REVISTA DE MÉTODOS CUANTITATIVOS PARA LA ECONOMÍA Y LA EMPRESA (24). Páginas 250–269.
Diciembre de 2017. ISSN: 1886-516X. D.L: SE-2927-06.
www.upo.es/revistas/index.php/RevMetCuant/article/view/2889

Análisis cualitativo comparativo difuso para determinar influencias entre variables socio-económicas y el rendimiento académico de los universitarios

Fedriani Martel, Eugenio M.
Departamento de Economía, Métodos Cuantitativos e Historia Económica
Universidad Pablo de Olavide (España)
Correo electrónico: efedmar@upo.es

Romano Paguillo, Inmaculada
Departamento de Economía, Métodos Cuantitativos e Historia Económica
Universidad Pablo de Olavide (España)
Correo electrónico: iromano@upo.es

RESUMEN

El objetivo de este artículo es explicar el rendimiento académico con la ayuda de una técnica novedosa: el análisis cualitativo comparativo difuso. Para hacerlo posible, se consideran diferentes variables que afectan a la educación superior, así como el rendimiento académico de los estudiantes universitarios. Los datos utilizados provienen de los diferentes grados impartidos en la Facultad de Ciencias Empresariales de la Universidad Pablo de Olavide, de Sevilla (España), desde el año 2009, aunque la técnica utilizada puede ser fácilmente adaptada a otros colectivos y situaciones.

Palabras claves: economía de la educación; educación superior; análisis cualitativo comparativo; espacio europeo de educación superior.
Clasificación JEL: I21; C02.
MSC2010: 91B15; 03B52.

Artículo recibido el 15 de diciembre de 2016 y aceptado el 19 de julio de 2017.

Fuzzy-Set Qualitative Comparative Analysis to Determine Effects from Socio-Economical Factors and University Students Performance

ABSTRACT

The objective of this paper is to explain academic performance with the aid of an innovative technique: Fuzzy-set Qualitative Comparative Analysis (fsQCA). To do so, different objective factors, which affect higher education, are analyzed and the academic performance of university students is also considered. Specifically, data of students on different degrees in the Faculty of Business Sciences of the Pablo de Olavide University of Seville since 2009 was used, but it is believed that the methodology described can be easily extrapolated.

Keywords: education economics; higher education; qualitative comparative analysis; European higher education area.
JEL classification: I21; C02.
MSC2010: 91B15; 03B52.

1. Introducción

Es indudable que existe una relación compleja entre Economía y Educación, pues son ámbitos que se afectan mutuamente y a través de múltiples variables interrelacionadas. Este hecho ha propiciado la aparición de numerosos trabajos de investigación que tratan de desentrañar tales influencias.

Por una parte, el interés por el análisis del rendimiento académico de los estudiantes viene justificado por la importancia de una buena elección de la carrera universitaria; un estudiante que elige con criterio tiene una mayor probabilidad de acabar con éxito, tanto en su vida académica como en su vida profesional, lográndose una mayor rentabilidad social y económica. De hecho, distintos autores (como [2], [3] y [14]) se reafirman en la rentabilidad social que genera una educación universitaria adecuada.

Por otra parte, en general, una mejora en la tasa de éxito de los estudiantes repercutiría en una disminución en los costes de las universidades y, con ello, en un ahorro económico a escala regional o nacional en los estudios universitarios [5]. Dicha mejora sería posible de conocerse las variables que determinan una gran parte de los resultados académicos finales y la dirección de su influencia en cada caso.

Abundando en lo anterior, las implicaciones de una buena elección de la carrera académica (a la hora de la matrícula universitaria) y la profesional van más allá del significado de una inversión en educación del propio individuo y, por ello, este factor (elegir correctamente la carrera) es muy influyente tanto en la economía universitaria como en diferentes aspectos considerados y analizados por la Economía de la Educación.

De todo esto se puede deducir que es relevante detectar los factores que explican el éxito o fracaso de los estudiantes, sobre todo en el primer curso académico (porque el primer año puede marcar el resto de la carrera y, además, porque es sobre el que se puede actuar más directamente mediante una modificación en la elección de la misma, por ejemplo).

Son numerosos los trabajos que se han desarrollado sobre los factores específicos que influyen en Educación; por ejemplo, se pueden consultar: [4], [12], [16] y [6]. En particular, varios de ellos analizan a estudiantes que estudian asignaturas de Matemáticas para Economía o Empresa; el motivo principal es que estas asignaturas son decisivas para el éxito académico y afectan como ninguna otra en el abandono académico prematuro.

En lo que respecta a las variables previas al ingreso en la Universidad, distintos autores han publicado diferentes trabajos sobre las asignaturas que el estudiante debería haber cursado antes de ingresar; en particular, las asignaturas de Matemáticas resultan esenciales para estudiantes del Grado en Economía, del Grado en Administración y Dirección de Empresas, etc. (cualquiera que sea la institución analizada) [6].

El principal objetivo de este trabajo es, consecuentemente, encontrar combinaciones de factores (o condiciones) que permiten garantizar el éxito académico de los estudiantes universitarios (resultado). Para ello, en este trabajo se consideran específicamente las variables y los datos descritos en [7], que se refieren a un colectivo formado por 1492 estudiantes de la Facultad de Ciencias Empresariales de la Universidad Pablo de Olavide, de Sevilla (España). Concretamente, la siguiente sección se dedica a una exposición resumida de los datos y variables consideradas. En otra sección se explica posteriormente la técnica utilizada para el análisis de dichos datos. Justo después se recoge la aplicación de la técnica a los datos y el artículo concluye con algunas de las conclusiones más destacadas del análisis.

2. Datos

Los datos utilizados en esta investigación han sido recopilados de distintas fuentes. Los datos económicos, asociados a los estudiantes a través de sus domicilios, se han consultado en distintas bases de datos y actualizados a partir del Instituto Nacional de Estadística (INE) [9, 10, 11], del Instituto Geográfico Nacional (IGN)(`www.ign.es`), del Instituto de Estadística y Cartografía de Andalucía (IECA) [8], de los Ayuntamientos de Sevilla y Dos Hermanas, de la herramienta web Google Maps `(www.google.es/maps)` y la página `(www.codigo-postal.info/sevilla/sevilla/7)`.

A partir de la información recopilada, se han desarrollado otras variables económicas definidas *ad hoc* y verificadas por expertos en la materia, atendiendo a las necesidades que el problema que se deseaba resolver sugería.

Por otra parte, las variables de carácter académico se han obtenido gracias a la colaboración de los profesores y coordinadores de las asignaturas del Área de Métodos Cuantitativos del Departamento de Economía, Métodos Cuantitativos e Historia Económica de la Universidad Pablo de Olavide, de Sevilla (UPO). Los datos de acceso e información previa de los estudiantes se han con-

seguido gracias al Área de Estudiantes y al Área de Gestión de Matrícula y Expediente Académico de Grado de la misma UPO.

Los datos académicos o educativos (a los que se acaba de hacer referencia) corresponden a un conjunto de 1492 estudiantes de Grado de la UPO, en concreto, de la Facultad de Ciencias Empresariales (FCE). Este grupo de estudiantes está delimitado por ser alumnos de las asignaturas obligatorias que imparte el Área de Métodos Cuantitativos del Departamento de Economía, Métodos Cuantitativos e Historia Económica en la FCE. A partir de las calificaciones obtenidas en dichas asignaturas, se calcula un índice del rendimiento académico de cada estudiante, según se explicita en [7]. *Grosso modo*, el indicador consiste en calcular la media aritmética de las notas conseguidas por el estudiante y multiplicarla por un "coeficiente penalizador", que depende del grupo en el que dicho estudiante es clasificado según una red neuronal artificial entrenada para generar subgrupos casi-homogéneos por número de asignaturas matriculadas y convocatorias agotadas. Una vez normalizadas las calificaciones entre 0 y 1, puesto que el coeficiente penalizador también está entre 0 y 1, los valores del indicador se mueven teóricamente entre 0 y 1.

Además de los valores del indicador, se conocen (entre otras) las siguientes características personales y educativas previas de los estudiantes. Todas ellas son conocidas a través de un formulario sencillo previo a que los estudiantes comiencen sus estudios universitarios. No obstante, en el conjunto de datos utilizado para este trabajo solo hay 858 casos en los que se conocen con fiabilidad los valores de todas y cada una de las variables. En el siguiente listado de variables, se indica entre corchetes el nombre original de cada variable, aunque algunas de ellas hubieron de recodificarse posteriormente (por lo que pueden no coincidir exactamente con las denominaciones de las tablas de resultados).

- Sexo *[Sexo]*: variable que toma los valores *mujer* y *hombre*, con un único valor posible para cada estudiante si bien puede haber datos perdidos.
- Edad *[Edad]*: valor numérico (creado a partir de la fecha de nacimiento de cada estudiante) que corresponde a la edad del estudiante en el momento de entrar por primera vez en los estudios universitarios.
- Municipio familiar *[Mun_f]*: municipio familiar del estudiante durante el primer curso de estudios universitarios.

- Dirección postal familiar *[Dir_f]*: dirección de la familia del estudiante durante el primer curso de universidad.

  A partir de las dos variables anteriores se obtienen las variables socioeconómicas del estudiante.

- Municipio durante el curso *[Mun_c]*: municipio del estudiante durante el curso, que no tiene que coincidir obligatoriamente con el municipio familiar.

- Dirección postal durante el curso *[Dir_c]*: dirección del estudiante durante el curso, siendo esta variables en algunos casos coincidente con la dirección postal familiar.

  A partir de las dos variables anteriores se obtiene la distancia en km y la distancia en tiempo desde este domicilio del estudiante hasta la UPO.

- Tipo de centro *[Tipo_centro]*: tipo de centro donde el estudiante ha cursado sus estudios justamente anteriores al acceso a la UPO. Puede tomar los siguientes valores:

  1. I.E.S. (centros públicos)
  2. C.D.P. (centros concertados o privados)
  3. Otros

- Nota del expediente Bachillerato *[Nota_exp]*: nota del expediente académico del estudiante en bachillerato; esta nota está comprendida entre 5 y 10.

- Nota fase general de la PAU *[Nota_GPAU]*: nota de la fase general de la PAU (Prueba de Acceso a la Universidad); esta nota está calculada como la media obtenida a partir de los exámenes de: Comentario de Texto, Idioma, Filosofía e Historia; el valor de esta nota está comprendido entre 0 y 10.

- Nota acceso *[Nota_Acceso]*: nota general de selectividad (propiamente, media de las calificaciones de los exámenes de la PAU), con valor comprendido entre 4 y 14.

Como se ha comentado, las variables anteriores permiten obtener información sobre las características socio-económicas de cada estudiante. Una vez realizados los cálculos y las modificaciones oportunas a la base de datos de ámbito económico, las variables utilizadas son las que a continuación se describen. Conviene aclarar que los datos corresponden a los municipios de Andalucía (en concreto, a un total de 840 municipios).

- Extensión del municipio *[ExtMun_ año]*: kilómetros cuadrados de la superficie de cada término municipal completo en el año 2010. Corresponde a la última información publicada en la base cartográfica numérica a escala 1:25.000 del IGN.

- Distancia a la capital provincial *[Dist_ año]*: kilómetros de distancia entre cada núcleo principal del municipio y la capital de provincia. La fuente utilizada es el INE.

- Altitud sobre el nivel del mar *[Altitud_ mar_ año]*: metros de altitud sobre el nivel del mar de un punto de la entidad singular principal. Es la información publicada del año 1999 en la base de datos de cartografía del IGN.

- Población *[Pob_ año]*: número de habitantes (a fecha del 1 de enero del año correspondiente) inscritos en el padrón municipal custodiado por el ayuntamiento del municipio. La información obtenida es desde el año 1996 al año 2011, siendo la fuente de obtención de dicha información el INE. La variable Población ha tenido que ser modificada para algunos análisis (con la ayuda de los Ayuntamientos de Sevilla y Dos Hermanas), para obtener información de la población por distritos de la capital de Sevilla así como para obtener información de los distintos núcleos poblacionales de Dos Hermanas; en particular, para obtener información sobre Montequinto, cuyo conocimiento es crucial para este trabajo por proximidad a la UPO y por el volumen de estudiantes que proceden de allí.

- Edad media de la población *[Edad_ Pob_ año]*: promedio en años de la edad del total de la población inscrita en el padrón municipal residente en el municipio correspondiente. La información corresponde al período desde el año 2001 hasta 2010 y ha sido obtenida del INE.

- Población extranjera *[Pob_ Ext_ año]*: número de habitantes extranjeros, a fecha del 1 de enero del año correspondiente, inscritos en el padrón municipal desde 1997 a 2011.

- IBI de naturaleza urbana: nº inmuebles oficina *[IBI_ ofic_ año]*: número de inmuebles de oficina que existen en cada municipio. La fuente de es la Dirección General del Catastro y corresponde al impuesto sobre bienes inmuebles y bienes de naturaleza urbana, del año 1999 hasta el año 2010.

- IRPF: renta neta declarada media *[RentaN_ año]*: la renta neta media se define como el cociente entre la renta neta total declarada y el número de declaraciones. La información que se facilita en esta variable está medida en euros y es la que se obtiene como suma de las rentas

declaradas según el tipo de rendimiento: rentas netas del trabajo, rentas netas de actividades empresariales, rentas netas de actividades profesionales y otros tipos de rentas. La fuente utilizada es la Agencia Tributaria y la información se refiere a los años 1999, 2000, 2002, 2003, 2004 y 2006.

- I.B.I. de naturaleza urbana: base imponible *[NumBase_año]*: valor de la base imponible, atendiendo a que se ha considerado como unidad urbana a todos los inmuebles con una relación de propiedad perfectamente delimitada a efectos fiscales. La variable ha sido elaborada a partir de la información de la Dirección General del Catastro, en concreto de sus estadísticas catastrales. Los años disponibles son desde 1998 hasta 2010.

- Energía *[Energía_año]*: datos procedentes de las facturaciones en megavatios por hora realizadas por la Compañía Sevillana de Electricidad a sus abonados. Se debe tener en cuenta que existen municipios que no poseen suministros, luego allí los datos son estimados. La fuente de la base de datos es la propia Compañía Sevillana de Electricidad, del año 1998 hasta 2010.

3. Metodología aplicada: el análisis cualitativo comparativo difuso

En otros trabajos se han comprobado las dificultades de aplicar métodos tradicionales al problema que se está afrontando, por lo que aquí se utiliza una técnica novedosa (fsQCA) para obtener resultados que, aunque parciales, son razonables, en cuanto a que no son incompatibles con los obtenidos por otros autores siguiendo otras estrategias. Pueden destacarse los resultados obtenidos recientemente mediante la aplicación de técnicas basadas en la inteligencia artificial; dichas técnicas están permitiendo obtener fiabilidad a pesar de tratarse de situaciones y modelos con gran número de variables interrelacionadas (véanse, por ejemplo, [7] y [15]).

Actualmente se reconocen, al menos, cinco formas de inteligencia artificial inspiradas en el funcionamiento del cerebro humano:

1. Ejecución automática de una respuesta predeterminada a cada posible entrada: este tipo sería el más básico y el análogo a los actos reflejos de los seres vivos.

2. Previsión de un conjunto de estados producidos por las acciones posibles y posterior búsqueda del estado efectivamente sucedido.

3. Algoritmos genéticos, que están más bien inspirados en el proceso de evolución de los seres vivos que se produce por la modificación y combinación progresiva de las cadenas de ADN.

4. Redes neuronales artificiales, que imitan el funcionamiento físico del cerebro de animales y humanos, sobre todo en lo que corresponde al aprendizaje y a la respuesta ante situaciones imprevistas de antemano.

5. Razonamiento mediante una lógica formal, que sería lo más análogo al pensamiento abstracto humano, pero la lógica a veces no puede aplicarse en su versión binaria, sino que debe ser "difusa" o "borrosa" (del inglés, *fuzzy*); hay que tener en cuenta que el cerebro humano es capaz de atender a la imprecisión de la realidad y en ocasiones tiene que medir características difíciles de cuantificar (qué es algo lejano, pobre, caro...).

En este último tipo de inteligencia artificial bien pudiera englobarse una técnica novedosa denominada Análisis Cualitativo Comparativo Difuso (fsQCA), que se explica brevemente a continuación por ser la que se aplica posteriormente al conjunto de datos considerado en esta investigación.

Por cuestiones prácticas, la siguiente introducción a la técnica fsQCA es superficial; hay varios artículos recientes que también pueden servir para hacerse una idea de su filosofía y funcionamiento (a este respecto, se pueden consultar, por ejemplo, [13], [17], [18] y [1]).

Pueden señalarse dos aspectos como clave para entender el interés de fsQCA en el ámbito de la Educación, de la Economía, de la Empresa y, en general, de las Ciencias Sociales: por una parte, que permite extraer conclusiones de los casos particulares (desde este punto de vista, se trata del equivalente cuantitativo del método del caso en Empresa), no como las técnicas estadísticas tradicionales, que no están diseñadas para justificar la validez de los estudios sobre muestras reducidas sino para operar bajo el paraguas de las propiedades asintóticas; por otro lado, facilita la incorporación de valoraciones imprecisas (variables subjetivas o de difícil medida exacta, variables en las que no se está muy seguro de los valores reales que toma, etc.), obteniéndose en muchos casos relaciones no simétricas, es decir, que pueden detectarse causas y consecuencias sin que necesariamente se estén produciendo relaciones de equivalencia (sino solo condiciones necesarias o suficientes).

Por otro lado, la diferencia fundamental entre la lógica tradicional (de dos valores de verdad: verdadero y falso) y la difusa se podría comparar con la diferencia entre precisión y relevancia (o significatividad); es decir, a veces no se necesita toda la información o que toda la información

sea precisa o exacta... sino que solo es necesario contar con la información que tiene realmente importancia. Este aspecto lo aporta la Lógica Difusa, en la que determinadas variaciones en los datos o resultados son relevantes mientras que otras variaciones no tienen la menor importancia.

Según [17], fsQCA surge para evitar los numerosos problemas que trae consigo la aplicación indebida de técnicas tradicionales como, por ejemplo, la Regresión Lineal Múltiple. Woodside [17] sostiene que la Regresión Lineal hace que los investigadores piensen de un determinado modo, que no siempre es el más apropiado. Por otro lado, afirma que en las regresiones suele confundirse ajuste (lo que realmente se hace) con predicción (lo que se desearía hacer).

Aparte de la crítica anterior, fsQCA también destaca como técnica interesante para analizar conjuntamente variables de diferentes tipos (aunque se requieran transformaciones) o cuando se necesita incorporar características cuantitativas continuas junto con otras discretas o cualitativas/categóricas.

Al contrario de lo que ocurre en otras técnicas, con fsQCA no es necesario suponer independencia entre las variables explicativas y tampoco supone la existencia de relaciones causa-efecto (pues se considera una lógica asimétrica). Es más, como se trata de un modelo cualitativo (aunque comparado), hay que hablar de "causas del efecto" y no de "efectos de las causas". Esta característica puede parecer una limitación de la técnica, cuando en realidad implica otra forma de entender la realidad, más basada en las relaciones entre las causas que en la influencia entre cada causa y los efectos. De hecho, de acuerdo con [13] y [18], el "efecto neto" no siempre es un concepto útil o válido en la investigación científica.

Otra ventaja de esta técnica es que tampoco es necesario suponer linealidad u otros tipos de relaciones *a priori* entre las variables explicativas y las explicadas. Finalmente, conviene resaltar que fsQCA permite conseguir significatividad con pocas observaciones. Esto es muy importante en Ciencias Sociales, porque es frecuente encontrar trabajos publicados en revistas prestigiosas en los que se extraen conclusiones con unos valores de los estadísticos muy poco significativos (R^2 o p-valor excesivamente bajos) o con muestras poco representativas.

Antes de describir la técnica, es pertinente realizar algunos breves comentarios sobre su origen. Por eso, en un párrafo se resume el fundamento del Análisis Cualitativo Comparativo (QCA). Fue desarrollado por el prestigioso sociólogo Charles C. Ragin. Primero, antes de 1990, desarrolló la técnica denominada csQCA (de las iniciales de *crisp set*, relativa a la Lógica Booleana). Se trataba

de “repensar” tipos de problemas. No se basaba en la idea de correlación sino en buscar relaciones lógicas entre “condiciones causales”. Es decir, se trata a los casos como “configuraciones de causas” y se valora cuáles de dichas configuraciones tienen una influencia en los resultados que se desea analizar.

En cuanto a fsQCA propiamente dicho, lógicamente, tiene unas hipótesis de aplicación (aunque razonables en el caso de esta investigación):

- Los resultados (“outcome”, que es el equivalente a “consecuencia” o “variable dependiente” en otras técnicas utilizadas para hacer inferencia) no suelen serlo de una sola condición (o una sola causa en el lenguaje habitual), sino de una combinación de ellas.
- Diferentes combinaciones de causas pueden proporcionar el mismo resultado final.
- Normalmente, no es posible tener casos de todas y cada una de las combinaciones posibles de causas, pero eso no debe ser impedimento para extraer conclusiones lógicas válidas.
- Las relaciones causales pueden no ser simétricas; es decir, algo puede ser causa sin ser la única y algo puede ser resultado sin ser el único. Desde otro punto de vista, una combinación causal no suele ser suficiente al 100 %.
- Un mismo conjunto de casos no debería utilizarse para explicar diferentes objetivos o diferentes resultados.

Hay determinadas circunstancias que parecen recomendar la aplicación de fsQCA. En cierto modo, fsQCA está a medio camino entre lo cualitativo y lo cuantitativo. Es una técnica que proporciona relaciones de causalidad difusas entre determinadas configuraciones y ciertos resultados. Presta más atención a los casos que a las variables: las variables relevantes se cambian por casos (o *paths*) relevantes. Por todo lo anterior, fsQCA es una técnica apropiada y particularmente potente cuando se estudian sistemas grandes y complejos, donde la interferencia de ocurrencias y de variables es importante. Algunas áreas (no excluyentes) de aplicaciones destacadas y actuales de fsQCA incluyen: descubrir patrones ocultos en los datos cualitativos, habitualmente mediante el uso de ordenadores; estudios de Sociología, donde las variables son usualmente subjetivas; análisis de datos para la toma de decisiones (empresariales...); etc.

En cuanto al sentido de la incorporación de la Lógica Difusa al QCA (o csQCA), es justo reconocer que fsQCA también parte de la Lógica Booleana, pero la mejora o potencia. Así, se

asigna a cada individuo un "índice/grado de pertenencia" al grupo (de modo que lo cualitativo se hace cuantitativo) que verifica las condiciones (en inglés, *recipe*, que se puede traducir por "receta" o "fórmula"). Un error muy común es considerar que este grado de pertenencia es una probabilidad. No se trata de eso, sino de asumir que cada individuo puede participar parcialmente de las características de un grupo (definido por la correspondiente receta).

Tras la aplicación de fsQCA, no siempre es posible encontrar equivalencias en los datos, pero a menudo sí es posible determinar condiciones necesarias o suficientes (o que lo son casi siempre).

En cierto modo, fsQCA se opone a la teoría de indicadores (sobre todo los unidimensionales), pues un indicador es un *output* para ordenar mientras que fsQCA no trata de dar puntuaciones a los individuos en la salida sino en la entrada.

3.1. Proceso de aplicación de fsQCA

El paso previo lógico consiste en determinar el problema y elegir los datos apropiados. Ha de tenerse en cuenta que el conjunto de casos puede variar a lo largo del proceso (normalmente se reduce). Después, conviene comprobar si los datos son "razonables"; hay que eliminar las partes de los datos que puedan ser problemáticas; a veces es pertinente dividir el conjunto de datos; debe comprobarse si el número de datos es adecuado (no muy grande ni muy pequeño...); etc.

A continuación, se convertirían las k variables/características en "difusas". Para dar este paso, que luego se explica bajo el nombre de "calibrado", suele ser necesario determinar el grado de pertenencia de cada caso a cada clase. Tanto este paso como los siguientes, pueden realizarse con la ayuda de un programa específico que puede encontrarse en `http://fsqca.com`.

Con el paso anterior, es posible establecer la *truth table* o "tabla de configuraciones" (sin configuraciones contradictorias), con k términos (cada uno o su complementario, conectados todos por conjunciones lógicas o "y lógicos").

Una vez establecida la tabla de configuraciones, hay que evaluar las 2^k configuraciones (son del tipo "si se da la configuración, entonces se obtiene el resultado"), junto con sus complementarios, estableciendo pertenencias.

No todas las configuraciones son significativas. Deben utilizarse el número de casos (eliminando los de poca frecuencia) y la Lógica Booleana (para prescindir de las cláusulas redundantes) para

reducir las 2^k configuraciones (en números absolutos y en términos que tenga cada condición, si es posible).

Finalmente, el investigador debe seleccionar las reglas con adecuadas "coherencia" y "cobertura" para extraer las conclusiones pertinentes e interpretarlas. Enseguida se comenta qué coherencias y coberturas pueden considerarse razonables en estudios del tipo que se aborda en este artículo.

3.2. Algunos elementos concretos relevantes del fsQCA

A continuación se realizan algunas explicaciones sobre el proceso anterior, a fin de facilitar su comprensión para los lectores que aún no estén familiarizados con él.

Comenzando por el calibrado (o ajuste), es como una normalización de los datos (tanto en las entradas como en las salidas), que se puede hacer en forma "binaria", "de intervalo" o "*fuzzy*". Como ya se ha comentado, consiste en estimar el grado de pertenencia de cada caso al grupo (o *recipe*). En el caso *fuzzy*, que es el que más interesa aquí, este es a menudo el punto más subjetivo del análisis, pues el investigador fija dónde están el 5 %, el 50 % y el 95 % de la distribución de pertenencia (y, en ocasiones, puede que no exista un criterio claro y objetivo para fijar dichos límites). Con el calibrado, las características se convierten en variables; en cierto modo, podría decirse que lo cualitativo se hace cuantitativo, lo discreto se convierte en continuo... Pero debe recordarse que las variables en escala ordinal o de intervalo se convierten en porcentaje de pertenencia, luego las variables se convierten de alguna manera en algo categórico; es decir, que también lo cuantitativo se hace aquí cualitativo.

En el fondo, lo que se hace con la calibración es distinguir entre variación relevante y variación irrelevante (lo que, en opinión de los defensores de fsQCA, hace que el "punto subjetivo" tenga menos importancia que en las regresiones o en los indicadores). Esto enlaza con el comentario que se hacía antes sobre precisión y relevancia.

Conviene comentar también que en un grupo con varias características consideradas, la pertenencia de un individuo a dicho grupo coincide con el menor valor de los de cada una de las variables individuales (sin necesidad de minorarlo).

Sigamos con unos comentarios sobre la coherencia (o *consistency*). Responde hasta qué punto (o grado) es coherente la hipótesis o el enunciado. Dicho con otras palabras, explica hasta qué grado

los casos comparten características del grupo de salida; es decir, el grado en que la pertenencia en la solución es subconjunto de la salida. Se le puede encontrar cierto parecido con la correlación, pero solo en un sentido. Se suele exigir que sea mayor que $0,74$ para extraer conclusiones válidas.

Finalmente, conviene explicar en qué consiste la cobertura (o *coverage*). Explica hasta qué punto cuenta el pertenecer a un grupo (*recipe*) de entrada para la variable dependiente (pertenecer a un grupo de salida). Tiene algo de similitud con el coeficiente de determinación R^2, pero lo que muestra en realidad es cuántos casos sustentan el resultado (esto es, el porcentaje de casos que cubre la solución). Se suele exigir que sea mayor que $0,25$ y menor que $0,65$ (menos no sería suficientemente significativo y más recomendaría el uso de algún tipo de regresión).

Una vez aplicado el análisis fsQCA, se pueden obtener 3 tipos de soluciones. Las más sencillas, simples o simplificadas son las que se suelen denominar "*parsimonious*"; las segundas son las "*intermediate*"; y las últimas se llaman "*complex*". Las últimas son las que contienen más condiciones simples en una misma solución.

4. Resultados

El primer paso en el análisis de la información consiste en evaluar cuáles de las variables disponibles son apropiadas y relevantes. En esa etapa previa se descubre que la relación entre las variables es compleja, que muchas de ellas no son directamente cuantitativas y que no se puede suponer independencia entre las variables explicativas, lo que justifica la utilización del análisis cualitativo comparativo difuso. Tras ejecutar el programa fsQCA.exe[1] sobre los datos anteriormente presentados, se obtienen los siguientes resultados parciales y finales. Primero hay que recodificar las variables, obteniendo las definiciones de pertenencia difusa presentadas en el Cuadro 1.

Una vez hecha la calibración de las variables, se genera la *truth table*, eliminando los casos con frecuencia baja (en este caso, se eliminaron las frecuencias 0 y 1, por lo que `frequency cutoff = 2.000000`, quedando el 80 % de los casos). Tras establecer el "umbral de consistencia" (en este caso, `consistency cutoff = 0.901040`, con lo que es superior a 0,9), se solicita el análisis estándar con todas las variables disponibles. De ahí se obtienen las siguientes dos soluciones (complejas, donde la conjunción lógica se denota por un asterisco y la tilde sirve para señalar la negación de una

[1] El programa está disponible en www.compasss.org y en www.fsqca.com.

Cuadro 1: Variables recodificadas para fsQCA

Variable	Definición	Sup.	Med.	Inf.
Sexo	mujer	1,95	1,5	1,05
km	lejos (en distancia)	0,25	0,03	0,01
sg	lejos (en tiempo)	0,3	0,1	0,01
Edad	mayor	30	19	16
Centro	privado	1,95	1,5	1,05
Bach	buen expediente	1	0,5	0,2
Selec_Gen	buen expediente	1	0,6	0,4
Selectiv	buen expediente	0,6	0,2	0
Nota_Acc	buen expediente	1	0,4	0
Poblac	gran municipio (pobl.)	1	0,05	0
Edad_P	municipio envejecido	0,9	0,8	0,65
Extranj	municipio internacional	1	0,05	0
Extens	municipio extenso	1	0,05	0
Dist_Cap	municipio alejado	1	0,08	0
Altitud	municipio elevado	1	0,3	0
Renta	municipio próspero	1	0,5	0,3
Catastral	municipio caro	0,25	0,05	0
Oficinas	municipio mercantil	1	0,08	0
Energ	municipio industrial	1	0,08	0
Output	alto rendimiento	1	0,5	0

propiedad). Aunque habría que analizarlas con más detalle, en cualquier caso, se puede afirmar que se obtienen combinaciones muy favorables para el rendimiento. La primera solución se escribiría:

```
~km*~sg*~edad*bach*selec_g*selectiv*nota_acc*poblac*extranj*~dist_cap*~altitud*
renta*catastral*oficinas*energ
```

Esta solución viene refrendada por los siguientes parámetros:

1. Raw coverage = $0,321658$
2. Unique Coverage = $0,033912$
3. Consistency = $0,917068$

Observando este resultado, se deduce que un estudiante que vive cerca de la UPO, tanto en distancia como en tiempo, con una edad joven, con buenas notas tanto en su expediente de Bachillerato como en las pruebas de selectividad (tanto en la parte genérica como en la especifica, luego, en consecuencia, con buena nota de acceso a la Universidad) va a tener un buen rendimiento académico en la FCE de la UPO (si cumple otras condiciones adicionales). Realmente, hasta aquí no parece una información muy sorprendente, pero también llaman la atención las siguientes variables socioeconómicas detectadas como favorables para el rendimiento en esta solución: su municipio tiene una población numerosa y un alto número de extranjeros, pero está relativamente cercana a la Capital y tanto el valor castastral como el de la renta del municipio son elevados, presentando también un gasto elevado de energía y numerosas oficinas de índole económico. La técnica garantiza que los estudiantes con todas estas características obtendrán un rendimiento académico elevado.

A continuación se presenta la otra solución obtenida, que tal vez permita seguir entendiendo las condiciones favorables para el rendimiento académico:

```
~km*~sg*~edad*bach*selec_g*selectiv*~poblac*~extranj*~dist_cap*~altitud*renta*
catastral*~oficinas*~energ
```

Esta solución viene caracterizada por los siguientes parámetros:

1. Raw coverage = $0,260870$

2. Unique Coverage = 0, 043938

3. Consistency = 0, 909926

Según esta segunda solución, un estudiante con las siguientes características obtendrá también un buen rendimiento académico al finalizar sus estudios (al menos en lo concerniente a las asignaturas de índole cuantitativa): el domicilio del estudiante está cercano a la UPO tanto en tiempo como en distancia, el estudiante es joven y tiene un buen rendimiento en Bachillerato y en Selectividad (tanto en la parte general como en la específica). En cuanto a las características socioeconómicas encontradas como relevantes, este estudiante vive en un municipio con muy poca población (tanto española como extranjera), que es cercano a la Capital, con una altitud cercana al nivel del mar y con una renta y un valor catastral del municipio muy elevados. El municipio gasta poca energía y posee pocas oficinas de carácter económico.

Según se puede comprobar, hay muchos casos que quedan fuera de estas soluciones, pero en ninguno de los dos casos se contradice el resto de análisis realizados por otros autores y ambas hacen pensar que son muchas las variables que afectan al rendimiento académico y que pueden servir para determinar perfiles educativos que implican éxito. En el caso de los dos conjuntos de características (condiciones) obtenidas, se comprueba que el éxito académico es posible tanto en municipios grandes como en pequeños, pero ambas soluciones se refieren a estudiantes con buen rendimiento académico previo y, lo que destaca aquí, en municipios con alto nivel económico (al menos, según la renta per cápita y el valor catastral).

Hay otras soluciones que no cumplen los criterios exigidos por los investigadores, por lo que se podrían ignorar. No obstante, la siguiente solución más cercana a aparecer (que queda casi en el límite de las que se han eliminado por ser escasamente significativas) es la siguiente:

```
~sexo*~km*~sg*~edad*bach*selec_g*selectiv*nota_acc*poblac*extranj*~dist_cap*
~altitud*renta*catastral*energ
```

Los parámetros asociados a esta solución son:

1. Raw coverage = 0, 212633

2. Unique Coverage $= 0,001411$

3. Consistency $= 0,904570$

Como se puede comprobar, aquí también aparecen individuos con buen rendimiento previo y que habitan en municipios con alto poder adquisitivo (alta renta y alto valor catastral).

5. Conclusiones

Cada día surgen análisis más rigurosos del rendimiento académico de los universitarios. Algunos de ellos utilizan las técnicas más innovadoras. Así, en [7] se utilizan redes neuronales artificiales para predecir el rendimiento académico y, más recientemente, [15] trata de predecir el abandono de los estudios con una técnica también basada en la inteligencia artificial.

En cuanto a la aplicación presentada, proporciona pistas para entender el fenómeno que se propone analizar: la relación estrecha entre Economía y Educación. La primera conclusión que debe destacarse de los resultados es que el problema es muy complejo, sobre todo, por la inmensa variabilidad en las relaciones entre las variables implicadas. También es considerable el número de variables en sí mismo, la distinta naturaleza de las características que hay que tener en cuenta, la dificultad de encontrar un conjunto de datos adecuado y fiable, etc.

Las variables referentes a género, a estudios previos, a pruebas de acceso o al nivel socioeconómico de los estudiantes son algunas de las que han sido utilizadas para tratar de explicar el rendimiento académico y las relaciones entre las variables que lo favorecen. Se ha demostrado que dichas variables tienen parcial influencia en el rendimiento académico y que la relación entre ellas es consistente en algunos casos. Como es lógico, no todos los estudiantes responden con el mismo rendimiento académico a las mismas características personales. También se detecta una fuerte dependencia del tipo de asignatura; por ello, sería necesario estudiar las distintas habilidades y destrezas que se adquieren con cada asignatura de ámbito cuantitativo y compararlas con los resultados del análisis, para tratar de relacionar los resultados alcanzados según el nivel real de una característica en particular.

Mediante el uso de una técnica novedosa (fsQCA), se ha estimado un modelo que explica adecuadamente cómo algunas condiciones (o causas) afectan a la variabilidad de un indicador de ren-

dimiento académico previamente definido. Llevar esto a la práctica permitiría incluso realizar recomendaciones de índole académica a los estudiantes que comienzan, a partir de datos previos y características objetivas. Consideramos que herramientas de este tipo (con los matices necesarios en cada caso) serían muy interesantes para las universidades que deseen poder orientar a los estudiantes que ingresan por primera vez en la institución.

Como línea futura de investigación, se propone realizar nuevos análisis con un conjunto de datos que incorpore otras facultades y titulaciones, tratando de comparar si las soluciones (o combinaciones de factores relevantes) propuestas por la técnica son apropiadas en otros contextos académicos, para poder trasladar estos resultados a otros estudios.

Referencias

[1] Aguilera, J.; Fedriani, E. M.; Delgado, B. (2014): "Institutional distance among country influences and environmental performance standardization in multinational enterprises". *Journal of Business Research* 67(11), pp. 2385–2392.

[2] Alba, R.; Segundo, S. M. (1995): "The return to education in Spain". *Economics of Education* 14(2), pp. 155–166.

[3] Barceinas, F.; Alonso, J. L.; Raymond, J. L.; Roig, J. L. (2000): "Los rendimientos de la educación en España". *Papeles de Economía Española* 86, pp. 128–148.

[4] Becker, W. E.; Walstad, W. B. (1987): *Econometric modelling in economic education research*. Boston: Kluwer Nijhoff Publishing.

[5] De la Fuente, Á.; Jimeno, J. F. (2012): *La rentabilidad privada y fiscal de la educación en España*. Madrid: Observatorio sobre Capital Humano en España, BBVA Research.

[6] Dolado, J. J.; Morales, E. (2007): "Which factors determine academic performance of undergraduate students in economics?". *CEPR Discussion Papers* 6237, 22 pp.

[7] Fedriani, E. M.; Hidalgo, M. A.; Romano, I. (2017): "The prediction of academic success of university students to optimize their performance", por aparecer.

[8] IECA (2013): *Sistema de Información Multiterritorial de Andalucía (SIMA)*. Sevilla: Instituto de Estadística y Cartografía de Andalucía.

[9] INE (2009): *Anuario Estadístico de España 2009*. Madrid: Instituto Nacional de Estadística.

[10] INE (2010): *Anuario Estadístico de España 2010*. Madrid: Instituto Nacional de Estadística.

[11] INE (2011): *Anuario Estadístico de España 2011*. Madrid: Instituto Nacional de Estadística.

[12] Navarro, M. L.; Marcenaro, O. D. (2003): "Condiciones de acceso y otras características del estudiante como determinantes del éxito en el primer curso universitario". En San Segundo, M. J.; Zorrilla, R. (eds.): *Economía de la Educación: AEDE XII*. Madrid: Asociación de Economía de la Educación, pp. 42–61.

[13] Ragin, C. (2008): *Redesigning social inquiry: Fuzzy sets and beyond*. Chicago: Chicago University Press.

[14] Raymond, J. (2002): "Convergencia real de las regiones españolas y capital humano". *Papeles de Economía Española* 93, pp. 109–121.

[15] Rovira, S.; Puertas, E.; Igual, L. (2017): "Data-driven system to predict academic grades and dropout". *PLoS ONE* 12(2), pp. 1–21.

[16] Tejedor, F. J.; García, A. (2007): "Causas del bajo rendimiento del estudiante universitario: propuesta de mejora en el marco del EEES". *Revista de Educación* 342, pp. 443–473.

[17] Woodside, A. G. (2013): "Moving beyond multiple regression analysis to algorithms: Calling for adoption of a paradigm shift from symmetric to asymmetric thinking in data analysis and crafting theory". *Journal of Business Research* 66(4), pp. 463–472.

[18] Woodside, A. G.; Zhang, M. (2013): "Cultural diversity and marketing transactions: Are market integration, large community size, and world religions necessary for fairness in ephemeral exchanges?". *Psychology and Marketing* 30(3), pp. 263–276.