamente, que cada uno de estos componentes *explicaría* la varianza de más de una variable[13].
b. El porcentaje de la varianza —de todas las variables inicialmente introducidas en el análisis— explicado por cada uno de los componentes. De manera coherente con los valores propios que aparecen en la columna anterior, nótese cómo dicho porcentaje alcanza sus máximos valores en los dos primeros componentes: el primero explicaría el 33.285% de la varianza total del modelo; el segundo, el 23.629%. Al estar

Total Variance Explained	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.329980	33.285432	33.285432	2.329980	33.285432	33.285432	2.278801	32.554301	32.554301
2	1.654000	23.628577	56.914009	1.654000	23.628577	56.914009	1.705180	24.359708	56.914009
3	0.906843	12.954895	69.868905						
4	0.775274	11.075341	80.944246						
5	0.558976	7.985369	88.929615						
6	0.493941	7.056296	95.985911						
7	0.280986	4.014089	100.000000						

Extraction Method: Principal Component Analysis.

ordenados dichos componentes de manera descendente, la varianza del modelo estaría menos explicada por los componentes situados hacia el final de la tabla[14].
c. El porcentaje acumulado de la varianza explicada, que, como es obvio, va incrementándose conforme se suman los porcentajes de varianza explicados por los componentes adicionales.

Los dos últimos grupos de la tabla anterior, Extracción y rotación de las sumas de cargas al cuadrado —*Extraction Sums of Squared Loadings* y *Rotation Sums of Squared Loading*—, hacen referencia a los porcentajes de varianza explicada tras la retención de los componentes o dimensiones latentes a las variables inicialmente consideradas, por lo que únicamente se muestran los datos correspondientes a los dos primeros componentes, cuyos valores propios son superiores a 1, tal y como indica la regla de Kaiser. A idénticas conclusiones podemos llegar si atendemos al gráfico insertado más abajo, donde se representan cada uno de los 7 componentes con relación a su valor propio:

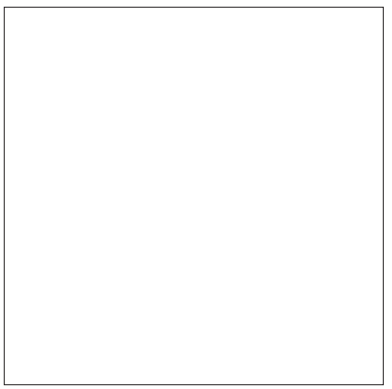
[13] Recordemos aquí que cada varianza es igual a 1, dada la estandarización previa.
[14] Obviamente, la columna a la que nos referimos se obtiene de dividir la varianza explicada por cada componente en números absolutos –véase la columna "Total"– entre la suma de la varianza del conjunto de las variables, en este caso 7.

Gráfico 1
Valores propios por componente

Si trazamos una línea horizontal paralela al eje de abscisas y que corte al eje de ordenadas donde el valor propio o *autovalor* es 1.0, y retenemos aquellos componentes cuyo valor propio se encuentra por encima de dicha recta, seleccionaremos finalmente los dos primeros componentes, tal y como habíamos concluido anteriormente. El número de componentes retenidos, subyacentes al conjunto de las 7 variables que introdujimos al principio, sería, por lo tanto, 2, y entre ambos explicarían el 56.914% de la varianza del modelo. Ahora bien, es posible considerar algunas diferencias en el porcentaje de varianza que explican ambos componentes, de manera independiente, antes y después de la rotación, como veremos enseguida.

La matriz de componentes y la matriz de componentes rotada[15]

La matriz de componentes que aparece más abajo, y que deriva del análisis ACP



llevado a cabo, ofrece la información requerida para una primera agrupación de las variables –antes de la rotación– en torno a los patrones o pautas latentes resultantes.

Cuadro 3
Matriz de componentes

Los valores de los pesos factoriales que aparecen en las celdas de dicha matriz –tras haber suprimido aquéllos iguales o inferiores a 0.30– no dejan lugar a dudas respecto a una inicial correspondencia entre variables y componentes: TR, TE, TG, PPD y PROD_D quedarían explicadas por el primero de los componentes, mientras que TPPI y PEP estarían representadas por el segundo de los componentes, aunque la última tendría una respetable correlación con el primero de los factores.

La relativa indeterminación que podría surgir en algunos casos –como sucede en cierto sentido con PEP– puede ser resuelta si consideramos una segunda matriz, la matriz de componentes rotada, que habitualmente permite una mayor discriminación de las variables con respecto a los componentes. La rotación, proceso que permite interpretar más fácilmente las asociaciones entre las variables y los componentes al lograr que las variables fuertemente correlacionadas entre sí presenten pesos factoriales elevados (en valor absoluto) y las menos correlacionadas obtengan pesos factoriales bajos (igualmente en valor absoluto) arroja en el presente análisis los resultados que se observan a continuación[16]:

	Component	
Component Matrix	**1**	**2**
TR	0.565031	
TE	0.660063	
TG	0.782940	
TPPI	0.752253	
PEP		0.900151
PPD	0.411473	0.819954
PROD_D	0.450179	

Extraction Method: Principal Component Analysis.
2 components extracted.

Cuadro 4
Matriz de componentes rotados

Como puede apreciarse, todas las variables presentan pesos factoriales algo más elevados que en la primera de las matrices, con excepción de la variable PROD_D, que registra un valor ligeramente inferior. En términos generales, podría afirmarse que la rotación efectuada, si bien no ha llevado a modificar la agrupación de las variables en torno a los componentes, sí ha permitido incrementar de manera óptima las saturaciones o proyecciones de las variables sobre dichos componentes. Esta consideración permite, pues, concluir la existencia de dos componentes o patrones subyacentes al conjunto de las siete variables inicialmente introducidas en el análisis que, a nuestro juicio, quedarían dispuestas del siguiente modo:

- Componente 1 o de "Adecuación docente", denominado así porque agruparía a las variables referentes a los resultados obtenidos por las instituciones educativas tanto en la formación de sus alumnos como en el perfil de sus docentes, esto es: alto rendimiento (TR) y éxito escolar (TE), elevada graduación (TG), adecuado perfil

[15] Con el fin de lograr una mayor claridad al agrupar las variables en torno a los componentes o patrones latentes, se ha pedido al programa de análisis estadístico empleado que suprima de la matriz de componentes aquellos pesos factoriales cuyo valor sea igual o inferior a 0.30, por lo que algunas de las celdas de las siguientes matrices aparecerán en blanco.
[16] Como puede leerse al pie de la matriz, se ha efectuado una rotación ortogonal de los factores, llamada Varimax, que maximiza la variación por columnas simplificando la lectura de los componentes al obtener los coeficientes estructura más altos o más bajos posibles —más 1 o menos 1—. Otras opciones de rotación ortogonal podrían ser, por ejemplo, la rotación Quartimax, que maximiza la variación filas, simplificando las variables; y la rotación Equamax, que busca un resultado intermedio.

formativo de los docentes (PPD) y alta producción de doctores (PROD_D), lo que, recordemos, indicaba el nivel de implicación del profesorado en la docencia de posgrado, pero también la dedicación y el éxito de sus estudiantes de doctorado[17].

- Componente 2 o de "Integración competitiva", acuñado de este modo por incluir

Rotated Component Matrix	Component 1	2
TR	0.585804	
TE	0.714185	
TG	0.785117	
TPPI	0.778434	
PEP		0.908196
PPD		0.901522
PROD_D	0.405805	

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
Rotation converged in 3 iterations.

dos variables referidas de manera directa a la vinculación de las instituciones educativas con el entorno científico y profesional, lo que incluye a la red empresarial. Los indicadores en este componente tienen que ver con la numerosa participación en proyectos de investigación (TPPI) y el elevado número de horas que los estudiantes dedican a las prácticas en empresas, según el plan de estudios; esto último, suponemos adicionalmente, permite una adecuación de los conocimientos teóricos recibidos por los mismos, al ser aplicados en el entorno en el que posteriormente se les exigirán resultados.

Nótese que, hasta este momento, el ACP nos ha permitido lo siguiente:

- Descubrir que nuestras siete variables originales pueden encontrarse relacionadas en dos conjuntos.
- Descubrir que tales conjuntos pueden ser etiquetados de forma teóricamente coherente, que han sido descubiertos, dentro de la línea teórica argumental que propusimos, en forma de dimensiones de la adecuación del quehacer de las instituciones o, si se prefiere, de su eficacia.
- Identificar el peso con el que las variables originales se relacionan con cada una de las dimensiones descubiertas.

Ahora, pues, disponemos de dos índices sintéticos: los componentes en sí mismos. Podríamos efectuar comparaciones directas y efectuar mapas que ordenaran a las instituciones por dos criterios: por un lado el de adecuación docente, y por

[17] Nótese que todos los coeficientes de correlación significativos son positivos, lo que indica que todas las variables se relacionan de forma directa: altos niveles de unas se asocian a altos niveles de otras, y viceversa.

otro el de integración competitiva. Nos resta efectuar, según nuestra pretensión principal en el presente artículo, la labor de obtener un índice final que aglutine ambos componentes, con el fin de dar cuenta no sólo de dimensiones parciales de la eficacia institucional, sino del concepto original entendido integralmente, esto es: "la eficacia institucional" en sí misma.

La síntesis de factores en un índice único: algunas notas teóricas para la construcción del índice a partir del ACP

Como hemos señalado, consideraremos la retención de dos factores explicativos derivados del ACP, con el fin de describir lo más claramente posible la forma en que ambos deben combinarse para obtener un único índice. Vamos a describir brevemente el procedimiento lógico que seguimos para la construcción de tal índice, enfatizando las premisas teóricas que lo conectan con otro tipo de análisis multivariable: el de regresión. Posteriormente describiremos este procedimiento en la práctica.

Una vez que hemos llegado a la matriz de componentes rotados estamos obteniendo las *correlaciones bivariadas* entre cada variable y componente del ACP. Estamos obteniendo, correspondientemente, el valor que tendrían los coeficientes beta estandarizados de sendas regresiones lineales simples entre las variables y los factores rotados[18]. Debe hacerse notar aquí que el resultado de los *coeficientes* es exactamente el mismo si utilizamos los siguientes procedimientos para ejecutar la regresión:

Regresiones simples del componente F_1 y, posteriormente, del componente F_2 sobre la variable z_i:

$F_1 = b_{11}z_{1.}$

(4)

$F_2 = b_{12}z_{1.}$

Regresión de la variable z_i sobre los factores F_1 y F_2:

$z_v = b_{11}F_1 + b_{12}F_2 + e.$

(5)

donde:

F_i: *i-ésimo* factor común extraído —dos en nuestro modelo—.

a: la constante que representa el punto de corte con el eje de ordenadas.

b_i: *i-ésimo* coeficiente que mide el efecto del cambio unitario de la variable independiente "i" sobre la dependiente.

z_i: la *i-ésima* variable del modelo.

e: el volumen de error en el análisis de regresión –entre otras cosas, la varianza única y propia de la variable z_i.

La primera pregunta que puede surgir a partir de las anteriores consideraciones es ¿por qué los modelos de regresión lineal simple de la variable z_i sobre el compo-

[18] No es necesario señalar que si deseamos conocer la importancia de la rotación finalmente efectuada por el ACP debemos mirar la Matriz de transformación de componentes —Component Transformation Matrix—, cuyos valores representan los cosenos de los ángulos existentes entre los factores originales y esos mismos componentes una vez rotados.

nente 1 y sobre el 2 no contienen el término de error, mientras que la regresión de los factores sobre la misma variable sí? Nada más sencillo de responder: el valor de cada coeficiente factorial en la matriz de componentes rotados queda enteramente determinado por el valor de cada variable, de manera que ésta explica el 100% de la varianza del componente, dado que este último es una combinación lineal de las variables observadas en cuanto poseen una determinada varianza común. Sin embargo, cada componente solamente explica una porción de la varianza de la variable, debido a que ésta posee un determinado porcentaje de varianza propia y única que podría ser vista como la conformación de una variable inobservada por el modelo[19].

La segunda pregunta, no menos importante, podría ser: ¿por qué arroja el mismo resultado una regresión simple de cada factor por separado sobre una variable que el de ambos factores en una regresión lineal múltiple sobre esa misma variable? Para responder esto solamente debemos traer a colación una premisa indispensable del ACP: los factores rotados por el método VARIMAX son ortogonales, vale decir incorrelacionados. De esta manera, igual da realizar con ellos sendas regresiones simples o una múltiple, dado que no existe efecto alguno compartido que cualquiera de ellos se pueda apropiar si suprimimos del análisis al otro. Este último argumento es fundamental para explicar porqué no podemos utilizar la matriz de componentes rotados para realizar una combinación lineal de indicadores que nos ofrezca como resultado un índice fiable[20].

Al llevar a cabo una regresión lineal simple –que produce los valores de la matriz de componentes rotados– estamos calculando en qué medida la variable dependiente puede ser predicha por la independiente mientras las otras variables intervienen libremente sin ser detectadas como tales, es decir, sin mantener constante los valores de los otros indicadores que pudieran estar afectando el comportamiento de la dependiente[21]. De esta manera, tanto los coeficientes no estandarizados como los estandarizados son la resultante de dos efectos *de la* variable, los que podríamos denominar *propio* y *apropiado*. El problema proviene de este segundo efecto dado que, erróneamente, estaremos concediendo a cada indicador utilizado un mayor peso del que realmente posee para conocer los valores que tomará la variable dependiente —el índice—: el suyo propio y el de las otras variables correlacionadas con él que no mantuvimos constantes en el proceso[22]. De esta forma, estaremos multiplicando los efectos de las variables, de manera que todas se verán *sobrerepresentadas,* pero fundamentalmente las que poseen mayor correlación con otros indicadores en el modelo.

Ahora bien, buscamos una combinación lineal de nuestras variables que evite el problema señalado, y que nos permita llegar a un número concreto –un índice– para cada caso objeto de estudio –las instituciones consideradas individualmente–, para lo que requerimos:

[19] De hecho, el valor en que cada indicador queda explicado por nuestro modelo de dos componentes puede verse, como decimos, en el cuadro "comunalidades". La resta entre los valores de la comunalidad inicial —1.000— y los de la columna "extracción" nos estaría señalando el volumen en que cada variable posee un factor único y no común para su explicación.
[20] Obviamente, las consideraciones que efectuamos sobre la matriz de componentes rotados son igualmente válidas en el caso de la matriz de componentes sin rotar.

1) Tener unos valores iniciales delimitados: los de las variables cuya información deseamos incluir en el índice sintético.
2) Conocer los patrones por los que estas variables se agrupan: los coeficientes adoptados por los factores subyacentes expresados en la matriz de componentes rotados en el ACP.
3) Identificar el peso de cada una de las variables en una potencial combinación lineal cuyo resultado final sea un número, el índice que buscamos, que debe resumir los valores indicados en (1) a través de los patrones de agrupamiento citados en (2).

Este número, el índice, posee la función de resumir para cada caso objeto de estudio el efecto de cada indicador según el volumen preciso en que éste, y solamente éste, determine el valor final adoptado por los dos factores seleccionados en el ACP y, consecuentemente, por el indicador final que representará a los mismos, lo que supone tener en cuenta el porcentaje de varianza compartida entre cada indicador concreto y los factores. Así, la combinación lineal que buscamos debe proceder entonces de manera que:

a) Cada variable se vea representada única y exclusivamente según su propio peso a la vista de los resultados del ACP.
b) Los factores se vean introducidos en el índice ponderados por su capacidad explicativa de las variables originales.

La construcción del índice en la práctica

En el ACP, una vez extraída la matriz rotada de componentes es posible obtener una última matriz: la de coeficientes de los componentes –"Component Score Coefficient Matrix". En ésta se obtienen los coeficientes b estandarizados de dos hipotéticas regresiones lineales múltiples que consideren todas las variables introducidas en el ACP como independientes, y cada factor como dependiente. En este momento –y solamente a partir de este momento– tenemos el peso concreto en el que cada variable determina la puntuación adoptada por ambos componentes para cada institución objeto de estudio o, podríamos decir, la medida en que cada componente representa a cada indicador individual, dado que conocemos el efecto de cada indicador –el influjo de cada componente– manteniendo constantes todos los demás.

Cuadro 5
Matriz de coeficientes de los componentes

Solamente nos falta considerar un elemento para llegar a un índice único óptimo:

[21] En este caso, la variable dependiente es el componente, y los otros indicadores que no observamos son el resto de indicadores que incluimos en el ACP y que evidentemente también determinan el valor del factor considerado.
[22] Más adelante, cuando insertemos la formulación final para la construcción del índice, expondremos formalmente el modelo erróneo al que aquí nos hemos referido.

la medida en que cada factor explica la varianza de las variables introducidas en el ACP o, puede leerse, la medida en que las variables poseen varianzas comunes, que son las que finalmente se verán representadas en el índice[23]. Formalmente necesitamos:

$$I = z_1((a_{11}\lambda_1) + (a_{12}\lambda_2)) + z2((a_{21}\lambda_1) + (a_{22}\lambda_2)) + z_3((a_{31}\lambda_1) + (a_{32}\lambda_2)) + z_4((a_{41}\lambda_1) + (a_{42}\lambda_2)) + z_5((a_{51}\lambda_1) + (a_{52}\lambda_2)) + z_6((a_{61}\lambda_1) + (a_{62}\lambda_2)) + z_7((a_{71}\lambda_1) + (a_{72}\lambda_2)).$$

(6)

Donde la operación $(a_{ji}\lambda_i)$ supone la multiplicación del peso o coeficiente factorial –de la matriz "Component Score Coefficient Matrix"– en la variable j para el Factor i, por el porcentaje de varianza total explicada en el caso de dicho Factor.

Obsérvese, en este sentido, que $((a_{11}\lambda_1) + (a_{12}\lambda_2))$ da lugar a la ponderación por la que finalmente quedará multiplicada la primera variable $–z_1–$, del conjunto de las siete utilizadas para construir el índice integral[24], realizando sucesivamente la misma operación con el resto de indicadores.

En términos matriciales, en los que tanto la lectura como las operaciones a ejecutar pueden quedar más claras, formalizaríamos de la siguiente manera:

$$I = X (A\lambda)$$

Component Score Coefficient Matrix	Component 1	2
TR	0.258890	-0.023231
TE	0.320482	-0.090205
TG	0.342646	0.024020
TPPI	0.343780	-0.027809
PEP	-0.085577	0.541585
PPD	0.033377	0.525196
PROD_D	0.169431	0.110193

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

(7)

donde A sería la matriz de coeficientes factoriales, l el vector que contiene las varianzas explicadas por factor y X nuestra matriz de datos original. De este modo, multiplicamos una matriz A (7 x 2) y un vector λ (2 x 1), lo que nos ofrece como resultado un vector para la ponderación por variable, llamémosle p, de orden 7 x 1. Al postmultiplicar este último por nuestra matriz de datos original –X (150 x 7)– obtenemos el vector con el índice que buscamos para cada institución educativa, de orden 150 x 1. Finalmente, sólo debemos estandarizar los valores de este

[23] Nuestro índice únicamente considerará el valor en que cada variable apunta al mismo fenómeno, a los mismos patrones latentes de nuestro interés, que en este caso hemos denominado "Adecuación docente" e "Integración competitiva". Recordemos que necesitamos Identificar el peso de cada una de las variables en una combinación lineal cuyo resultado final debe resumir los valores de las variables a través de los patrones de agrupamiento generados por los componentes principales.

resultado para obtener nuestro índice de media 0.000 y varianza 1.000.

De esta manera, hemos resumido la información de 7 variables en una única medida, representada por el componente final, que sintetiza los dos componentes originales retenidos en el ACP.

Ahora estamos en condiciones de volver al problema que dejamos irresuelto anteriormente, el que hacía referencia a la expresión formal del error que cometeríamos utilizando los coeficientes de la matriz rotada para construir el índice. Tal construcción daría como resultado la siguiente formulación:

$$I = z_1((a_{11}\gamma_{11}\lambda_1) + (a_{12}\gamma_{12}\lambda_2)) + z_2((a_{21}\gamma_{21}\lambda_1) + (a_{22}\gamma_{22}\lambda_2)) + z_3((a_{31}\gamma_{31}\lambda_1) + (a_{32}\gamma_{32}\lambda_2)) + z_4((a_{41}\gamma_{41}\lambda_1) + (a_{42}\gamma_{42}\lambda_2)) + z_5((a_{51}\gamma_{51}\lambda_1) + (a_{52}\gamma_{52}\lambda_2)) + z_6((a_{61}\gamma_{61}\lambda_1) + (a_{62}\gamma_{62}\lambda_2)) + z_7((a_{71}\gamma_{71}\lambda_1) + (a_{72}\gamma_{72}\lambda_2)). \quad (8)$$

donde $(a_{11}g_{11}l_1)$ estaría constituido por el coeficiente factorial y la varianza explicada por el Factor 1 en el caso de la variable 1, pero a esto se añadiría γ_{11}, que podemos denominar el efecto apropiado por la variable 1 para explicar el Factor 1, debido a que inicialmente sólo tuvimos en cuenta este indicador para efectuar la predicción de dicho factor. De esta manera, con γ_{11} denotamos las correlaciones entre z_1 y todas las otras variables que poseen peso factorial en el primer Factor, correlaciones que son añadidas al modelo cada vez que insertamos un nuevo indicador para calcular el índice. Podemos pensar claramente que la distorsión del resultado final estará en relación, pues, al tamaño de las correlaciones entre variables, debido a que los coeficientes hallados se han obtenido sin mantener constantes el resto de indicadores. El tamaño de tal distorsión vendrá dado, de esta manera, por el peso combinado de todas las g_{ij}.

Para concluir debemos señalar que, a pesar de que nuestros argumentos para legitimar el procedimiento de construcción del índice final obtenido puedan parecer complejos, la forma para llevar a cabo los cálculos en SPSS es ciertamente sencilla. En realidad únicamente necesitamos multiplicar la puntuación factorial guardada para cada caso con la varianza explicada por cada componente y efectuar la sumatoria. El programa, con los valores que aparecen en nuestros resultados, es:

```
COMPUTE índice = (compon1 * 0.33285432) + (compon2 * 0.23628577).
EXECUTE.
```

Posteriormente, sólo se requiere estandarizar el resultado para que nuevamente obtengamos una serie de promedio 0 y varianza 1. Una sencilla orden en relación con las ventajas que se obtienen, en la medida en que su resultado crea un índice único que permite cubrir el conjunto del concepto "eficacia" tal como decidimos

[24] Simbolizamos con "z" debido a que previamente al cálculo del índice hemos estandarizado los valores de cada variable para que posea media igual a 0 y desviación típica –y varianza– de 1.

definirlo. Además, no necesitamos preocuparnos por los límites máximo y mínimo ni por el significado numérico de los valores alcanzados por las diferentes instituciones, en la medida en que su lectura es clara dado el procedimiento empleado: el 0 indica que la institución se encuentra en el promedio de eficacia. Los valores positivos indican una alta eficacia, mayor a medida que crece la puntuación, y los negativos, que esas instituciones concretas se encuentran por debajo del estándar en el conjunto de aquéllas consideradas.

Referencias

CAMACHO ROSALES (1995). *Análisis multivariado con SPSS/PC*+, Barcelona, EUB.

CASTRO POSADA, Juan y Mª Purificación Galindo (2000). *Estadística multivariante. Análisis de correlaciones,* Salamanca, Amarú.

CORTÉS, Fernando y Rosa Mª Rubalcava (1993). "Consideraciones sobre el uso de la estadística en las ciencias sociales. Estar a la moda o pensar un poco", en Méndez, Ignacio y Pablo González Casanova (Coords.), *Matemáticas y Ciencias sociales*, México, Miguel Ángel Porrúa.

GREENE, William H. (1993). *Econometric Analysis*, London, Prentice Hall.

HAIR, Anderson, Tatham & Black (1999). *Análisis multivariante,* Madrid, Prentice Hall.

Ministerio de Educación, Cultura y Deporte de España, *Catálogo de indicadores del sistema universitario español*, http://www.mec.es/.

MORA Y ARAUJO, Manuel y Paul Lazarsfeld *et al.* (1971). *Medición y construcción de índices,* Buenos Aires, Nueva Visión.

PINDYCK, Robert S. & Daniel L. Rubinfeld (1984). *Econometric Models and Economic Forecast,* Auckland, McGraw-Hill.

SCHWARTZMAN, Simón (Comp.) (1977). *Técnicas avanzadas en ciencias sociales,* Buenos Aires, Nueva Visión.

TINSLEY, Howard & Steven Brown (2000). *Applied Multivariate Statistics and Mathematical Modeling,* California, Academic Press.