Este libro está especialmente dirigido a estudiantes o profesionales de la investigación social que han seguido algún curso introductorio de estadística y que desean comenzar su aprendizaje de las técnicas multivariables. Todos los capítulos tienen un doble nivel de lectura. Siguiendo los ejemplos ilustrativos el lector podrá comprender la utilidad de la técnica en cuestión. Aun cuando haya alguna parte de cada capítulo que por su complejidad matemática resulte difícil de seguir, a través de los ejemplos se puede llegar a comprender básicamente cuál es el objetivo de cada técnica de análisis, cómo procede, cuáles son los resultados y cómo se interpretan. Ello, unido a la amplia divulgación que existe hoy en día de los programas estándar de análisis de datos en ordenador, permitirá hacer uso de cada técnica con el fin de dar solución a los problemas de investigación. Para un lector matemáticamente más exigente, cada capítulo incluye información suficiente para seguir el algoritmo básico de cálculo que se utiliza en la técnica.

Los autores:

Joan Manuel BATISTA FOGUET, profesor del departamento de Técnicas Cuantitativas de Gestión de la Escuela Técnica Superior de Ingenieros Industriales de Barcelona.

A. P. M. COXON, profesor de Métodos de Investigación en Sociología en el University College de Cardiff (Universidad de Gales).

José Miguel GARCIA SANTESMASES, profesor del Departamento de Estadística e Investigación Operativa de la Universidad Complutense de Madrid.

Charles JONES, profesor de Sociología en la Universidad de Toronto.

Emilio MARTINEZ RAMOS, profesor de la Facultad de Ciencias de la Información y Director General de Emopública.

Juan Javier SANCHEZ CARRION, profesor del Departamento de Métodos y Técnicas de Investigación Social (Universidad Complutense de Madrid: Facultad de CC.PP. y Sociología).

William E. SARIS, profesor de Métodos y Técnicas de Investigación en Ciencia Política en la Universidad de Amsterdam.

Albert SATORRA BRUCART, profesor del Departamento de Estadística y Econometría de la Universidad de Barcelona.

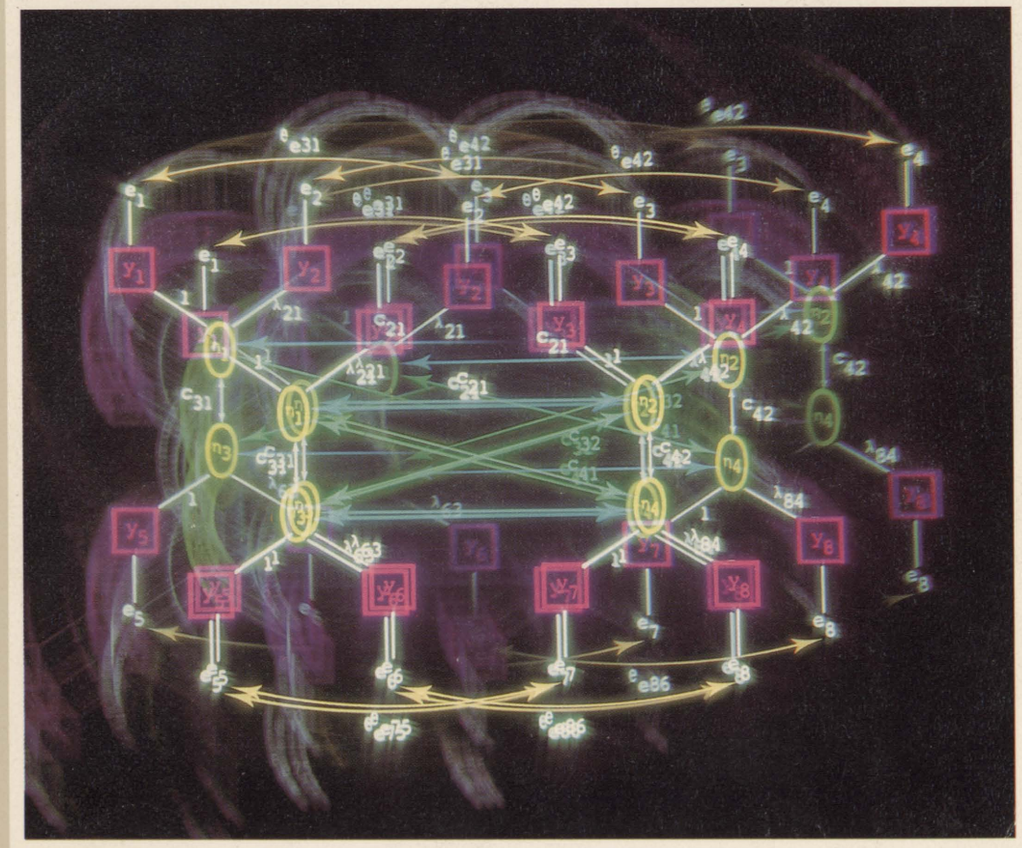
L. Henk STRONKHORST, trabaja en el Netherland Central Bureau of Statistics, Amsterdam.

Edit. Juan Javier Sánchez Carrión

introducción a las técnicas de análisis multivariable aplicadas a las ciencias sociales

introducción a las técnicas de análisis multivariable aplicadas a las ciencias sociales

Edición a cargo de: Juan Javier Sánchez Carrión



Centro de Investigaciones Sociológicas

INTRODUCCION A LAS TECNICAS DE ANALISIS MULTIVARIABLE APLICADAS A LAS CIENCIAS SOCIALES

INTRODUCCION A LAS TECNICAS DE ANALISIS MULTIVARIABLE APLICADAS A LAS CIENCIAS SOCIALES

Editor: Juan Javier Sánchez Carrión

CENTRO DE INVESTIGACIONES SOCIOLOGICAS
MADRID-1984

I.S.B.N.: 84-7476-124-9
Depósito legal: M. 40.717-1984
Impreso en España por:
Rumagraf, S. A. - Nicolás Morales, 34
28019 - Madrid

Indice

Introducción

El libro que ahora introducimos pretende cubrir una laguna que existe en el campo de las ciencias sociales. Si bien resulta relativamente sencillo de encontrar un libro que cubra los conocimientos introductorios de estadística, ya no es tan fácil hallar un manual que recoga las técnicas multivariables más usuales del análisis de los datos y que esté escrito específicamente para personas que sin una fuerte base matemática se introducen en el tema desde su interés por la investigación social.

Este libro está especialmente dirigido a estudiantes o profesionales de la investigación social que han seguido algún curso introductorio de estadística y que desean comenzar su aprendizaje de las técnicas multivariables. Hemos pretendido que todos los capítulos tengan un doble nivel de lectura. Siguiendo el ejemplo ilustrativo que se incluye en cada capítulo el lector no técnico podrá comprender la utilidad de la técnica en cuestión. Es decir, aun cuando haya alguna parte de cada capítulo que por su complejidad matemática resulte difícil de seguir, a través de los ejemplos se puede llegar a comprender básicamente cuál es el objetivo de cada técnica de análisis, cómo procede, cuáles son los resultados y cómo se interpretan. Ello, unido a la amplia divulgación que existe hoy en día de los programas estándar de análisis de datos en ordenador, permitirá hacer uso de cada técnica con el fin de dar solución a los problemas de investigación. Para un lector matemáticamente más exigente, cada capítulo incluye información suficiente para seguir el algoritmo básico de cálculo que se utiliza en la técnica.

El libro está dividido en 3 Secciones y 9 Capítulos. La clasificación que se ha hecho sigue un criterio puramente práctico, en función de la aplicación de las técnicas. Aun cuando técnicas de diferentes secciones tengan fundamentos estadísticos semejantes, por ejemplo los modelos lineales, el análisis factorial y el análisis discriminante son todas derivaciones del modelo general lineal, sus objetivos de investigación pueden ser diferentes. Básicamente se pueden distinguir tres operaciones: reducir los datos, clasificarlos y explicarlos.

Cuando se recoge mucha información sobre los individuos (u otras unidades de análisis), parte de esta información es redundante. Las técnicas de *reducción* pretenden eliminar esa información redundante y quedarse con lo esencial. Así, el gusto por el orden como forma de resolver los problemas sociales, las creencias en *una* verdad frente a opiniones discrepantes o en la falta de autoridad como causa de los problemas de la juventud se pueden interpretar por el carácter autoritario de las personas que

sustentan estas opiniones. En otras situaciones de investigación, cuando los individuos contestan a varias preguntas se puede pensar que no todos son diferentes entre sí, sino que en función de sus respuestas cabe *clasificarlos* en diferentes grupos más o menos homogéneos. Por último, cuando se disponga de una teoría que lo justifique cabe superar el carácter generalmente descriptivo de las dos operaciones previas, para tratar de explicar en base a una serie de variables independientes y sus interrelaciones la variabilidad que se observa en una variable dependiente (por qué no todo el mundo gana lo mismo, o piensa igual, etc.); para ello habrá que *ajustar algún modelo* a los datos observados.

Dado que los datos que se producen en las ciencias sociales son diversos (nominales, ordinales o intervales), en las tres secciones incluimos técnicas que se ajustan al variado nivel de información del que se dispone. Por ejemplo, en la Sección I (Reducción de los datos) incluimos el Análisis Factorial, que requiere de información interval; el Análisis Factorial de Correspondencias, que parte de las tablas de contingencia; y las Escalas Multidimensionales, que aunque producen resultados métricos tan sólo asumen un nivel ordinal de los datos.

Con el fin de situar y dar unidad a cada uno de los capítulos del libro, al comienzo de cada sección hemos escrito unas páginas donde se pasa revista a cada técnica ofreciendo un breve resumen del contenido de los capítulos.

Digamos como sencilla y breve contextualización epistemológica del libro que no abogamos por unas ciencias sociales que vayan a encontrar su estatus científico en el uso indiscriminado de un repertorio de técnicas de investigación desprovisto de todo conocimiento sustantivo (teórico) del tema que sea objeto de estudio, pero rechazamos la concepción de esas disciplinas como meras formulaciones teoréticas que jamás descienden a ser contrastadas en la práctica de la investigación.

Por último sólo nos queda agradecer su colaboración a cada uno de los autores que han hecho posible este libro y al CIS el interés mostrado al publicarlo.

Autores:

Joan Manuel BATISTA FOGUET es doctor Ingeniero Industrial (Universidad Politécnica de Barcelona), licenciado en Psicología (Universidad Central de Barcelona) y diploma en Social Science Data Analysis and Collection (Universidad de Essex). Es profesor del departamento de Técnicas Cuantitativas de Gestión de la Escuela Técnica Superior de Ingenieros Industriales de Barcelona.

Antony P. M. COXON es profesor de Métodos de Investigación en Sociología en el University College de Cardiff (Universidad de Gales). Estudió Sociología y Filosofía en la Universidad de Leeds y ha dado clases en esta Universidad, el Massachusetts Institute of Technology y la Universidad de Edimburgo. En la Escuela de Verano de la Universidad de Essex ha impartido cursos de Escalas Multidimensionales durante varios años. En 1983 fue Distinguished Visiting Hooker Professor en la Universidad de McMaster (Canadá). Tony Coxon está interesado, *inter alia,* en la Sociología de las Profesiones, los problemas de medida en Sociología y la Sociología de la Religión.

José Miguel GARCIA SANTESMASES es doctor en Ciencias Matemáticas (Univer-

sidad Complutense de Madrid) y profesor del Departamento de Estadística e Investigación Operativa de esta misma Universidad. Está interesado en temas de Estadística Computacional.

Charles JONES es profesor de Sociología en la Universidad de Toronto. Estudió Ciencias Naturales y Antropología Social en la Universidad de Cambridge y Psicología Social en la London School of Economics. Ha enseñado en la Universidad de Edimburgo, el Massachusetts Institute of Technology y la Universidad de McMaster. Está interesado en temas de mobilidad social, el uso de la estadística en Sociología y la Sociología de la Educación. Da clases de Escalas Multidimensionales en la Universidad de Essex.

Emilio MARTINEZ RAMOS es doctor en Ciencias Económicas (Universidad de Madrid) y profesor de la Facultad de Ciencias de la Información. Es Director General de Emopública, Instituto de Investigación de Mercados y de Opinión Pública.

Juan Javier SANCHEZ CARRION es doctor en Sociología (Universidad Complutense de Madrid), Técnico Publicitario y diploma en Social Science Data Analysis and Collection (Universidad de Essex). Profesor del Departamento de Métodos y Técnicas de Investigación Social (Universidad Complutense de Madrid: Facultad de CC.PP. y Sociología). Está interesado en las Técnicas de Investigación Social y en el estudio del contenido de los medios de comunicación de masas.

William E. SARIS es profesor de Métodos y Técnicas de Investigación en Ciencia Política en la Universidad de Amsterdam. Está especializado en el campo de la Teoría de la Decisión y en los Modelos de Ecuaciones Estructurales.

Albert SATORRA BRUCART es profesor del Departamento de Estadística y Econometría de la Universidad de Barcelona. Es doctor en matemáticas por la Universidad de Barcelona y ha realizado estudios e impartido enseñanzas en la Universidad de Essex dentro de los programas de la Summer School of Social Science Data Analysis. Sus trabajos de investigación actual versan sobre problemas relativos al contraste estadístico en modelos de ecuaciones estructurales.

L. Henk STRONKHORST recibió su doctorado en Sociología de las Areas en Desarrollo en la Universidad Libre de Amsterdam. Obtuvo su Ph. D. en la Universidad de Arizona en Tucson con una tesis denominada «Conflicto colectivo en América Latina, 1946-1975». Es coautor con W. E. Saris del libro *Introducción a los modelos causales en la investigación no-experimental: el enfoque LISREL.* Actualmente trabaja en el Netherland Central Bureau of Statistics, donde analiza varios bancos de datos y trata de mejorar la calidad de las estadísticas sociales.

SECCION I

LA REDUCCION DE LOS DATOS

1. Introducción

En esta sección vamos a presentar el Análisis Factorial, el Análisis Factorial de Correspondencias y las Escalas Multidimensionales. Las tres técnicas tienen en común el hecho de que intentan representar los estímulos (coches, naciones, etc.) o las variables (religión, actitudes políticas, etc.) objeto de estudio en un espacio r-dimensional, donde r es menor que el número de estímulos o de variables. Esta representación tiene una doble lectura. Por un lado se pueden estudiar los estímulos o variables que aparecen juntos, definiendo grupos homogéneos; por otro se pueden dejar de lado los estímulos o variables prestando atención a (interpretando) las dimensiones obtenidas en el análisis. Veamos cada uno de estos aspectos considerando las técnicas en cuestión.

Bajo el título de *análisis factorial* se incluyen modelos diferentes como son el análisis de componentes principales, el análisis factorial exploratorio y el análisis factorial confirmatorio. En los tres casos el punto de partida es la matriz de correlaciones o de varianzas-covarianzas de las diferentes variables incluidas en el análisis —por lo tanto el nivel de medida que se requiere es de tipo interval. El objetivo de los tres tipos de análisis es obtener e interpretar un conjunto reducido, r, de componentes o factores latentes (no observados empíricamente) que expliquen la covariación existente entre las p variables originales, siendo $r < p$.

Si se observa que los niños que tienen buenas notas en matemáticas —ejemplo que utiliza Batista— suelen también tener buenas notas en ciencias naturales (las dos variables covarían o están correlacionadas), y que aquéllos con buenas notas en francés también son alumnos brillantes en latín, se puede pensar que existe una capacidad lógico formal y otra habilidad verbal que son las que explican que aquel niño que posea la primera saque buenas notas en las dos primeras materias y el que posea la segunda lo haga en las dos últimas. Matemáticas, ciencias naturales, francés y latín serían la p variables observadas y las habilidades lógico-formal y verbal los polos de uno de los componentes o factores que explica las relaciones que se observan entre las notas.

Observando la representación gráfica de las 4 variables del ejemplo anterior en un espacio bidimensional, donde los ejes son los componentes o factores (apartado 2.5) se puede ver que las 4 variables aparecen en dos grupos: por un lado, matemáticas y ciencias naturales, y, por otro, francés y latín. Una posible lectura consistiría en estudiar los grupos formados en el análisis. Alternativamente, o complementariamente, se puede estudiar el significado de los ejes del espacio representacional. En este caso el

autor de este sencillo ejemplo, Reuchlin, etiqueta uno de los ejes como «inteligencia general» y el otro, tal como hemos indicado, «habilidad lógico-formal frente a habilidad verbal».

Una vez que se han obtenido los dos componentes o factores podemos considerarlos como dos nuevas variables que sustituyen a las 4 originales. En función de los coeficientes que proporcionan la combinación lineal de las variables originales en la determinación de las componentes o de los factores *(factor score coefficients)* y de los valores de las variables se puede determinar una nueva puntuación factorial de los individuos en cada uno de los nuevos factores, sustituyendo así las 4 variables originales por los valores en estos dos nuevos componentes.

Matemáticamente el análisis de componentes y el análisis factorial difieren en la forma de obtener los ejes. En el primer caso se trata simplemente de hacer una transformación lineal ortogonal de las variables originales, definida por los vectores propios de la matriz de varianza-covarianza. Así, las variables originales, $(x_1 \ldots x_p)$, se transforman linealmente en un nuevo conjunto de variables compuestas y estandarizadas, $(y_1 \ldots y_p)$, no correlacionadas entre sí y cuya varianza decrece a partir de la primera componente. Se trata simplemente de seleccionar las primeras componentes que expliquen una cantidad «suficiente» de la varianza de las variables originales.

Cuando se trata del análisis factorial el modelo hace el supuesto de que cada variable está compuesta de una parte común con las otras variables y otra específica, en la que se incluye tanto su propia especificidad como el posible error de medida. La parte común puede ser explicada por una serie de factores comunes, que son los que hay que calcular. Puesto que cada variable sólo tiene una parte en común con el resto, la diagonal de la matriz de varianza-covarianza que sirve de base para la extracción de los factores debe de reflejar este hecho. Ello se consigue sustituyendo los 1 (unos) de la matriz utilizada en las componentes principales por un valor estimado de esa parte común, normalmente el coeficiente de correlación múltiple de cada variable con las restantes. En la práctica ésta es la diferencia.

Batista muestra el procedimiento analítico de cálculo de las componentes principales y de los factores, incluyendo las sentencias del paquete SPSS que permiten su cálculo en el ordenador y las correspondientes salidas con los resultados. Además del cálculo de los factores y de las componentes el autor trata el tema de la rotación de los ejes con el fin de facilitar la interpretación de los resultados. Igualmente se sigue un ejemplo común para ambos métodos con el fin de poder comparar los resultados.

El tercer método que se explica en el capítulo permite introducir una problemática nueva a la que hemos mostrado hasta ahora (cálculo e interpretación de los ejes). Tanto en el análisis de las componentes como en el análisis factorial se parte de un desconocimiento teórico del objeto de estudio y se selecciona una muestra de variables tratando de descubrir los componentes o los factores necesarios para explicar sus interrelaciones. Este uso del método se dice que es exploratorio. Ahora bien, cuando se conoce el proceso que genera la covariación es posible formular hipótesis acerca de la estructura causal que existe entre las variables observadas y los factores no observados que las explican. Para ello hay que construir un modelo causal (véase sección III) en el que se especifican las relaciones entre las variables, restringiendo los valores de algunos parámetros. A condición de que el modelo causal esté identificado es posible estimar sus parámetros (los antiguos pesos de los factores en las variables). El método de esti-

mación utilizado es el de máxima-verosimilitud. Una de las ventajas de este método es que permite contrastar la idoneidad del modelo propuesto, cosa que no era posible con los modelos precedentes de estimación mínimo cuadrática. El modelo factorial confirmatorio es un caso particular del modelo de medida en los sistemas de ecuaciones estructurales explicado por Saris en este mismo libro. Batista sigue la notación utilizada por Saris y Satorra en sus trabajos (véase sección III), que no es otra que la notación de LISREL.

El segundo de los trabajos que incluimos en esta sección corresponde al Análisis Factorial de Correspondencias, desarrollado por García Santesmases. Desde un punto de vista práctico, la principal diferencia de esta técnica con el análisis factorial clásico se basa en la naturaleza cualitativa de las variables que ahora se utilizan. Mientras que un requisito del análisis factorial era el uso de variables medidas a un nivel interval, en el análisis de correspondencias las variables son de tipo categórico (nominales u ordinales).

Tomando el ejemplo de las correspondencias múltiples podemos mostrar el objetivo de esta técnica. A partir de las tablas de contingencia entre las variables *X, Y, Z,* por ejemplo (tabla de Burt), se trata de ver el tipo de dependencias que se establecen entre ellas. Buscaremos cuáles son las categorías de *X* y de *Z,* por ejemplo, que más influyen en (o que se asocian con) una determinada categoría de *Y,* pudiendo determinar así el estereotipo de los individuos (objetos) que se clasifican en esa categoría de *Y.* Así, en el ejemplo que utiliza Santesmases sobre los emigrantes iberoamericanos en España, en base a las variables utilizadas en el estudio se muestra cómo los «cubanos» quedan descritos en términos de «refugiados», siendo «estudiantes» o estando «desempleados», habitando preferentemente en «Madrid» y estando muy pocos de «turismo». Igual se podría hacer con el estereotipo de otra nacionalidad y por comparación entre sus respectivos estereotipos (lo que unos tienen y de lo que los otros carecen) ver las categorías que les discriminan —este sería un uso del análisis de correspondencias parecido al análisis discriminante.

Como resultado del análisis de correspondencias también se obtienen representaciones gráficas que permiten visualizar los resultados a los que acabamos de hacer referencia, de forma que la alteración producida por estas representaciones en los datos de partida sea mínima. Si se miden las distancias que hay entre las categorías en los datos de partida utilizando la distancia de Benzecri[1], el objetivo del análisis es representar las categorías de todas las variables en un espacio del menor número posible de dimensiones respetando las distancias originales. Las deformaciones producidas para cada categoría en la solución factorial se pueden evaluar mediante la observación de las contribuciones relativas de las categorías a los factores. Al igual que en el análisis factorial clásico se puede estar interesado en la interpretación de los ejes factoriales. También en este caso hay que recurrir a las categorías que mejor definen cada uno de los ejes, viendo las contribuciones absolutas de las categorías (el equivalente a las saturaciones del análisis factorial).

En aquellos casos en los que se haga imposible la reducción de la dimensionalidad del problema, debido a que el número de factores necesarios para explicar un tanto

[1] Tal como explica Santesmases ésta es una distancia basada en la Gi-cuadrado que se elige por tener algunas propiedades que la hacen conveniente.

por ciento elevado de la inercia sea muy grande, se pueden utilizar los factores en sustitución de las variables cualitativas, considerando así nuestro problema descrito en términos de variables cuantitativas.

En términos prácticos vemos que no hay diferencia entre las dos técnicas (análisis factorial y análisis de correspondencias), si exceptuamos el punto que hace verdaderamente interesante al análisis de correspondencias: su capacidad para trabajar con variables cualitativas (nominales u ordinales), sin ningún tipo de transformación. Donde sí hay diferencias es en la elección de la métrica utilizada. En las correspondencias las categorías vienen representadas por sus perfiles en un espacio de tantas dimensiones como categorías tiene la otra variable (en el supuesto de correspondencias simples). A esos puntos, dotados con la métrica de Benzecri, se les somete a la misma metodología que en el análisis de componentes principales: localización de subespacios de máxima inercia. En las componentes principales o en el análisis factorial los puntos vienen representados por sus valores en un espacio de tantas dimensiones como variables existan, siendo la métrica utilizada la euclídea. El objeto del análisis es la localización de subespacios de máxima varianza mediante la diagonalización de la matriz de correlaciones.

En el tercer capítulo de esta sección Coxon y Jones introducen las escalas multidimensionales no-métricas. En esta técnica se asume que el nivel de medida es ordinal, aun cuando en la solución del análisis (por solución se entiende la distancia entre los estímulos en un espacio *r*-dimensional) se recupera la información métrica subyacente a los mismos.

Dado un número de estímulos (coches, países, etc.) que difieren respecto de una serie de propiedades o dimensiones (velocidad, precio, etc. en el caso de los coches; renta, sistema político, etc. en el caso de las naciones), el modelo *básico* de las escalas multidimensionales trata de ver cuál es el número mínimo de esas dimensiones que permiten explicar la variabilidad de los estímulos (según las respuestas de los entrevistados los coches o los países son diferentes) y cuáles sus coordenadas en esas mismas dimensiones. En el espacio multidimensional los estímulos están representados por puntos, correspondiendo su posición al grado o cantidad de «propiedad(es)» o «dimensión(es)» que posean; mientras que la distancia entre dos estímulos (entre dos puntos en el espacio) está en función de su grado de (di)similaridad: cuanto más semejantes sean, más próximos estarán los dos estímulos en el espacio.

En el modelo básico del análisis multidimensional la variabilidad que se observa en las respuestas de los entrevistados a la hora de evaluar los estímulos[2] se adscribe a la propia variabilidad que en éstos mismos existe. Sólo se miden los estímulos, tratando de lograr que sea en el menor número posible de dimensiones. Se desprecia la singularidad de las respuestas de cada entrevistado y se calcula una media para el conjunto de las respuestas.

En los análisis de las *diferencias individuales* y de las *preferencias* (ambos se explican en el trabajo de Coxon y Jones) las diferencias que hay en las respuestas de los

[2] Una pregunta que diera lugar a información susceptible de ser analizada mediante las escalas multidimensionales podría ser del tipo siguiente: formar parejas con una lista de países; ordenándolas por orden de semejanza; o puntuar de 0 a 100 las parejas de países, reservando el 0 para aquella pareja totalmente semejante y el 100 para aquella otra que sea totalmente diferente. (Véase una explicación e ilustración del modelo básico en Sánchez Carrión, 1984).

entrevistados también se tienen en cuenta, considerando que no sólo hay variabilidad entre los estímulos sino también en su percepción por parte de los sujetos, por lo que ambos (estímulos y sujetos) son objeto de medición. En el primer caso (diferencias individuales) para ver cómo perciben los mismos estímulos diferentes individuos. En el segundo (análisis de las preferencias) se estudian cuáles son las preferencias, frente a un grupo de estímulos, de cada uno de los entrevistados.

Respecto del análisis factorial se pueden señalar las siguientes diferencias. Por una parte, ambas técnicas difieren en el tipo de datos que utilizan: interval en el análisis factorial y ordinal en el análisis multidimensional no-métrico. Sin embargo, dado que también hay análisis factorial no-métrico y escalas métricas, es más importante la diferencia que se establece entre ambas técnicas en términos de los modelos subyacentes respectivos. El análisis factorial está basado en un modelo de productos escalares mientras que las escalas multidimensionales descansan en un modelo de distancias. Puesto que el análisis factorial trata los datos como productos escalares, una solución perfecta será aquella en la que los productos escalares entre los puntos en el espacio multidimensional se correspondan con los valores de los datos originales. O dicho de otra manera, siguiendo la explicación de Batista, cuando las correlaciones entre los puntos reproducidas a partir de las ponderaciones factoriales sean iguales a las correlaciones originales. En las escalas multidimensionales no métricas la solución correcta será aquella en la que las distancias entre los puntos se iguale con (sea una transformación monótona de) los valores de los datos de partida (las [di]similaridades entre los estímulos) (véase Rabinowitz, 1975).

2. Componentes principales y análisis factorial (exploratorio y confirmatorio)

por *Joan Manuel Batista Foguet*

Cuando el objeto de estudio es el hombre, la investigación científica en Psicología, Sociología, Biología, etc., requiere instrumentos adecuados para tratar la enorme y diversa cantidad de información generalmente disponible. Los métodos estadísticos multivariables implementados en ordenador, permiten mediante cálculo matricial conjugar todos los aspectos del estudio —diseño, estimación y contrastación de relaciones, y evaluación del término de error— en un único análisis. En la presentación de la técnica de Análisis de Componentes Principales y del modelo de Análisis Factorial, se ha puesto el énfasis tanto en la necesidad de matrices como en la del ordenador.

En primer lugar, un estudio para definir una tipología socioeconómica de municipios en Cataluña introduce a las técnicas de Análisis Multivariable, cuyo objetivo es reducir, sin parcializar, la información proporcionada por el conjunto inicial de variables. Los conceptos que aparecen en este apartado se establecen a través de nociones de Estadística básica, para formalizarlos a continuación.

Un ejemplo que se resuelve «manualmente», permite utilizar la técnica de Análisis de las Componentes Principales para introducir conceptos propios del modelo de Análisis Factorial, e identificar posteriormente los resultados en el listado que proporciona un programa estadístico ad-hoc.

Desde una perspectiva algebraica se describe el concepto de estructura de una matriz de correlaciones, derivada del análisis previo que se efectúa sobre esta matriz: a continuación se obtienen las componentes a partir de las variables originales, y se apuntan algunas de las relaciones que se precisarán ulteriormente.

El tercer apartado presenta el modelo de Análisis Factorial, explicitando sus diferencias con la técnica de análisis anterior, desde los puntos de vista: histórico, formal y práctico.

En primer lugar, se definen la ecuación y los supuestos que conjuntamente determinan el modelo, así como la estructura que éste implica *a priori* en la matriz de correlaciones. Más adelante, las mismas correlaciones entre un grupo de variables, cuya subyacente estructura común es conocida si bien su solapamiento es reducido, se someten a dos análisis paralelos: Factorial y de Componentes Principales, con la finalidad de observar las distintas repercusiones que deben aparecer en los resultados. Por último, ya que en la valoración de estos análisis han prevalecido tradicionalmente criterios propios de la técnica de Componentes Principales, se sugiere aquilatar, como

adecuado complemento de aquellos criterios, los residuos que permanecen al reproducir la matriz de correlaciones según el modelo Factorial estimado.

Por fin, en el apartado cuarto, se introduce el Análisis Factorial Confirmatorio a partir de la referencia exploratoria que del modelo se ha dado en capítulos anteriores.

Se establece primero el papel que el modelado estadístico multivariable desempeña en la investigación científica, y se demarca su utilización exploratoria o confirmatoria. En segundo lugar, la introducción de restricciones en el modelo de Análisis Factorial permite, por un lado, resolver con mayor precisión teórica lo resoluble desde la perspectiva exploratoria; por otro, especificar gran cantidad de modelos de medida a partir del Factorial.

Por último, se repasan someramente las etapas del modelado estadístico: especificación, identificación, estimación y verificación, para así asentar conceptos básicos de planteamientos más generales de análisis de ecuaciones de estructura lineal, LISREL. Al igual que otros temas tratados sólo sucintamente, se proporciona la bibliografía adecuada para su consulta.

Este trabajo se ha escrito pensando en cursos para post-graduados, cuya formación no sea esencialmente matemática. Por ello, resulta apropiado como complemento en cursos de especialización en facultades de Ciencias Sociales, de la conducta y biológicas, o en seminarios introductorios al análisis estadístico multivariable.

2.1. Introducción histórica

En el siglo XIX, en Londres, en un intento para identificar criminales a partir de sus características físicas, se propone tomar un conjunto de 12 medidas del cuerpo de estos delincuentes. Galton critica este procedimiento alegando que varias de las medidas tomadas estarían altamente intercorrelacionadas (extremidades, etc.), proporcionando por consiguiente información redundante. Posteriormente, McDonald, colaborador de K. Pearson, midió 7 características del cuerpo en 3.000 criminales y publicó los resultados en una matriz de correlaciones. Pearson estaba convencido que los índices ideales, que resumirían lo esencial de las medidas utilizadas para la identificación, se corresponderían con los ejes perpendiculares del elipsoide de inercia, obtenido cuando estas medidas se plasmasen en el espacio de las 7 variables observables (v.o.).

En 1933, Hottelling, basándose en aquel artículo de Pearson, propone un algoritmo para hallar dichos ejes, lo cual supuso un gran avance en la diagonalización de matrices simétricas, pues en esencia requería obtener los vectores y valores propios de la matriz de correlaciones de las p variables originales.

2.2. El Análisis de Componentes Principales. Una aplicación

La concepción de esta técnica descansa en la utilización del Análisis Multivariable (AM), que tiene por finalidad reducir la dimensión de un conjunto de p variables observables. Resulta pues adecuada para describir, clasificar, predecir, etc. (Lebart, Morineau, Tabard, 1977). Usualmente esta técnica de análisis se aplica con el fin de obte-

TABLA 1. Relación de indicadores compuestos utilizados en el análisis multivariable.

* 1.	PO79 Población municipal 1979 (% del total)
* 2.	DE79 Densidad demográfica 1979 (Hab/Ha)
* 3.	CV75 Crecimiento vegetativo 1975 (‰)
4.	CV76 Crecimiento vegetativo 1976 (‰)
5.	NA76 Nacimientos 1976 (‰)
6.	MO76 Muertes 1976 (‰)
* 7.	MA76 Matrimonios 1976 (‰)
* 8.	CA70 Crecimiento anual acumulativo 1950/70
* 9.	CA75 Crecimiento anual acumulativo 1971/75
*10.	CA81 Crecimiento anual acumulativo 1976/81
11.	FAMI Familias en núcleo diferente principal (%)
12.	SL72 Superficie cultivada 1972 (% con respecto a la censada)
*13.	SLLA Superficie cultivada 1972 (% por habitante)
*14.	PAPR Población activa sector I 1975 (%)
*15.	PASE Población activa sector II 1975 (%)
*16.	PATE Población activa sector III 1975 (%)
*17.	NR75 Nivel renta 1975 (índice Banesto)
18.	NR70 Nivel renta 1970 (índice Banesto)
19.	NR65 Nivel renta 1965 (índice Banesto)
*20.	TH79 Teléfonos 1970 (por 1.000 habitantes)
*21.	DP76 Total gastos muncipal 1976 (por habitante) (TELF)
*22.	EH79 Consumo energía eléctrica 1979 (por habitante) (ENEL)
23.	TH81 Turismos 1981 (por 1.000 habitantes)
*24.	TETU Teléfonos + turismos (por 1.000 habitantes)
25.	BH79 Bancos y Cajas 1979 (por habitante)
26.	FH78 Farmacias 1978 (por habitante)
27.	AH79 Ambulancias 1979 (por habitante)
28.	LH79 Camas hospitalarias 1979 (por 1.000 habitantes)
29.	AP78 Alumnos preescolar 1978/79 (por 1.000 habitantes)
30.	AE78 Alumnos EGB 1978/79 (por 1.000 habitantes)
31.	AB78 Alumnos BUP 1978/79 (por 1.000 habitantes)
32.	AF78 Alumnos FP 1978/79 (por 1.000 habitantes)
*33.	EQES AP78 + AE78 + AB78 + AF78
34.	NI73 Nuevas inversiones 1964/73 (% del total)
35.	NI78 Nuevas inversiones 1974/78 (% del total)
*36.	AI73 Ampliación inversiones 1964/73 (% del total)
37.	AI78 Ampliación inversiones 1974/78 (% del total)
38.	EI80 Empleo industrial 1980 (% población)
*39.	EMIN Empleo industrial 1980 (% del total)
40.	VS80 Viviendas secundarias 1981 (% del total)
*41.	VISE Viviendas secundarias 1981 (por habitante)
42.	PFLI Población flotante 1981 (% del total)
*43.	PFLO Población flotante 1981 (por habitante)
*44.	ACCE Accesibilidad a la red de carreteras
45.	ALCA Altura.

ner un número más reducido de variables compuestas que se denominan componentes, y que por su mayor relevancia conceptual pueden, en posteriores aplicaciones, sustituir a las primitivas variables (Batista y Estivill, 1983; Moser y Wolf, 1961; Jollife, 1972, 1973; Daling y Tamura, 1970: 260-268).

Considerar la forma natural de la nube de puntos, implica que la transformación idónea de las v.o. es la ortogonal definida por los vectores propios de la matriz de covarianzas. Matemáticamente esto se traduce en transformar linealmente el vector de v.o., $x = (x_1, \ldots, x_p)$, en un nuevo conjunto de variables compuestas estandarizadas, $y = (y_1, \ldots, y_p)$, que están intercorrelacionadas entre sí y cuya variancia decrece a partir de la primera componente. La técnica de ACP está orientada a explicar, en el sentido de la regresión, la mayor proporción de variancia de las v.o. mediante el menor número posible de componentes.

Entre las usuales aplicaciones del ACP, mencionadas anteriormente, merece destacarse aquella que trata de establecer la dimensionalidad latente de un conjunto de v.o. Discernir lo esencial posibilita prescindir de la información redundante que existiese en las p variables originales, y considerar únicamente la proporcionada por un conjunto menor de m ($< p$) variables no observables, conocidas como Componentes Principales.

El estudio a que se refieren Batista y Estivill (1983) trata de definir una clasificación tipológica de los municipios de Cataluña, de acuerdo con su comportamiento respecto a una serie de variables socioeconómicas consideradas relevantes. En primer lugar se utiliza el ACP (Lebart y Morineau, 1982; Nie, Hull *et al.*, 1975) para decidir el reducido número (m) de componentes, que determina, bien una estructura causal subyacente o bien un patrón de covariación estable entre las v.o., y permite explicar gran parte de la variabilidad del conjunto original. Posteriormente, en base a la clasificación de los municipios, según su situación en las $m = 6$ primeras componentes, se procede a un Análisis de las Agrupaciones (Cluster Analysis) que partirá a la población de municipios en clases razonablemente tipificadas. Algunos antecedentes de la conjunción de ambas técnicas pueden encontrarse en Green, Frank y Robinson (1967), Everitt, Gourlay y Kendell (1971); Lebart, Morineau y Fenelon (1979).

La matriz de datos X(Nxp) de la que finalmente se parte, incluye $p = 22$ v.o. (tabla 1), y $N = 296$ municipios con un número de habitantes superior a 1.500 (95% de la población de Cataluña distribuida de forma representativa).

Ya que las variables se midieron en escalas de intervalo o de razón, la matriz varianzas-covarianzas de las p v.o. estandarizadas coincidirá con la matriz de correlaciones de Pearson, R (tabla 2), que es el punto de partida del análisis estadístico de los datos. Por otro lado, resulta adecuada la matriz R, como base del análisis posterior pues, en general, los gráficos bivariantes de v.o. no exhibían patrones de relación claramente no lineales.

Como se verá más adelante, uno de los resultados del análisis de la matriz R, es la obtención de los coeficientes de las ecuaciones de la tabla 4 que expresan la transformación lineal de las $p = 22$ v.o., en el conjunto de 22 componentes. De la tabla 3 se seleccionan las 5 primeras más relevantes.

Cuanto mayor sea el coeficiente *(factor score)* más realce tiene la variable en la componente particular de que se trate. A partir de estas ecuaciones puede evaluarse la

TABLA 2. Matriz de correlaciones de Pearson entre las 22 variables medidas.

	PO79	DE79	CV75	MA76	CA70	CA75	CAB1	PAPR	PASE	PATE	SLLA	EMIN	AI73	PFLO	VISE
PO79	1.00	0.75	0.37	-0.18	0.34	0.28	0.03	-0.32	0.17	0.29	-0.22	0.93	0.53	-0.10	-0.13
DE79	0.75	1.00	0.42	-0.25	0.47	0.32	0.06	-0.31	0.24	0.17	-0.23	0.64	0.31	-0.09	-0.12
CV75	0.37	0.42	1.00	-0.46	0.63	0.58	0.29	-0.44	0.37	0.18	-0.38	0.37	0.29	0.09	0.00
MA76	-0.18	-0.25	-0.46	1.00	-0.50	-0.48	-0.40	0.46	-0.48	0.00	0.36	-0.21	-0.18	-0.02	0.02
CA70	0.34	0.47	0.63	-0.50	1.00	0.51	0.31	-0.49	0.48	0.08	-0.40	0.34	0.26	-0.02	-0.08
CA75	0.28	0.32	0.58	-0.48	0.51	1.00	0.45	-0.50	0.39	0.23	-0.43	0.30	0.29	0.13	0.06
CAB1	0.03	0.06	0.29	-0.40	0.31	0.45	1.00	-0.31	0.24	0.17	-0.31	0.04	0.04	0.20	0.20
PAPR	-0.32	-0.31	-0.44	0.46	-0.49	-0.50	-0.31	1.00	-0.85	-0.38	0.75	-0.38	-0.26	0.00	0.05
PASE	0.17	0.24	0.37	-0.48	0.48	0.39	0.24	-0.85	1.00	-0.16	-0.60	0.26	0.19	-0.15	-0.16
PATE	0.29	0.17	0.18	0.00	0.08	0.23	0.17	-0.38	-0.16	1.00	-0.35	0.26	0.16	0.26	0.19
SLLA	-0.22	-0.23	-0.38	0.36	-0.40	-0.43	-0.31	0.75	-0.60	-0.35	1.00	-0.26	-0.19	-0.13	-0.06
EMIN	0.93	0.64	0.37	-0.21	0.34	0.30	0.04	-0.38	0.26	0.26	-0.26	1.00	0.59	-0.13	-0.16
AI73	0.53	0.31	0.29	-0.18	0.26	0.29	0.04	-0.26	0.19	0.16	-0.19	0.59	1.00	-0.10	-0.13
PFLO	-0.10	-0.09	0.09	-0.02	-0.02	0.13	0.20	0.00	-0.15	0.26	-0.13	-0.13	-0.10	1.00	0.80
VISE	-0.13	-0.12	0.00	0.02	-0.08	0.06	0.20	0.05	-0.16	0.19	-0.06	-0.16	-0.13	0.80	1.00
NR75	0.12	0.09	0.32	-0.30	0.20	0.39	0.42	-0.43	0.17	0.51	-0.46	0.16	0.10	0.57	0.51
DP76	0.07	0.03	0.22	-0.13	0.13	0.22	0.21	-0.20	0.01	0.36	-0.29	0.10	0.05	0.75	0.64
TELF	0.06	0.04	0.23	-0.13	0.14	0.22	0.29	-0.28	0.04	0.47	-0.35	0.09	0.02	0.59	0.43
TFTU	-0.01	-0.03	0.17	-0.07	0.05	0.14	0.27	-0.24	-0.01	0.45	-0.30	0.02	-0.04	0.61	0.48
ENEL	-0.08	-0.10	0.09	-0.10	-0.01	0.10	0.28	-0.13	-0.04	0.31	-0.23	-0.08	-0.08	0.74	0.64
EQES	0.21	0.15	0.22	-0.14	0.19	0.27	0.28	-0.36	0.14	0.44	-0.33	0.25	0.16	0.01	-0.04
ACCE	-0.22	-0.28	-0.45	0.51	-0.45	-0.46	-0.35	0.55	-0.61	0.05	0.42	-0.27	-0.21	-0.04	0.00

	PO79	DE79	CV75	MA76	CA70	CA75	CAB1	PAPR	PASE	PATE	SLLA	EMIN	AI73	PFLO	VISE

	NR75	DP76	TELF	TFTU	ENEL	EQES	ACCE
PO79	0.12	0.07	0.06	-0.01	-0.08	0.21	-0.22
DE79	0.09	0.03	0.04	-0.03	-0.10	0.15	-0.28
CV75	0.32	0.22	0.23	0.17	0.09	0.22	-0.45
MA76	-0.30	-0.13	-0.13	-0.07	-0.10	-0.14	0.51
CA70	0.20	0.13	0.14	0.05	-0.01	0.19	-0.45
CA75	0.39	0.22	0.22	0.14	0.10	0.27	-0.46
CAB1	0.42	0.21	0.29	0.27	0.28	0.28	-0.35
PAPR	-0.43	-0.20	-0.28	-0.24	-0.13	-0.36	0.55
PASE	0.17	0.01	0.04	-0.01	-0.04	0.14	-0.61
PATE	0.51	0.36	0.47	0.45	0.31	0.44	0.05
SLLA	-0.46	-0.29	-0.35	-0.30	-0.23	-0.33	0.42
EMIN	0.16	0.10	0.09	0.02	-0.08	0.25	-0.27
AI73	0.10	0.05	0.02	-0.04	-0.08	0.16	-0.21
PFLO	0.57	0.75	0.59	0.61	0.74	0.01	-0.04
VISE	0.51	0.64	0.43	0.48	0.64	-0.04	0.00
NR75	1.00	0.76	0.70	0.65	0.59	0.36	-0.33
DP76	0.76	1.00	0.68	0.66	0.63	0.16	-0.19
TELF	0.70	0.68	1.00	0.94	0.54	0.26	-0.14
TFTU	0.65	0.66	0.94	1.00	0.58	0.21	-0.12
ENEL	0.59	0.63	0.54	0.58	1.00	0.15	-0.07
EQES	0.36	0.16	0.26	0.21	0.15	1.00	-0.11
ACCE	-0.33	-0.19	-0.14	-0.12	-0.07	-0.11	1.00

	NR75	DP76	TELF	TFTU	ENEL	EQES	ACCE

TABLA 3. Vectores propios normalizados asociados a los cinco mayores valores propios de la matriz de correlaciones de la tabla A.2 (Resultados del subprograma FACTOR- PA1 del SPSS).

COMPONENTES PRINCIPALES DE LOS MUNICIPIOS DE CATALUÑA
File NO NAME (Creation date = 3-Aug-83)

Factor score coefficients

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
P079	0,06513	−0,10914	0,30166	0,11665	−0,07393
DE79	0,06250	−0,10894	0,19237	0,18408	−0,00440
CV75	0,09422	−0,06839	−0,02391	0,16535	0,23606
MA76	−0,07974	0,06592	0,17620	−0,14398	−0,16634
CA70	0,08539				
CA75	0,09706				
CA81	0,07663				
PAPR	−0,10919				
PASE	0,07450				
PATE	0,07367				
SLLA	−0,10285				
EMIN	0,06999				
AI73	0,04990				
PFLO	0,06342				
VISE	0,04725				
NR75	0,11407				
DP76	0,09170				
TELF	0,09542				
TFTU	0,08575	0,13331	0,04252	−0,08949	−0,09500
ENEL	0,06955	0,14335	−0,00642	0,07879	−0,04644
EQES	0,06520	−0,01213	0,07250	−0,41693	0,31641
ACCE	−0,08506	0,06961	0,17518	−0,14558	0,14345

TABLA 4

$$
\begin{aligned}
y_1 &= .06513\,x_1 + .06250\,x_2 + \ldots - .08506\,x_{22} \\
y_2 &= -.10514\,x_1 - .10894\,x_2 - \ldots + .06961\,x_{22} \\
y_3 &= .30166\,x_1 + .19327\,x_2 - \ldots + .17518\,x_{22} \\
y_4 &= .11665\,x_1 + .18406\,x_2 + \ldots - .14558\,x_{22} \\
y_5 &= -.07393\,x_1 - .00440\,x_2 + \ldots + .14345\,x_{22}
\end{aligned}
$$

puntuación de cada municipio en una componente específica, sustituyendo los valores que éste toma en cada una de las 22 v.o.

La importancia de cada componente, y_k, es función de la magnitud de la varianza, λ_k, que ésta logra explicar —en el sentido de la regresión— de la varianza total, p.

De la tabla 5, se recogen a continuación las cantidades absolutas y relativas de varianza explicada por las cinco componentes principales.

Componente (y_k)	Varianza explicada (λ_k)	Ratio $\frac{\lambda}{p}$ (%)
y_1	6.7840	30.84
y_2	4.5805	20.82
y_3	2.1661	9.85
y_4	1.3911	6.32
y_5	1.0588	4.81
		72.64

En la misma tabla puede comprobarse que la varianza extraída por todas las componentes, coincide con la varianza total,

$$\lambda_1 + \lambda_2 + \lambda_3 + \ldots + \lambda_{22} = 6.7840 + \ldots + 0.0005 = 22$$

Obtenidas las componentes, es útil disponer de una ordenación de los municipios y de las v.o. según estas nuevas dimensiones. En efecto, las coordenadas o proyecciones de los individuos (unidades del análisis) sobre las, $m = 6$, primeras componentes, que son los nuevos ejes ortogonales (tabla 6) permiten; 1.°) Facilitar la interpretación de las componentes (tabla 7), y 2.°) Sustituir por un número reducido de componentes incorrelacionadas, el conjunto inicial de v.o. no independientes, ya sea como regresores en el modelo lineal (Jollife, 1972, 1973; Daling y Tamura, 1970: 260-268), ya para facilitar el proceso de agrupamiento de los individuos (Green, Frank y Robinson, 1967; Everitt, Gourlay y Kendell, 1971, Batista y Estivill, 1983).

Adicionalmente la proyección (a_{ik}) de una v.o. (x_i) sobre una componente particular (y_k) evalúa el grado de relación lineal entre las mismas, pues, como se verá en el apartado 2.3.3 coincide con su coeficiente de correlación. La matriz $A(pxm)$, definida por el conjunto de estas correlaciones se denomina matriz de saturaciones o ponderaciones factoriales (tabla 8) y permite representar las v.o. en el subespacio que las componentes determinan (tabla 9) o expresarlas en función de las componentes principales, según ecuaciones estocásticas del tipo:

$$\begin{aligned}
x_1 &= -.44y_1 + .50y_2 + \ldots .08y_5 + e_1 \\
x_2 &= -.43y_1 + .50y_2 + \ldots .01y_5 + e_2 \\
&\cdots\cdots\cdots \\
&\cdots\cdots\cdots \\
x_{22} &= +.58y_1 - .32y_2 + \ldots -.16y_5 + e_{22}
\end{aligned}$$

TABLA 5. Valores propios y porcentaje de la variancia total que explican las quince primeras componentes.

EDITION DES VALEURS-PROPRES

SOMME DES VALEURS-PROPRES 22.00000381

HISTOGRAMME DES PREMIERES VALEURS-PROPRES

	VALEUR-PROPRE	POURCENTAGE	POURCENTAGE CUMULE	
1	6.78396845	30.84	30.84	**********
2	4.58050585	20.82	51.66	**********
3	2.16610432	9.85	61.50	**********
4	1.39110470	6.32	67.83	**********
5	1.05878806	4.81	72.64	**********
6	0.80119973	3.64	76.28	**********
7	0.69710279	3.17	79.45	**********
8	0.66131008	3.01	82.45	**********
9	0.55502397	2.52	84.98	**********
10	0.50375825	2.29	87.27	**********
11	0.48287237	2.19	89.46	**********
12	0.37680340	1.71	91.18	**********
13	0.36587447	1.66	92.84	**********
14	0.32776672	1.49	94.33	**********
15	0.30636427	1.39	95.72	**********

EDITION SOMMAIRE DES VALEURS-PROPRES DE 16 A 22

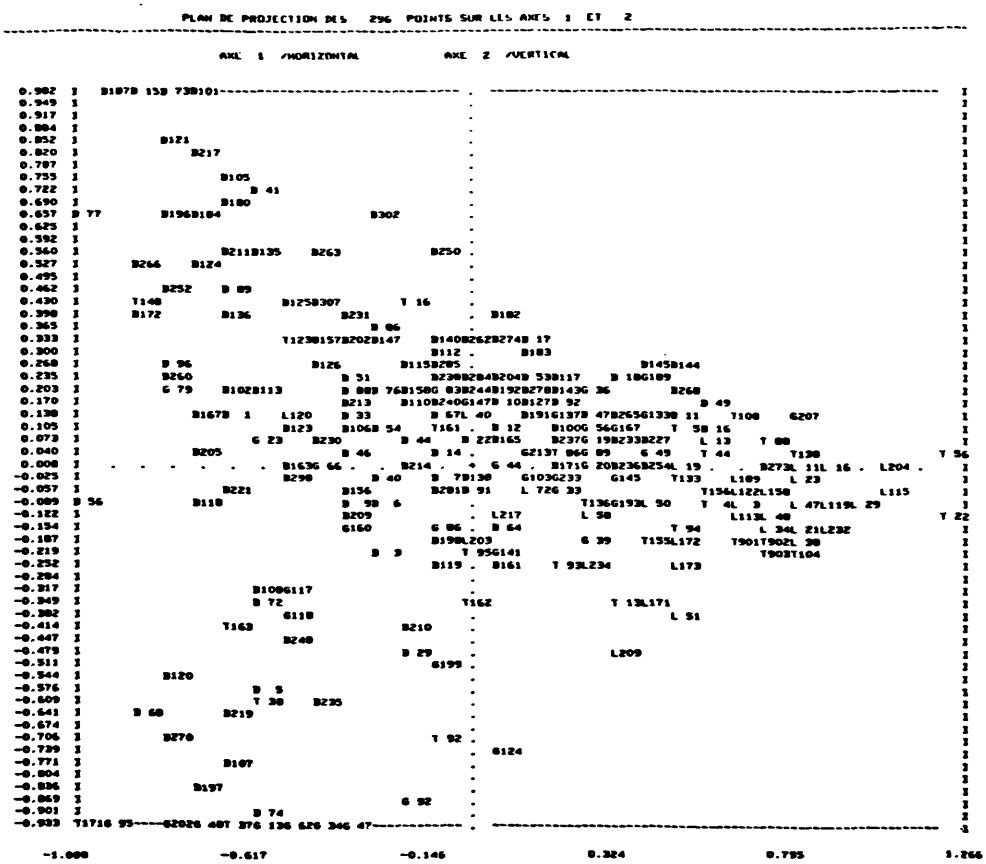
0.27950725 0.25406459 0.18576765 0.12766898 0.04861223 0.04528928 0.00054610

TABLA 6. Edición de coordenadas y contribuciones de los municipios.
(a) Código municipal. (b) Peso relativo de cada elemento. (c) Cuadrado de la distancia (χ^2) en el origen del elemento (indicador de su carácter periférico). (d) Coordenadas del municipio sobre las seis primeras componentes. (e) Contribución absoluta (sobre 100) del elemento. (f) Contribución relativa del municipio.

EDITION DES COORDONNEES ET DES CONTRIBUTIONS DES INDIVIDUS

(a)	(b)	(c)	(d)						(e)						(f)					
NOMS	MASSES	DISTO	COORDONNEES						CONTRIBUTIONS ABSOLUES*100						CONTRIBUTIONS RELATIVES					
			F1	F2	F3	F4	F5	F6	F1	F2	F3	F4	F5	F6	F1	F2	F3	F4	F5	F6
1	1.000	1.11	-0.62	0.14	-0.46	-0.07	-0.55	0.13	0.42	0.03	0.71	0.03	2.10	0.17	0.35	0.02	0.19	0.00	0.27	0.02
3	1.000	0.34	-0.25	-0.22	-0.23	0.01	-0.07	0.14	0.07	0.08	0.19	0.00	0.04	0.18	0.18	0.14	0.16	0.00	0.01	0.06
5	1.000	1.22	-0.54	-0.58	-0.21	0.11	0.21	-0.17	0.32	0.55	0.15	0.06	0.32	0.26	0.24	0.28	0.04	0.01	0.04	0.02
6	1.000	0.31	-0.19	-0.11	0.07	-0.22	0.21	0.01	0.04	0.02	0.02	0.26	0.30	0.00	0.12	0.04	0.02	0.16	0.14	0.00
7	1.000	0.34	-0.11	-0.05	-0.22	0.00	0.06	0.10	0.01	0.00	0.16	0.00	0.02	0.09	0.03	0.01	0.14	0.00	0.01	0.03
9	1.000	0.32	-0.28	-0.11	-0.27	-0.05	-0.06	0.12	0.08	0.02	0.26	0.01	0.03	0.14	0.24	0.04	0.23	0.01	0.01	0.05
10	1.000	0.21	0.11	0.17	-0.29	-0.05	0.16	0.11	0.01	0.05	0.28	0.02	0.17	0.12	0.06	0.13	0.39	0.01	0.11	0.06
11	1.000	0.60	0.60	0.12	-0.10	0.08	0.35	-0.10	0.39	0.02	0.03	0.04	0.84	0.09	0.60	0.02	0.02	0.01	0.20	0.02
12	1.000	0.25	0.12	0.09	-0.16	-0.02	0.14	-0.02	0.02	0.01	0.09	0.00	0.14	0.00	0.06	0.03	0.10	0.00	0.08	0.00
14	1.000	0.15	-0.08	0.02	-0.19	-0.02	0.14	-0.08	0.01	0.00	0.12	0.00	0.13	0.06	0.04	0.00	0.23	0.00	0.12	0.04
15	1.000	4.94	-1.08	1.26	1.26	0.41	0.20	0.09	1.28	2.57	5.41	0.89	0.29	0.08	0.24	0.32	0.32	0.03	0.01	0.00
16	1.000	0.83	0.67	0.08	-0.07	0.15	0.10	-0.01	0.49	0.01	0.01	0.13	0.07	0.00	0.54	0.01	0.01	0.03	0.01	0.00
17	1.000	0.33	0.15	0.31	-0.32	0.13	-0.02	-0.09	0.02	0.15	0.35	0.10	0.00	0.07	0.07	0.29	0.31	0.05	0.00	0.02
18	1.000	0.48	0.43	0.21	-0.18	-0.06	0.33	-0.02	0.20	0.07	0.11	0.02	0.74	0.00	0.38	0.09	0.07	0.01	0.22	0.00
22	1.000	0.34	-0.02	0.04	0.14	-0.50	0.05	0.05	0.00	0.00	0.07	1.32	0.02	0.03	0.00	0.01	0.06	0.73	0.01	0.01
29	1.000	0.54	-0.17	-0.48	-0.10	0.04	0.19	0.22	0.03	0.38	0.04	0.01	0.26	0.45	0.06	0.44	0.02	0.00	0.07	0.09
31	1.000	0.22	0.07	0.08	-0.17	-0.18	0.04	-0.09	0.01	0.01	0.10	0.18	0.01	0.08	0.02	0.03	0.13	0.15	0.01	0.04
33	1.000	0.34	-0.31	0.13	-0.15	-0.11	-0.03	-0.08	0.11	0.03	0.07	0.06	0.00	0.06	0.28	0.05	0.06	0.04	0.00	0.02
35	1.000	1.16	-0.51	-0.58	0.31	-0.43	0.11	-0.13	0.29	0.54	0.34	0.99	0.09	0.16	0.23	0.29	0.08	0.16	0.01	0.02
38	1.000	0.38	0.37	0.22	-0.21	-0.02	0.36	0.00	0.15	0.08	0.15	0.00	0.92	0.00	0.36	0.13	0.11	0.00	0.34	0.00
40	1.000	0.19	-0.26	-0.05	-0.14	-0.05	0.21	0.03	0.08	0.00	0.07	0.01	0.32	0.01	0.37	0.01	0.11	0.01	0.24	0.00
41	1.000	2.40	-0.57	0.71	-0.54	0.30	-0.29	-0.63	0.35	0.82	0.99	0.49	0.59	3.66	0.13	0.21	0.12	0.04	0.03	0.16
44	1.000	0.20	-0.12	0.07	-0.20	-0.11	0.02	0.07	0.02	0.01	0.13	0.07	0.00	0.04	0.07	0.02	0.20	0.07	0.00	0.02
46	1.000	0.24	-0.28	0.04	-0.25	-0.15	-0.05	0.01	0.08	0.00	0.22	0.13	0.01	0.00	0.31	0.01	0.26	0.10	0.01	0.00
47	1.000	0.45	0.35	0.11	-0.12	-0.17	0.41	-0.05	0.13	0.02	0.05	0.16	1.17	0.02	0.27	0.03	0.03	0.07	0.37	0.01
49	1.000	0.67	0.66	0.16	-0.11	0.11	0.36	0.03	0.47	0.04	0.05	0.07	0.90	0.01	0.64	0.04	0.02	0.02	0.19	0.00
51	1.000	0.44	-0.31	0.21	-0.03	-0.31	-0.20	0.05	0.10	0.07	0.00	0.50	0.29	0.02	0.21	0.10	0.00	0.21	0.09	0.00
53	1.000	0.63	0.18	0.22	-0.38	0.09	0.33	0.18	0.04	0.08	0.49	0.04	0.77	0.30	0.05	0.08	0.23	0.01	0.17	0.05

TABLA 7. (a) Gráfico de los municipios en el espacio de las dos primeras componentes.



Obviamente, la introducción del término de error en estas ecuaciones está justificado por cuanto se han omitido 17 de las 22 componentes finales.

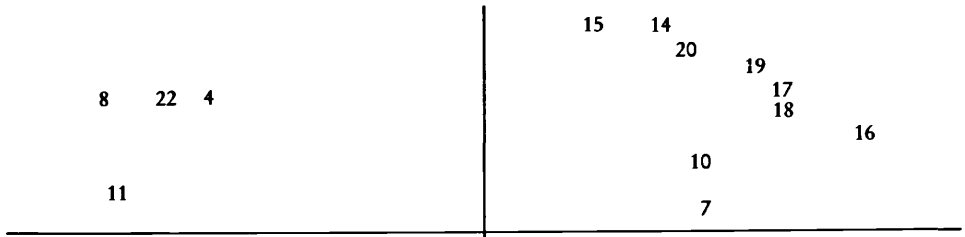
La trascendencia de esta matriz de saturaciones *A*, se debe a que posibilita deducir el alcance que en el análisis tienen cada v.o. y cada componente. Como se verá en el apartado 2.3.5, de las columnas de *A* puede deducirse la contribución de cada componente a la varianza total, y de las filas, puede obtenerse el porcentaje de la varianza de cada v.o. que explican las 5 primeras componentes —este % coincide con el concepto de comunalidad (tabla 10), propio del Análisis Factorial del apartado 2.3.2.

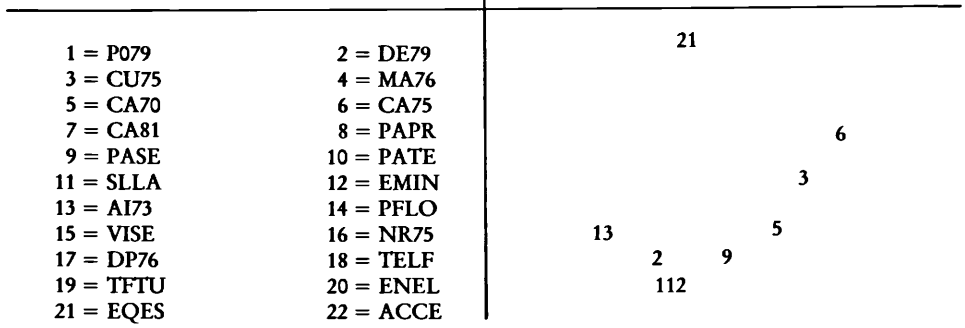
TABLA 8. (SPAD). (a) Identificación variable. (b) Coordenadas de la variable en el subespacio determinado por las seis primeras componentes. (c) Vectores propios de la matriz de correlaciones, R, en la tabla 3. (Se obtienen multiplicando las columnas de la tabla 3 por la raíz cuadrada del valor propio correspondiente.) Elevados al cuadrado representan la contribución absoluta de la variable en la inercia total del eje, y la suma por cada componente (en columna) debe valer la unidad. (d) Coeficientes de correlación entre componentes y factores —saturaciones—, coinciden con las coordenadas (b), siempre que se analiza R. Pueden obtenerse multiplicando (las columnas de la tabla 3 [o de (c)]) por el valor propio (o su raíz cuadrada) correspondiente. Elevados al cuadrado representan la contribución relativa de la variable al eje, siendo su suma para todos los ejes (en fila) igual a la unidad.

EDITION DES COORDONNEES ET DES CONTRIBUTIONS DES VARIABLES

(a)		(b)						(c)						(d)					
NOMS	E.TYPE	COORDONNEES						PROJECTION ANCIENS AXES UNITE (CARRE=CONTRIBUTION ABSOLUE)						CORRELATION VARIABLE-FACTEUR (CARRE=CONTRIBUTION RELATIVE)					
		F1	F2	F3	F4	F5	F6	F1	F2	F3	F4	F5	F6	F1	F2	F3	F4	F5	F6
PO79	0.499	-0.44	0.50	0.65	0.16	0.08	0.03	-0.17	0.23	0.44	0.14	0.08	0.04	-0.44	0.50	0.65	0.16	0.08	0.03
DE79	24.141	-0.43	0.50	0.42	0.26	0.01	-0.28	-0.16	0.23	0.28	0.22	0.01	-0.31	-0.43	0.50	0.42	0.26	0.01	-0.28
CV75	5.060	-0.64	0.31	-0.05	0.23	-0.26	-0.27	-0.25	0.15	-0.03	0.20	-0.25	-0.30	-0.64	0.31	-0.05	0.23	-0.26	-0.27
MA76	3.307	0.54	-0.30	0.38	-0.19	0.18	-0.06	0.21	-0.14	0.26	-0.17	0.17	-0.07	0.54	-0.30	0.38	-0.19	0.18	-0.06
CA70	17.301	-0.58	0.43	-0.16	0.21	-0.18	-0.31	-0.22	0.20	-0.11	0.18	-0.18	-0.35	-0.58	0.43	-0.16	0.21	-0.18	-0.31
CA75	1.422	-0.66	0.25	-0.16	0.10	-0.34	0.07	-0.25	0.12	-0.11	0.09	-0.34	0.07	-0.66	0.25	-0.16	0.10	-0.34	0.07
CA81	2.160	-0.52	-0.06	-0.32	-0.03	-0.49	0.24	-0.20	-0.03	-0.22	-0.02	-0.48	0.27	-0.52	-0.06	-0.32	-0.03	-0.49	0.24
PAPR	16.847	0.74	-0.33	0.23	0.33	-0.30	-0.02	0.28	-0.16	0.16	0.28	-0.30	-0.02	0.74	-0.33	0.23	0.33	-0.30	-0.02
PASE	15.775	-0.51	0.48	-0.50	-0.08	0.42	0.05	-0.19	0.23	-0.34	-0.07	0.40	0.06	-0.51	0.48	-0.50	-0.08	0.42	0.05
PATE	8.964	-0.50	-0.23	0.45	-0.47	-0.16	-0.06	-0.19	-0.11	0.31	-0.40	-0.15	-0.07	-0.50	-0.23	0.45	-0.47	-0.16	-0.06
SLLA	51.403	0.70	-0.14	0.21	0.33	-0.28	0.00	0.27	-0.07	0.14	0.28	-0.27	0.00	0.70	-0.14	0.21	0.33	-0.28	0.00
EMIN	0.454	-0.48	0.51	0.59	0.10	0.14	0.12	-0.18	0.24	0.40	0.09	0.13	0.14	-0.48	0.51	0.59	0.10	0.14	0.12
AI73	0.731	-0.34	0.41	0.40	0.09	0.05	0.47	-0.13	0.19	0.27	0.08	0.05	0.53	-0.34	0.41	0.40	0.09	0.05	0.47
PFLO	243.666	-0.43	-0.75	0.04	0.31	0.06	0.06	-0.16	-0.35	0.03	0.26	0.05	0.06	-0.43	-0.75	0.04	0.31	0.06	0.06
VISE	26.944	-0.32	-0.71	0.01	0.35	0.06	0.18	-0.12	-0.33	0.01	0.30	0.05	0.20	-0.32	-0.71	0.01	0.35	0.06	0.18
NR75	1.396	-0.77	-0.43	0.01	-0.07	-0.02	0.09	-0.30	-0.20	0.01	-0.06	-0.01	0.10	-0.77	-0.43	0.01	-0.07	-0.02	0.09
DP76	2464.310	-0.62	-0.59	0.11	0.20	0.13	0.02	-0.24	-0.27	0.07	0.17	0.12	0.03	-0.62	-0.59	0.11	0.20	0.13	0.02
TELF	141.835	-0.65	-0.54	0.11	-0.13	0.06	-0.28	-0.25	-0.25	0.08	-0.11	0.06	-0.31	-0.65	-0.54	0.11	-0.13	0.06	-0.28
TFTU	166.502	-0.58	-0.61	0.09	-0.13	0.10	-0.26	-0.22	-0.29	0.06	-0.11	0.10	-0.29	-0.58	-0.61	0.09	-0.13	0.10	-0.26
ENEL	353.722	-0.47	-0.66	-0.01	0.11	0.05	0.17	-0.18	-0.31	-0.01	0.09	0.05	0.19	-0.47	-0.66	-0.01	0.11	0.05	0.17
EQES	68.149	-0.44	0.06	0.16	-0.58	-0.33	0.14	-0.17	0.03	0.11	-0.49	-0.32	0.15	-0.44	0.06	0.16	-0.58	-0.33	0.14
ACCE	36.886	0.58	-0.32	0.38	-0.20	-0.16	-0.09	0.22	-0.15	0.26	-0.17	-0.15	-0.10	0.58	-0.32	0.38	-0.20	-0.16	-0.09

TABLA 9. Situación de las variables según sus coordenadas (tabla 8) en el espacio determinado por las dos primeras componentes (Resultados del subprograma FACTOR-PA1 del SPSS).





1 = P079 2 = DE79
3 = CU75 4 = MA76
5 = CA70 6 = CA75
7 = CA81 8 = PAPR
9 = PASE 10 = PATE
11 = SLLA 12 = EMIN
13 = AI73 14 = PFLO
15 = VISE 16 = NR75
17 = DP76 18 = TELF
19 = TFTU 20 = ENEL
21 = EQES 22 = ACCE

TABLA 10. Porcentaje de la varianza de cada variable explicado por las cinco primeras componentes. Variables mejor representadas en este sistema más parco: P079, PAPR y PASE; la peor representada es la AI73, que según la tabla 8 (c) es relevante en la sexta componente (Resultados del subprograma FACTOR-PA1 del SPSS).

Variable	(Communality)
P079	0,90637
DE79	0,66895
CV75	0,62516
MA76	0,60138
CA70	0,62417
CA75	0,65389
CA81	0,62606
PAPR	0,91772
PASE	0,91904
PATE	0,75183
SLLA	0,73415
EMIN	0,87083
AI73	0,45203
PFLO	0,84931
VISE	0,73621
NR75	0,78341
DP76	0,79240
TELF	0,74538
TETU	0,74659
ENEL	0,66954
EQES	0,67056
ACCE	0,64358

TABLA 11

a_{ik}^2	y_1	y_2	y_3	y_4	y_5	$h_i^2 = \sum_{k=1}^{5} a_{ik}^2$
x_1	$(-.44)^2$	$(\ .50)^2$	$(.65)^2$	$(\ .16)^2$	$(\ .08)^2$	.9064
x_2	$(-.43)^2$	$(\ .50)^2$	$(.42)^2$	$(\ .26)^2$	$(\ .01)^2$	.6689
. . .			. . .			. . .
x_{22}	$(\ .58)^2$	$(-.32)^2$	$(.38)^2$	$(-.20)^2$	$(-.16)^2$	.6436
$\lambda_k = \sum_{i=1}^{22} a_{ik}^2$	6,784	4,580	2,166	1,391	1,059	$\sum_{k=1}^{5} \lambda_k = \sum_{i=1}^{22} h_i^2 = 15{,}9875$

En la tabla 5 puede comprobarse que el porcentaje de varianza explicado por las $m = 5$ componentes principales es efectivamente

$$\frac{\Sigma\lambda_k}{P} = \frac{15.9805}{22} = .7264$$

La selección de variables, que inicialmente se hizo con la finalidad de cubrir el dominio objeto de estudio, redunda en aumentar la heterogeneidad de las mismas (DP76 y PASE; ACCE y VISE). Esto ocasiona la obtención de componentes con un poder explicativo reducido —las tres primeras sólo representan el 61,5% de la variabilidad total (tabla 5)— y por lo tanto debe procederse con cautela al validar los nuevos ejes. A continuación se detalla la «etiquetación» de aquellas componentes cuya interpretación teórica es clara.

La componente I es un indicador de «tamaño» que podría llamarse *calidad de vida urbana.* Las correlaciones más elevadas de la componente (tabla 8), se dan con la NR75 (− 0,77), PAPR (0,74), SLLA (0,70), CA75 (− 0,66) y TELF (− 0,65). Esta primera componente discrimina claramente entre los municipios con unos indicadores de renta agregada más elevados (nivel de renta, número de teléfonos por habitante, gastos municipales, etc.) y con una dinámica demográfica expansiva entre los años 70-75, y los municipios más rurales y deprimidos, agrarios y con mala accesibilidad. La jerarquía definida por las coordenadas municipales correspondientes a la primera componente se puede asimilar a una ordenación de acuerdo con un factor de tamaño cualitativo determinado por el nivel de vida urbana. Los valores más significativos según la pri-

mera coordenada factorial, corresponden a un conjunto de municipios formado, por una parte, por los que tienen un fuerte componente urbano (área metropolitana de Barcelona, Tarragona, Gerona, Mataró...) y por otra, todos aquellos núcleos más turísticos caracterizados por un alto nivel de vida y de consumo (Lloret, Castell d'Aro, Tossa, Salou...). En el extremo opuesto encontramos los núcleos más rurales y deprimidos de todo el conjunto analizado (Batea, la Fatarella, Serós, Isona, etc.).

Hay que tener en cuenta a la hora de analizar los resultados obtenidos, que la variable de más alta correlación con la primera componente la NR75, se refiere al nivel de renta por cápita municipal estimado para el año 1975, año en que la crisis industrial acababa de empezar.

La componente II es un indicador que muestra el grado de *especialización económica* de cada municipio. Esta componente presenta una alta correlación (tabla 8) con las variables PFLO (−0,75), VISE (−0,71), ENEL (−0,66) y TFTU (−0,61). La ordenación proporcionada por la proyección de los municipios sobre el eje 2, sitúa en primer lugar todos los municipios con una fuerte especialización turística (Castell d'Aro, Tossa, Lloret, Cadaqués...). Al tratarse de un factor ortogonal con respecto al primero, discrimina entre los municipios altamente significativos según aquella primera componente principal. Esto se ve claramente en la representación de la nube de puntos en el plano de los ejes 1 y 2 (tabla 7). Obsérvese la forma cónica abierta hacia los municipios de alta calidad de vida urbana, lo cual permite distinguir entre los núcleos eminentemente turísticos (Salou, Lloret, Sitges, Tossa, Sant Pol, Llançà...) y las áreas urbanas de especialización equilibrada (Sabadell, Mataró, Tarragona, Girona, Reus, Lleida...). Al extremo opuesto del gráfico, se sitúa el conjunto de todos aquellos municipios con una baja calidad de vida urbana, rurales y deprimidos que, lógicamente, no incluye ningún núcleo turístico ni ninguna área urbana importante. De aquí que todos se sitúen en la zona de mínima significación con respecto al eje 2.

La componente III se caracteriza (tabla 8) por las variables PO79 (0,65), EMIN (0,59) y PASE (−0,50). Se trata de una componente que destaca el *tamaño urbano*. Discrimina claramente entre los núcleos principales de la red urbana, de acuerdo con la población y un cierto desarrollo de las funciones terciarias (Sabadell, Badalona, Tarragona, Gerona, Reus...), en contraposición a los pequeños municipios industrializados y de rápido crecimiento demográfico (Sant Andreu de la Barca, Barberà, Montmeló, Canovelles...).

La aplicación del ACP a un conjunto inicial de variables, con la finalidad de sintetizar la información recogida por éstas, nos ha permitido adelantar conceptos esenciales del Análisis Estadístico Multivariable. A continuación formalizamos estas ideas sirviéndonos de las propiedades del álgebra matricial, para referirnos a ella al introducir el modelo de Análisis Factorial.

2.2.1. *Algoritmo de Cálculo*

El psicólogo francés M. Reuchlin (1964), comenta un sencillo ejemplo en el contexto del Análisis Factorial (AF), cuyos datos se utilizarán aquí para desarrollar el

ACP. La tabla 12 describe las puntuaciones de ocho alumnos en las asignaturas de Matemáticas (x_1), Ciencias Naturales (x_2), Francés (x_3) y Latín (x_4).

TABLA 12. Notas obtenidas por ocho alumnos en cuatro asignaturas.

x_1	x_2	x_3	x_4
13	12.5	8.5	9.5
14.5	14.5	15.5	15
5.5	7	14	11.5
14	14	12	12.5
11	10	5.5	7
8	8	8	8
6	7	11	9.5
6	6	5	5.5

Debe seleccionarse en primer lugar la medida de dependencia entre las variables: covariancias o correlaciones, pues las unidades afectan al concepto de distancia y por tanto a la variancia; así, no se obtienen soluciones equivalentes según se utilice una u otra con el simple reescalamiento de las componentes (Morrison, 1976). En nuestro caso se ha decidido efectuar el análisis en base a correlaciones, por tanto se supondrá a partir de ahora que las v.o. han sido previamente estandarizadas ($E(x) = 0$; $V(x) = 1$).

La matriz de correlaciones, R, entre estas variables es:

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{array} \begin{array}{c} \begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \end{array} \\ \begin{bmatrix} 1 & & & \\ 0.9829455 & 1 & & \\ 0.250345 & 0.4175258 & 1 & \\ 0.5429605 & 0.6823637 & 0.9483579 & 1 \end{bmatrix} \end{array} \qquad [1]$$

Se trata de obtener las combinaciones lineales $V(pxp)$ de las v.o.

$$Y = X\ V \qquad [2]$$

donde $Y(Nxp)$ y $X(Nxp)$

tales que cumplan aquellas condiciones mencionadas en el apartado 2.2. Así, sobre la primera componente

$$y_1 = Xv_1 \qquad [3]$$

debe proyectarse la mayor cantidad de varianza posible; es decir

$$V(y_1) = E(y_1'y_1) = E(v_1'X'Xv_1) = v_1'Rv_1 = \lambda_1 \qquad [4]$$

debe ser máximo, con

$$v_1' v_1 = 1 \qquad [5]$$

La segunda componente, que debe maximizar la varianza explicada residual (después de la extracción de y_1), se obtendría calculando el v_2 tal que

$$y_2 = Xv_2 \qquad [6]$$

$$\left.\begin{array}{l} y_2 \text{ está incorrelacionada con } y_1 \\ v(y_2) = \lambda_2, \text{ es máximo} \\ v_2' v_2 = 1 \end{array}\right\} \qquad [7]$$

análogamente se obtendrían las restantes componentes. El sistema de ecuaciones [2] junto con las restricciones [7], asociadas a cada ecuación, se resuelven por el método de los multiplicadores de Lagrange que conduce a la diagonalización de la matriz, R, de correlaciones entre la v.o. (Mulaik, 1972).

Es inmediato demostrar (Maxwell, 1977) que, λ_1, varianza de la primera componente, coincide con el mayor valor propio de R, siendo los coeficientes de la combinación lineal, los elementos del vector propio, v_1, normalizado y asociado a dicho valor propio. En general, se obtiene

$$RV = VD \qquad [8]$$

$$\text{con } D = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_p / \lambda_{k+1} < \lambda_k\} \qquad [9]$$

$$\text{y } V'V = 1 \qquad [10]$$

En nuestro caso los vectores y valores propios de R se han obtenido con un programa en BASIC (Hewlett Packar 9390 A)

$$V = \begin{bmatrix} v_1 & v_2 & v_3 & v_4 \\ .479846 & -.552457 & -.604236 & -.317084 \\ .531096 & -.402197 & .739926 & .093190 \\ .442562 & .631486 & .104691 & -.628016 \\ .540207 & .367910 & -.276494 & .704535 \end{bmatrix} \qquad [11]$$

$$D = \begin{bmatrix} 2.930085 & & & \\ (\lambda_1) & 1.067980 & & \\ & (\lambda_2) & .001400 & \\ & & (\lambda_3) & .000534 \\ & & & (\lambda_4) \end{bmatrix} \qquad [12]$$

2.2.2. *Interpretación de los valores propios*

Al examinar la matriz de correlaciones R, se observa, que la diagonal principal contiene las varianzas de las p, v.o., y los valores exteriores representan la parte com-

partida o explicada en común por pares de variables; estas correlaciones pueden entenderse por tanto, como una parte no fundamental o redundante.

No obstante, si se toma como base la matriz de vectores propios, V, que discrimina al máximo las v.o., la matriz R «se verá» en forma diagonal; en efecto, de [8] se tiene:

$$V'RV = D \qquad [13]$$

que generaliza [4]. Esta matriz, D no es más que una transformación ortogonal de R, por lo que contendrá, aunque resumida, toda la información relevante que aquella incluía, es decir, lo esencial, «su naturaleza o espectro», que se resume en el concepto de traza, Tr, valor que se conserva en toda transformación similar.

$$Tr(R) = \sum_{i=1}^{p} r_{ii} = p = \sum_{k=1}^{p} \lambda_k = Tr(D) \qquad [14]$$

La ecuación [14] aclara que toda la varianza del conjunto de v.o. ha sido «extraída» por las componentes (en la transformación); tiene sentido, pues, interpretar el valor propio, λ_k, como la parte de la varianza que el k-ésimo eje principal «explica», y el ratio $\frac{\lambda_k}{p}$ como índice de la importancia de esta componente en una descripción más parca del conjunto de v.o. Sin embargo, es la nueva versión de [8];

$$R = VDV' \qquad [15]$$

que estructura R en términos de vectores y valores propios, la que permite interpretar geométricamente estos conceptos de forma que resulten más manejables. Así, se desarrolla [15] según las direcciones principales de V,

$$R = \lambda_1 v_1 v_1' + \lambda_2 v_2 v_2' + \ldots + \lambda_p v_p v_p' \qquad [16]$$

cada término del segundo miembro determina las proyecciones sobre cada componente ($Rv_k = v_k\lambda_k$), es decir, expresa que la varianza del conjunto de v.o. proyectada sobre el vector v_k es precisamente λ_k.

Ya que en nuestro ejemplo, λ_3 y λ_4, son despreciables, puede prescindirse de las componentes asociadas a estos valores propios y reducir, en consecuencia, la dimensión de $p = 4$ a $m = 2$. En efecto, al reproducir según [16] la matriz de correlaciones

$$\hat{R} = \lambda_1 v_1 v_1' + \lambda_2 v_2 v_2' \qquad [17]$$

$$\hat{R} = \lambda_1 \begin{bmatrix} .230 & & & \\ .255 & .282 & & \\ .212 & .235 & .169 & \\ .259 & .286 & .239 & .292 \end{bmatrix} + \lambda_2 \begin{bmatrix} .304 & & & \\ .222 & .162 & & \\ .348 & .254 & .399 & \\ .203 & .148 & .232 & .135 \end{bmatrix}$$

$$= \begin{bmatrix} .674 & & & \\ .747 & .826 & & \\ .621 & .688 & .574 & \\ .759 & .840 & .700 & .855 \end{bmatrix} + \begin{bmatrix} .324 & & & \\ .237 & .173 & & \\ .371 & .271 & .426 & \\ .216 & .158 & .248 & .144 \end{bmatrix} \quad [18]$$

La matriz de residuos, $R - \hat{R}$, determina la posible distorsión en la que se hubiera podido incurrir al describir las v.o., x_1, x_2, x_3 y x_4, mediante dos únicas componentes y_1 e y_2.

$$R - \hat{R} = \begin{bmatrix} .002 & & & \\ .001 & .001 & & \\ .000 & .000 & .000 & \\ .000 & .000 & .000 & .001 \end{bmatrix} \quad [19]$$

2.2.3. *Relación entre variables originales y componentes*

La matriz R simétrica y semidefinida positiva, puede ser analizada siempre en función de una matriz A, llamada matriz factorial, y que se obtiene de [15] según:

$$R = VD^{1/2}D^{1/2}V' = AA' \quad [20]$$

con

$$A = VD^{1/2} \quad [21]$$

Este hecho, como se verá en el próximo apartado, fue utilizado por Thurstone (1931) para descomponer la matriz de correlaciones en el producto de una matriz A de menor rango, por su transpuesta A'. Dado que las componentes son ortogonales, ya se mencionó en el apartado que esta matriz A está constituida por las correlaciones entre x e y, que reciben el nombre de «saturaciones» o «ponderaciones», y relacionan ambos conjuntos de variables según la ecuación:

$$\underset{(p,1)}{x} = \underset{(p,m)}{A} \quad \underset{(m,1)}{y} \quad [22]$$

en nuestro caso, A es;

$$A = VD^{1/2} =$$

$$= \begin{bmatrix} .47984 & -.55146 & -.60423 & -.31708 \\ .53110 & -.40220 & \ldots\ldots\ldots & \\ .44256 & .63144 & \ldots\ldots\ldots & \\ .54021 & .36791 & \ldots\ldots\ldots & \end{bmatrix} \begin{bmatrix} 1.71175 & & & \\ & 1.03343 & & \\ & & \ldots\ldots\ldots & \\ & & \ldots\ldots\ldots & \end{bmatrix} \quad [23]$$

que sólo contiene dos columnas al considerar únicamente las componentes asociadas a los dos primeros vectores propios;

$$A = \begin{bmatrix} .8213 & .5699 \\ .9041 & .4156 \\ .7575 & .6526 \\ .9247 & .3802 \end{bmatrix} \qquad [24]$$

Esta, a su vez, constituye también una base de vectores propios ortogonales de R, ya que se ha obtenido al multiplicar por los escalares, $D^{-1/2}$, la matriz, V, de sus vectores propios normalizados. La condición de ortogonalidad equivalente para esta matriz de saturaciones es

$$A'A = (VD^{1/2})'(VD^{1/2}) = D^{1/2}V'VD^{1/2} = D^{1/2}ID^{1/2} = D \qquad [25]$$

matriz diagonal de valores propios.

Por último, [22] se verificará sólo si se considera igual número de componentes que de v.o.; en cualquier otro caso se cumple tanto mejor cuanto más despreciables sean los residuos de [19]. Estas condiciones forzaron la introducción del término de error en las ecuaciones de la tabla 4, sistema que ahora puede expresarse de forma general y sencilla, añadiendo en [22] el vector de errores e:

$$x = Ay + e \qquad [26]$$

2.2.4. *Obtención de las componentes*

Hasta el momento se ha puesto el énfasis en no perder información al reducir la dimensión a sólo dos componentes, sin embargo, no se ha explicitado cómo obtenerlas.

Si [22], se premultiplica por A',

$$A'x = A'Ay \qquad [27]$$

queda que

$$y = (A'A)^{-1}A'x \qquad [28]$$

ecuación que Horts (1965) presenta como de regresión en la obtención de las componentes, y que en el caso del ACP queda muy simplificada. En efecto, [25] permite escribir [28] como,

$$y = D^{-1}A'x = D^{-1}(VD^{1/2})'x = D^{-1/2}V'x \qquad [29]$$

que por un lado, ilustra la ecuación [2], y por otro, incorpora explícitamente la condición [10] sobre la normalización de los vectores propios de V.

Los coeficientes que finalmente proporcionarán la combinación lineal de las v.o. para obtener las componentes son:

$$D^{-1/2}V' = \begin{bmatrix} \frac{1}{1.71175} & \\ & \frac{1}{1.03343} \end{bmatrix} \begin{bmatrix} .47985 & .53110 & .44256 & .54021 \\ -.55146 & -.40220 & .63149 & .36791 \\ -.60423 & .73993 & .10469 & -.27649 \\ -.31708 & .09319 & -62802 & .70453 \end{bmatrix}$$

$$= \begin{bmatrix} .28032 & .31027 & .25854 & .31559 \\ -.53362 & -.38919 & .61106 & .35601 \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \qquad [30]$$

La ecuación [29], que sintetiza la tabla 3, es ahora:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} D^{-1/2}V' \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} .28032x_1 + .31027x_2 + .25854x_3 + .31559x_4 \\ .53362x_1 - .38919x_2 + .61106x_3 + .35601x_4 \end{bmatrix} \qquad [31]$$

Por último, al comparar este apartado con el anterior observamos cierta analogía en el proceso. Mientras allí, el producto de matrices factoriales, AA', reproduce las correlaciones entre v.o., y la multiplicación matricial, $A'A$, en cambio concierne a las correlaciones entre componentes, aquí, la ecuación [29] indica cómo obtener las componentes mediante $D^{-1/2}V'$, mientras su homónima, la [22] reproduce las v.o. según $VD^{1/2}$.

Se detallan a continuación los resultados obtenidos al utilizar el programa de ACP del package SPSS (Nie, Hull, Jenkins, Steinbrenner y Bent, 1975), Factor Analysis (PA1), con los datos de la matriz de correlaciones [1]. El sencillo conjunto de instrucciones requerido es;

TABLA 13. Programa de ACP para los datos de [1].

```
 1  FILE NAME          ACPE1CL
 2  VARIABLE LIST      MA, CN, FR, LA
 3  INPUT MEDIUM       DISK
 4  N OF CASES         8
 5  VAR LABELS         MA, MATEMATIQUES/
 6                     CN, CIENCIES NATURALS/
 7                     FR, FRANCES/
 8                     LA, LLATI
 9  FACTOR             VARIABLES = MA, CN, FR, LA/
10                     TYPE = PA1/
11                     N FACTORS = 2/
12                     ROTATE = NOROTATE
13  OPTIONS            3, 9
14  STATISTICS         ALL
```

y los resultados obtenidos fueron:

TABLA 14. Valores propios y porcentajes de varianza explicada por las componentes, calculadas a partir de la matriz [1] de correlaciones *R*.

VARIABLE	EST COMMUNALITY	FACTOR	EIGENVALUE	PCT OF VAR	CUM PCT
MA	1.00000	1	2.93008	73.3	73.3
CN	1.00000	2	1.06798	26.7	100.0
FR	1.00000	3	0.00140	0.0	100.0
LA	1.00000	4	0.00053	0.0	100.0

que exceptuando la columna de «comunalidades», objeto del próximo apartado, reflejan, en la columna EIGENVALUE los resultados de [12] y bajo el título PCT OF VAR, el ratio $\frac{\lambda_k}{p}$, al que se refería el apartado 2.2.2, como porcentaje de varianza explicado por la k-ésima componente.

La matriz factorial A, obtenida ya en [24] es;

TABLA 15. Matriz factorial de las proyecciones sobre las dos primeras componentes de las v.o., x_1, x_2, x_3, x_4.

FACTOR MATRIX USING PRINCIPAL FACTOR, NO ITERATIONS

	FACTOR 1	FACTOR 2
MA	0.82138	−0.56989
CN	0.90910	−0.41564
FR	0.75756	0.65260
LA	0.92470	0.38021

VARIABLE	COMMUNALITY
MA	0.99944
CN	0.99923
FR	0.99977
LA	0.99963

y los coeficientes de las variables estandarizadas que en [31] proporcionaban las dos primeras componentes, se especifican en la tabla 16.

TABLA 16. Coeficientes que proporcionan la combinación lineal de las v.o. en la determinación de las componentes.

FACTOR SCORE COEFFICIENTS

	FACTOR 1	FACTOR 2
MA	0.28032	−0.53362
CN	0.31027	−0.38919
FR	0.25854	0.61106
LA	0.31559	0.35601

El resultado de la extracción factorial consiste en resumir la matriz de correlaciones R, por una matriz de saturaciones A que la reproduce razonablemente (AA'). Esta matriz factorial A, puede representarse gráficamente según se dijo en 2.2.3, identificando sus elementos como las proyecciones de las v.o. sobre una serie de componentes ortogonales, pues el criterio de Hottelling, utilizado aquí —del Factor Principal[1]—, estima secuencialmente las componentes, después de eliminar la «influencia» de las obtenidas previamente, de esta forma se consiguen las componentes que explican la mayor varianza posible. No resulta difícil entender, por tanto, que la primera componente sea habitualmente de tipo general, es decir, con saturaciones positivas en todas las variables, siendo las restantes de tipo bipolar. Este hecho es el que refleja la figura de la página siguiente.

2.2.5. *Interpretación de los resultados. Rotación*

Aun en primeras aplicaciones —exploratorias— del ACP, no es habitual que los investigadores adopten posturas meramente nominalistas, sino que buscan implicaciones en el mundo real para interpretar y etiquetar las componentes obtenidas en función de las variables con las que se relacionan.

Aceptar como definitiva esta primera solución factorial [24] puede originar confusionismo. Así, por ejemplo, Ch. Spearman (1904), en el contexto del AF y a partir de unos pocos tests de rendimiento en estudiantes, llevó a cabo inferencias psicométricas sobre todo el dominio de tests del rendimiento humano, lo cual le condujo a concebir la mente organizada jerárquicamente y regida por un factor general de la inteligencia (factor G).

No obstante, debe tenerse en cuenta que hasta el momento sólo se ha establecido, en función de los valores propios λ_k, el número de factores (componentes) necesarios

[1] Este método o su aproximación geométrica, «método del centroide», que Thurstone (1935) desarrolla para evitar la resolución de la ecuación característica en la diagonalización de R, proponen a la varianza residual como entidad a minimizar. En Harmann (1976), Mulaik (1972), o Cuadras (1981) se consideran, entre otros, criterios específicos del modelo de AF (MINRES), junto con procedimientos basados en la inferencia estadística; en particular, debe tenerse en cuenta el método basado en la estimación máximo verosímil, a la que D. N. Lawley (1940) redujo el problema de la extracción factorial.

para describir aceptablemente las v.o. Este criterio de maximizar la varianza explicada (λ_k) por los sucesivos factores, redunda en la obtención de componentes altamente complejas que ponderan en todas las variables, como sucede al primer factor de nuestro ejemplo (véase fig. 1).

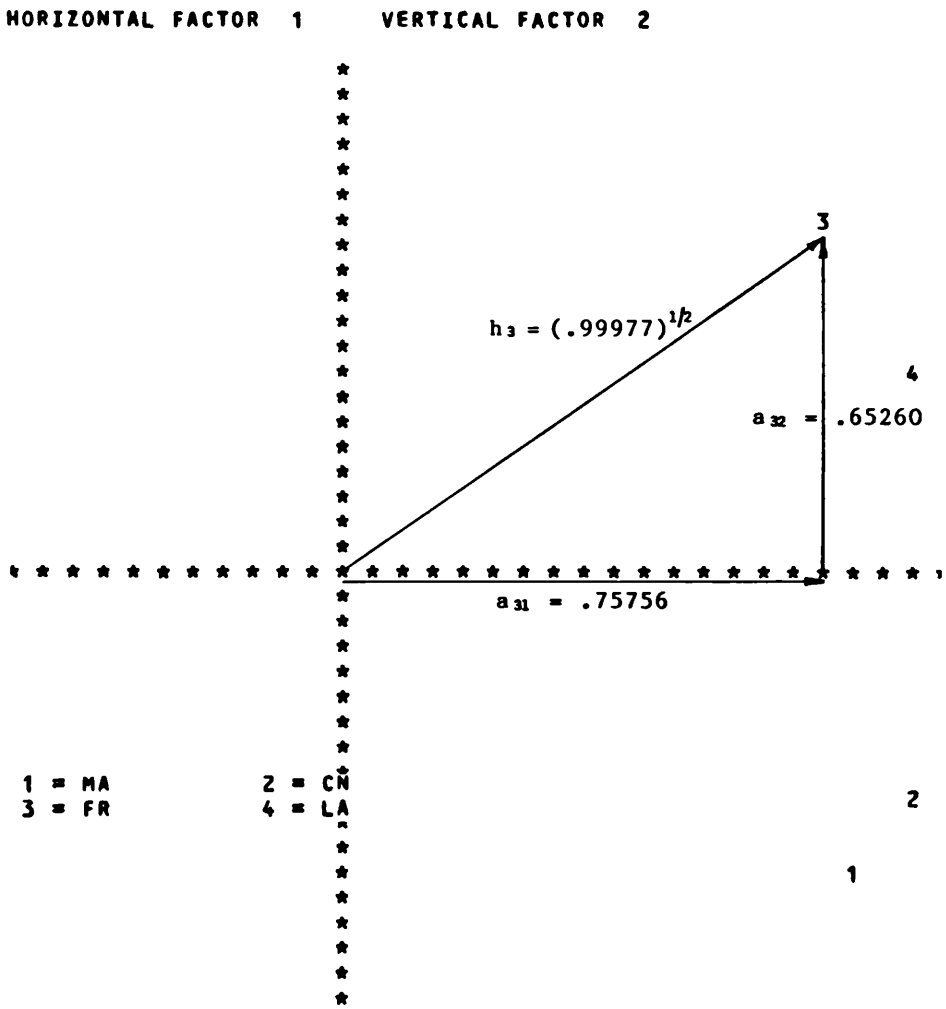


FIGURA 1. Representación de las v.o. en el espacio definido por las dos primeras componentes. Proyecciones de una variable (Francés) sobre ambos ejes.

Afortunadamente cuando el número de componentes es mayor que la unidad, el conjunto de saturaciones, A, no es único, pudiéndose transformar ortogonalmente (T)

en otros conjuntos equivalentes, $A_1 = AT$, en el sentido de que ambos explican igualmente las intercorrelaciones entre las variables originales; en efecto:

$$R = AA' = AIA' = (AT)(T'A') = A_1A_1' \qquad [32]$$

$$\text{siendo } TT' = I \qquad [33]$$

Este hecho se utiliza con frecuencia para simplificar los primeros resultados del análisis y hacerlos más interpretables.

Los pioneros de las aplicaciones del *AF* determinaban la matriz T que proporcionase la rotación apropiada, inspeccionando los $\binom{m}{2}$ gráficos de las variables en el espacio determinado por cada par de factores (fig. 1). En la actualidad la implementación en programas de ordenador de métodos analíticos que proporcionan, según distintos criterios (Harman, 1976), rotaciones objetivas de los ejes factoriales extraídos inicialmente, han relegado aquellos subjetivos y laboriosos métodos manuales.

A continuación se detallan los resultados obtenidos en nuestro caso, al rotar la matriz A obtenida en [24] según el criterio VARIMAX de Kaiser (1958: 187-200); éste, en esencia, trata de situar las componentes de forma que cada una tenga grandes saturaciones en pocas variables siendo el resto de reducido tamaño. Este método, a pesar de ser «psicológicamente ciego» ha sido ampliamente utilizado en este campo, debido a que evita la solución de un factor general proporcionando óptimos resultados con número de componentes reducido.

TABLA 17. Solución final, matriz de saturaciones rotada (A_1).

VARIMAX ROTATED FACTOR MATRIX

	FACTOR 1	FACTOR 2
MA	0.98900	0.14599
CN	0.94751	0.31852
FR	0.10635	0.99422
LA	0.41458	0.90981

TABLA 18. Matriz de rotación ortogonal, T, permite el giro de la solución inicial (tabla 15) a la final.

TRANSFORMATION MATRIX

	FACTOR 1	FACTOR 2
FACTOR 1	0.72955	0.68392
FACTOR 2	−0.68392	0.72955

TABLA 19. Coeficientes de las v.o. en el cálculo de las componentes giradas.

FACTOR SCORE COEFFICIENTS

	FACTOR 1	FACTOR 2
MA	0.56947	−0.19758
CN	0.49253	−0.07173
FR	−0.22930	0.62262
LA	−0.01325	0.47557

El lector puede comprobar estos resultados: girando los puntos de la matriz de saturaciones de la tabla 15; rotando los ejes de la tabla 16; o verificando la igualdad [32].

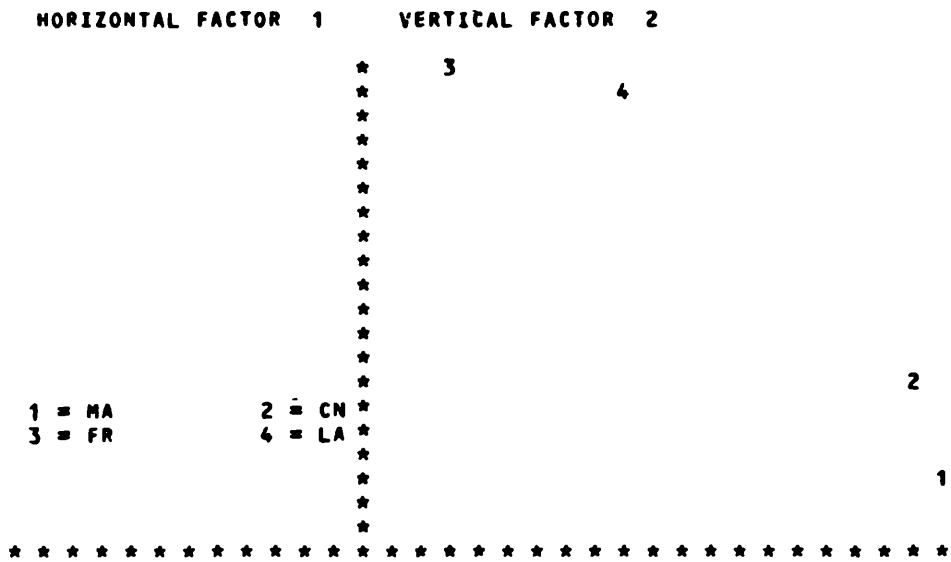


FIGURA 2. Representación de las v.o. en el espacio de las componentes giradas.

La interpretación de los resultados, en algunos casos, puede hacerse sin necesidad de rotar. Así, el hecho de que todos los coeficientes de correlación en la ecuación [1] sean positivos indica que los estudiantes cuyas puntuaciones están por encima de la media en una asignatura tienden a estarlo también en las demás. Esta relativa facilidad en la adquisición de conocimiento se traduce en un primer factor con saturaciones positivas en todas las variables que puede etiquetarse como de «inteligencia general», siendo el factor bipolar el que sirve para discriminar los dos grupos de variables en asignaturas que indican habilidad verbal (FR, LA) y otras (MA, CN) que reflejan capacidad lógico-formal.

El ejemplo de REUCHELIN que se ha seguido para ilustrar el ACP, es en realidad la aplicación típica del modelo de AF, por ello, se ha utilizado para introducir conceptos propios del AF, que será objeto del próximo apartado. No obstante, el autor es reticente a la habitual interpretación de los ejes de la figura 2 en «Ciencias» y «Letras», primero, porque las puntuaciones en las que se basa el estudio conciernen sólo a ocho estudiantes, y en segundo lugar, la dicotomía entre «ciencia» y lo que «no es ciencia», requeriría un profundo estudio que escapa a la información que las cuatro v.o. pueden proporcionar.

Por último, en el capítulo cuarto, al tratar las aplicaciones confirmatorias y el concepto de identificación, se aclara que la palabra rotación en realidad es sólo un eufemismo que esconde una indeterminación en la solución, inherente al modelo de Análisis Factorial.

2.3. El Análisis Factorial

2.3.1. *Introducción histórica*

En 1869 Galton, en su libro *Genio Hereditario,* afirma que, por analogía a otras características físicas, la inteligencia debería tener un valor medio constante para el conjunto de habitantes en Gran Bretaña, y las desviaciones con respecto a este valor promedio debían regirse según la ley Normal. Posteriormente esta idea, considerablemente más elaborada por Pearson y Spencer, se traduce en considerar la capacidad mental normalmente distribuida en un continuo, debido a la supuesta existencia de un factor general de tipo cognitivo y a las diferencias en la constitución genética.

En base a estas ideas, mencionadas ya en el apartado 2.2.5, Spearman propone el primer modelo de AF, conocido con el nombre de «modelo de los dos factores» —el general y el único—. En efecto, sobre 27 alumnos midió un conjunto de variables de tipo intelectual (rendimiento escolar, razonamiento matemático, etc.), que según él debían ser mediciones de un factor de tipo cognitivo común a todas ellas que explicaría las intercorrelaciones existentes, y otro factor específico de cada medida que sería el responsable de lo residual. El solapamiento entre las variables utilizadas corroboraba, para Spearman, la evidencia de un factor general de orden superior o integrativo de las funciones intelectuales humanas. Este modelo que incluye un factor común (o varios), puede generalizarse para explicar las correlaciones (o grupos de intercorrelaciones) de cualquier otro conjunto (o subconjuntos) de variables.

2.3.2. *El modelo de Análisis Factorial*

Dado un conjunto de p características medibles sobre N individuos, el objetivo del AF es la obtención e interpretación de un conjunto más reducido (m) de factores latentes (no observables) que expliquen la covariación existente entre dichas p v.o.

El modelo supone que toda variable (x_i) se compone de dos partes: una común con otras v.o. (C_i), que puede ser expresada por ello en términos de factores comunes;

y otra única (U_i) que incluye tanto la especificidad propia de la variable como el llamado «error de medida». Así, cada variable observable se descompone en:

$$x_i = C_1' + U_i \qquad [34]$$

y la puntuación del j-ésimo individuo en esta v.o.

$$x_{ij} = C_{ij} + U_{ij} = \Sigma a_{ik} f_{kj} + d_{ij} \qquad [35]$$

que para los N individuos es:

$$X = AF + D \qquad [36]$$

siendo $X(pxN)$, $A(pxm)$, $F(mxN)$ y $D(pxN)$.

Si se consideran ahora las p v.o. recogidas en el vector aleatorio $x(px1)$, la ecuación [35] puede generalizarse como

$$x = Af + \delta \qquad [37]$$

siendo $f(mx1)$ y $\delta(px1)$ los vectores de los factores comunes y únicos, respectivamente.

Para que la «partición» efectuada en [37] resulte correcta deberían cumplirse algunos supuestos sobre las variables que allí se relacionan:

1. Incorrelación entre ambas fuentes de variación.

$$E(f\delta') = 0 \qquad [38]$$

2. Incorrelación: 1.° entre errores de medida asociados a distintas v.o.; 2.° interespecificidades, es decir, inclusión de todos los factores comunes a las p v.o.; 3.° entre el error de medida y la especificidad.

$$E(\delta\delta') = \theta \text{ diagonal} \qquad [39]$$

2.3.3. *Descomposición de la matriz de varianzas-covarianzas. Ecuación fundamental del Análisis Factorial*

Diferenciar lo que comparten de lo que no es común en el conjunto de p v.o., representa la diferencia conceptual que existe entre la ecuación [37] del modelo, definida por tanto a priori, y la [26] obtenida como resultado de aplicar la técnica de ACP a un conjunto de v.o. con el objetivo de reducir su dimensión. Sin embargo, es la descomposición de la matriz de correlaciones (varianzas-covarianzas, en general), R, de las v.o. que el modelo de AF (ecuaciones [37], [38] y [39]) comporta, lo que en la práctica distingue el AF del ACP; en efecto,

$$R = E(xx') = E(Af + \delta)(Af + \delta)' = E(Aff'A') + E(\delta\delta') = A\Phi A' + \theta \qquad [40]$$

siendo $\Phi = E(ff')$, la matriz de correlaciones de los factores comunes[2].

La ecuación [40], que estructura la matriz de correlaciones R en función de los parámetros del modelo, A, Φ y θ, difiere de [20], exclusivamente, en la partición que induce en los elementos de la diagonal (varianzas); así según [40], la variancia, σ_i^2, de cada v.o. es

$$1 = \sigma_i^2 = \sum_k a_{ik}^2 + \theta_i \qquad [41]$$

En el segundo miembro de la ecuación [41], $\sum_k a_{ik}^2$ establece la varianza de la v.o. que explican los factores comunes en conjunto, y se llama comunalidad, h_i^2; el término complementario, θ_i, se conoce por unicidad de la variable.

Los elementos exteriores a la diagonal principal de AA' no se alteran de la ecuación [20] a la [40], pues expresan lo que comparten cada par de variables, x_i, x_j, en cada una de las direcciones ortogonales, k, es decir, su correlación

$$\sum_k a_{ik}a_{jk} \cong r_{ij} \qquad [44]$$

2.3.4. *Extracción factorial*

El método del Factor Principal es sin duda el que más se ha utilizado para obtener los factores; consiste esencialmente en la aplicación directa del mencionado algoritmo de Hotelling (1933: 417-441) a la matriz de correlaciones «reducida», R^0, obtenida a partir de R por sustitución de las unidades de la diagonal principal con comunalidades estimadas, que usualmente son los cuadrados de los coeficientes de correlación múltiple de la v.o. con las demás.

$$R^o = R - \theta = AA' \qquad [45]$$

La utilización del mismo algoritmo de resolución conduce en algunos casos a erróneas conclusiones sobre la equivalencia de la técnica de ACP y el modelo de AF. En nuestro caso, los resultados obtenidos al considerar el modelo de AF para explicar las intercorrelaciones entre las v.o. del ejemplo, no difieren mucho de los que se obtuvieron en el apartado anterior.

---

[2] La literatura del AF distingue habitualmente entre el modelo de factores ortogonales u oblicuos, según la matriz Φ sea la identidad o no. En el desarrollo posterior se considera por simplicidad $\Phi = I$, pues, únicamente en este caso la relación entre v.o. y factores está perfectamente determinada por la matriz factorial A. En efecto, si se postmultiplica [37] por

$$xf' = Aff' + \delta f' \qquad [42]$$

y por tanto

$$E(xf') = AE(ff') + E(\delta f') = A\Phi \qquad [43]$$

en cualquier otro caso, $\Phi \neq I$, la ecuación [43] recibe el nombre de estructura factorial.

TABLA 20. (a) Comunalidades estimadas, para sustituir en AF a los elementos de la diagonal principal de R. (b) Valores propios y porcentajes de varianza explicada correspondientes a los factores que el AF establece.

VARIABLE	EST COMMUNALITY	EIGENVALUE	PCT OF VAR	CUM PCT
MA	0.99777	2.93008	73.3	70.
CN	0.99755	1.06798	26.7	100.
FR	0.99866	0.00140	0.0	100.
LA	0.99898	0.00053	0.0	100.
	(a)	(b)		
CONV	CONVERGENCE REQUIRED	1 ITERATIONS[3]		

Esta identidad entre los valores propios de la tabla 20 y los obtenidos por ACP redunda en la igualdad de matrices factoriales; esto es debido a que las comunalidades estimadas toman valores muy próximos a la unidad. Las diferencias, aunque no importantes, radican en las distintas combinaciones lineales de las v.o. que proporcionan las componentes en la tabla 16, y los factores en la tabla 21.

TABLA 21. Coeficientes de la combinación lineal de las v.o. en la composición de los dos factores comunes que el AF de la matriz [1] establece.

FACTOR SCORE COEFFICIENTS

	FACTOR 1	FACTOR 2
MA	0.26733	−0.51785
CN	0.25991	−0.39208
FR	0.10925	0.64417
LA	0.49866	0.31754

2.3.5. *Un ejemplo ilustrativo: ACP versus AF*

Con la finalidad de diferenciar, también a efectos prácticos, la técnica de ACP y el modelo de AF, se han tomado los datos de un ejemplo que Lawley y Maxwell (1971) presentan para ilustrar el enfoque confirmatorio del AF.

La tabla adjunta describe la matriz de correlaciones de seis asignaturas escolares: Francés (FR); Inglés (AN); Historia (HS); Aritmética (AR); Algebra (AL) y Geometría (GM), que estos autores consideraron en una muestra de 220 alumnos.

[3] Ya que generalmente, tanto las comunalidades como el número de factores a extraer son desconocidos, los distintos procedimientos de extracción factorial (véase Harman, 1976: cap. 5; Cuadras, 1981: cap.3) requieren que al menos uno de estos valores se establezca a priori, y determinan el otro por procedimientos iterativos.

TABLA 22. Matriz de correlaciones de las seis asignaturas consideradas en Lawley y Maxwell (1971).

	FR	AN	HS	AR	AL	GM
FR	1.00000	0.43900	0.41000	0.28800	0.32900	0.24800
AN	0.43900	1.00000	0.35100	0.35400	0.32000	0.32900
HS	0.41000	0.35100	1.00000	0.16400	0.19000	0.18100
AR	0.28800	0.35400	0.16400	1.00000	0.59500	0.47000
AL	0.32900	0.32000	0.19000	0.59500	1.00000	0.46400
GM	0.24800	0.32900	0.18100	0.47000	0.46400	1.00000

En la tabla 22 de partida pueden distinguirse dos agrupaciones entre las seis v.o. (FR, AN, HS y AR, AL GM), los cual parece indicar la existencia de dos constructos subyacentes que explicarán estas elevadas correlaciones, y se traducirán en componentes o factores en los dos análisis que a continuación se presentan. Si bien éstos no proporcionarán patrones de relación entre las v.o. esencialmente distintos, los resultados que se obtengan en el ACP deberán estar «hinchados» con respecto a los obtenidos según el modelo de AF. En efecto, en el ACP las «comunalidades» de las v.o. se determinaban una vez extraídas las componentes; por el contrario, los factores en el AF se establecen en función de las estimaciones de las comunalidades efectuadas como prólogo del Análisis.

TABLA 23. (a) Diagonal principal de la matriz *R*, punto de partida del ACP. (b) Valor propio y porcentaje de varianza asociado (PCT) a cada componente. (c) Varianza o comunalidad de cada variable, explicada por el conjunto de las 2 componentes establecidas en el análisis previo.

VARIABLE	EST COMMUNALITY	EIGENVALUE	PCT OF VAR	CUM PCT
FR	1.00000	2.73289	45.5	45.5
AN	1.00000	1.12977	18.8	64.4
HS	1.00000	0.61517	10.3	74.6
AR	1.00000	0.60122	10.0	84.7
AL	1.00000	0.52480	8.7	93.4
GM	1.00000	0.39615	6.6	100.0

(a) (b)

VARIABLE	COMMUNALITY
FR	0.63437
AN	0.55825
HS	0.67387
AR	0.71570
AL	0.69431
GM	0.58614

(c)

La tabla anterior describe la parte de varianza de las v.o. que se proyecta sobre las componentes. Sin embargo, si previamente suponemos que las variables tienen cierta estructura [34] subyacente, la matriz de correlaciones, R, que se analice, deberá reflejar, en el sentido de [40], este hecho, cuyas consecuencias se detallan a continuación:

TABLA 24. Cuadrados de los coeficientes de correlación múltiple de cada variable con las restantes, que son las estimaciones iniciales de las comunalidades incluidas en la diagonal principal de R^*, punto de partida del AF. (b) Comunalidad de cada v.o. explicada por los dos factores comunes considerados. (c) Valores propios y porcentajes de la comunalidad total asociados a cada factor.

VARIABLE	EST COMMUNALITY
FR	0.30010
AN	0.29659
HS	0.20610
AR	0.41970
AL	0.41775
GM	0.29524

CONVERGENCE REQUIRED 11 ITERATIONS

(a)

VARIABLE	COMMUNALITY	FACTOR	EIGENVALUE	PCT OF VAR	CUM PCT
FR	0.48663	1	2.22214	79.0	79.0
AN	0.40868	2	0.59135	21.0	100.0
HS	0.35649				
AR	0.61783				
AL	0.56888				
GM	0.37497				
	(b)		(c)		

TABLA 25. Matrices factoriales que resultan de los ACP (a), y AF (b), efectuados sobre los datos de la tabla 22.

FACTOR MATRIX USING PRINCIPAL FACTOR, NO ITERATIONS

	FACTOR 1	FACTOR 2
FR	0.65782	0.44905
AN	0.68842	0.29039
HS	0.51737	0.63734
AR	0.73831	−0.41303
AL	0.74388	−0.37545
GM	0.67831	−0.35501

(a)

WITH ITERATIONS

	FACTOR 1	FACTOR 2
FR	0.58650	0.37769
AN	0.59411	0.23603
HS	0.43134	0.41284
AR	0.71135	−0.33438
AL	0.70142	−0.27730
GM	0.58406	−0.18397

(b)

La tabla 11 ilustraba cómo podían deducirse la varianza extraída por las componentes y la comunalidad de cada variable. Análogamente pueden derivarse estos resultados de las matrices factoriales que en la tabla 25 corresponden al ACP y al AF. En efecto, sobre la primera componente se proyecta

$$\lambda_1 = \sum a_{i1}^2 = (.65782)^2 + \ldots + (.67831)^2 = 2.73289 \qquad [46]$$

$$\text{siendo } \sum \lambda_k = p = 6 = Tr(R) \qquad [47]$$

y sobre el primer factor

$$\lambda_1 = \sum a_{i1}^2 = (.58650)^2 + \ldots + (.58406)^2 = 2.22214 \qquad [48]$$

$$\text{siendo } \sum_k \lambda_k = \sum_i h_i^2 = (.48663)^2 + \ldots + (.37497)^2 = Tr(R) \qquad [49]$$

la varianza de la primera variable, que las dos primeras componentes explican, es

$$\sum_k a_{1k}^2 = (.65782)^2 + (.44905)^2 = .63437 \qquad [50]$$

mientras que la comunalidad de dicha v.o., explicada en el espacio de los factores comunes, sería

$$\sum_k a_{1k}^2 = (.58650)^2 + (.37769)^2 = .48663 \qquad [51]$$

A diferencia del ejemplo anterior, se distingue aquí perfectamente la importancia relativa de la primera componente, $\frac{\lambda_1}{p} = .455$, y la del primer factor, $\frac{\lambda_1}{\sum h_i^2} = .79$.

En ambos casos la aparición de un primer factor común a todas las v.o. y el segundo bipolar, sugieren la necesidad de rotar.

El solapamiento entre las v.o. que la matriz de correlaciones (tabla 22) revela, no hace verosímiles los elevados valores de la tabla 25 (a). El modelo de AF debería evaluarse, por consiguiente, con distinto criterio del proporcionado exclusivamente por: el tamaño de los valores propios; o la comunalidad explicada por los factores extraídos; o la magnitud de las saturaciones factoriales. Resultaría más apropiado como índice de la calidad del modelo, uno que valora el desajuste entre la matriz de correlaciones reducida y la reproducida a partir de las ponderaciones factoriales. En nuestro caso esta última es:

$$AA' = \begin{bmatrix} .4866 & & & & & \\ .4376 & .4087 & & & & \\ .4089 & .3537 & .3565 & & & \\ .2909 & .3437 & .1688 & .6178 & & \\ .3066 & .3513 & .1881 & .5917 & .5689 & \\ .2731 & .3036 & .1760 & .4770 & .4607 & .3750 \end{bmatrix} \qquad [52]$$

TABLA 26. Resultados del ACP (a, c, e), frente a los obtenidos mediante AF (b, d, f).

VARIMAX ROTATED FACTOR MATRIX

	FACTOR 1	FACTOR 2
FR	0.22531	0.76394
AN	0.34939	0.66044
HS	−0.00259	0.82089
AR	0.83308	0.14727
AL	0.81360	0.17989
GM	0.64990	0.15424

(a)

VARIMAX ROTATED FACTOR MATRIX

	FACTOR 1	FACTOR 2
FR	0.23006	0.65856
AN	0.32321	0.55156
HS	0.08611	0.59083
AR	0.76653	0.17396
AL	0.72358	0.21286
GM	0.57365	0.21424

(b)

TRANSFORMATION MATRIX

	FACTOR 1	FACTOR 2
FACTOR 1	0.77441	0.63269
FACTOR 2	−0.63269	0.77441

(c)

TRANSFORMATION MATRIX

	FACTOR 1	FACTOR 2
FACTOR 1	0.78840	0.61516
FACTOR 2	−0.61516	0.78840

(d)

FACTOR SCORE COEFFICIENTS

	FACTOR 1	FACTOR 2
FR	−0.06507	0.46010
AN	0.03245	0.35843
HS	−0.21032	0.55664
AR	0.44052	−0.11219
AL	0.42105	−0.08514
GM	0.39102	−0.08631

(e)

FACTOR SCORE COEFFICIENTS

	FACTOR 1	FACTOR 2
FR	−0.05482	0.43170
AN	0.02457	0.28433
HS	−0.08482	0.33337
AR	0.46889	−0.09037
AL	0.37885	−0.03538
GM	0.19835	0.01219

(f)

siendo A la matriz factorial estimada en la tabla 25 (b) o 26 (b). Los residuos que se derivan del modelo de dos factores propuesto, son

$$R^{o} - AA' = \begin{bmatrix} .0000 & & & & & \\ .0014 & -.0001 & & & & \\ .0011 & -.0027 & .0000 & & & \\ .0029 & .0103 & -.0048 & .0000 & & \\ .0224 & -.0313 & -.0019 & .0033 & .0000 & \\ .0249 & .0264 & .0050 & .0070 & .0033 & .0000 \end{bmatrix} \quad [53]$$

La representación geométrica de los resulados obtenidos en la tabla 26 (b), es

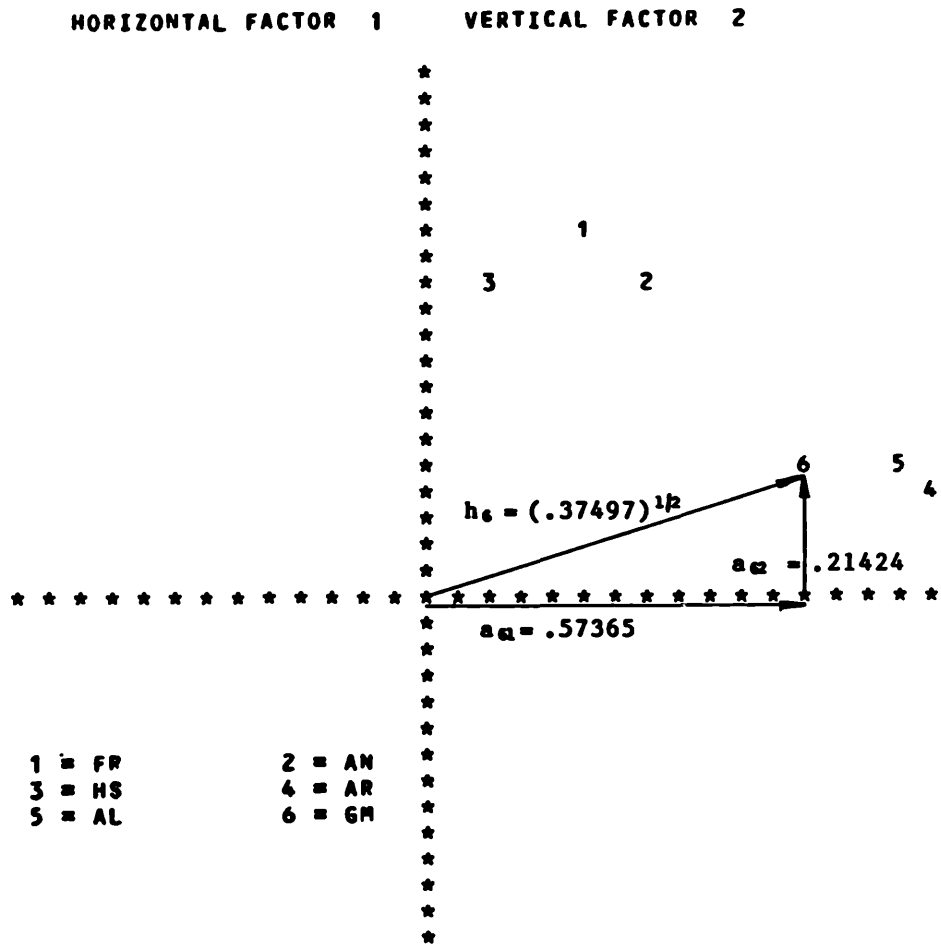


FIGURA 3. Representación de las v.o. en el espacio de los factores comunes. Descomposición de la comunalidad de la variable «Geometría».

2.4. Utilización de modelos para confirmar teorías

2.4.1. *Introducción metodológica a la utilización de modelos estadísticos*

Todo trabajo científico debe enfrentarse con el problema de la inferencia de entidades o procesos que no puede observar directamente a partir de los datos disponibles. Habitualmente la inferencia científica se lleva a cabo de acuerdo con el paradigma hipotético-deductivo, en el que:

i) Se supone un modelo para «estructurar» lo no observable.
ii) Se deducen consecuencias observables para el modelo propuesto.

iii) Se realiza una investigación empírica con el objeto de demostrar que las consecuencias esperadas en las observaciones, son las que realmente aparecen en los datos.

Este proceso ha sido tácitamente observado en el apartado anterior. En efecto,

i) Se asume la existencia de dos factores: uno de habilidad verbal, y otro de habilidad matemática. Estas hipótesis verbales pueden trasladarse a hipótesis causales como se hace en el diagrama de la fig. 5, ó en las especificaciones del modelo [54 ÷ 57] equivalente.
ii) Ya que el modelo es identificable (véase apart. 2.4.3) pueden estimarse los parámetros del mismo. Estas estimaciones nos permitirán reproducir en el sentido de [45], las observaciones, correlaciones en nuestro caso, como más adelante detallaremos.
iii) Por otro lado, en el contexto adecuado, se han recogido (tabla 22) datos empíricos relevantes, cuya contrastación con los que se derivan en la etapa anterior (ii) permite la verificación del modelo supuesto.

Así pues, toda inferencia, y en particular la causal, supone confirmar o rechazar hipótesis según los datos recogidos (véase el esquema de la fig. 4, inspirado en el de Saris y Stronkhorts, 1981). La interpretación de los datos requiere un conjunto limitado de asunciones sobre la generación de los mismos. Al mínimo número de supuestos que es capaz, como instrumento metodológico, de estructurar los datos según una cierta teoría, es lo que entendemos como modelo, y en la utilidad de éste (véase figura adjunta) radica el único índice para su valoración.

2.4.2. *El Análisis Factorial Confirmatorio*

Si con la pretensión de «recubrir» cierto dominio, del que prácticamente se carece de conocimiento teórico previo, se selecciona una muestra representativa de variables para someterla a Análisis Factorial, se dice que la técnica se aplica con carácter puramente exploratorio. Esta utilización del modelo de AF es a la que se ha referido el apartado anterior; su finalidad está en descubrir el número de factores o dimensiones necesarias para explicar las interrelaciones entre el conjunto de variables medidas. Esta perspectiva exploratoria del AF que se inició en Spearman y que continúa con Thurstone, culmina con la obra de Harman (1976).

Realmente, este Análisis Factorial no representa más que una transformación matemática de la información contenida en la matriz de correlaciones a unas ecuaciones que pueden (o no) resultar más interpretables que esta matriz. No obstante, al desconocerse el proceso que produce la covariación, la interpretación es difícil, y en general los factores no representan más que tautológicas reformulaciones de los nombres de las variables originales.

En otras situaciones en las que el investigador ya tiene información sobre las variables y sus interrelaciones, podrá permitirse formular hipótesis (quasi) definitivas acerca de la naturaleza de los constructos subyacentes; es decir, podrá establecer un modelo de la estructura causal de estas variables, en el que pueden haberse especifica-

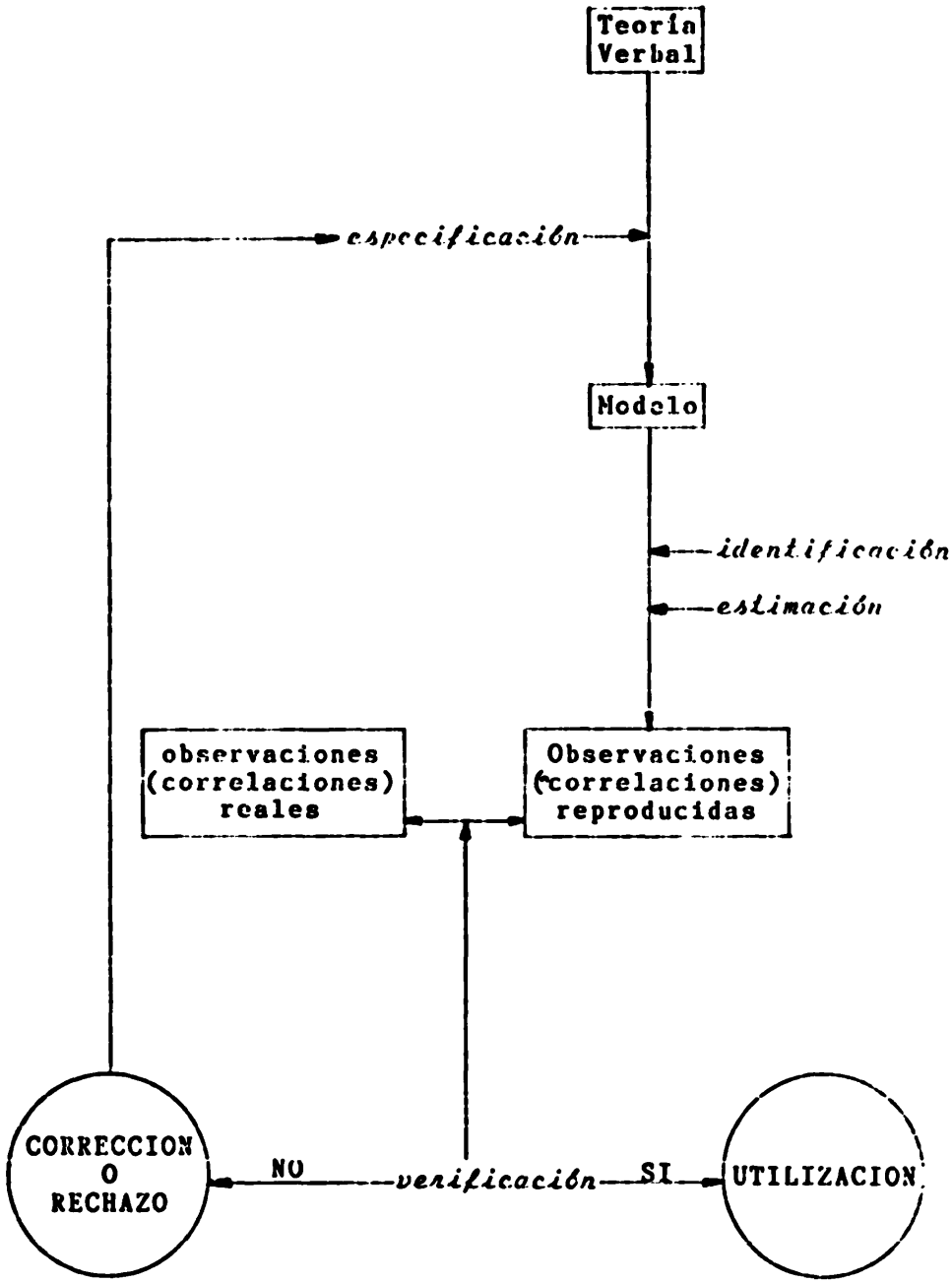


FIGURA 4. Esquema del proceso de modelado estadístico sin diseño.

do o restringido los valores de algunos parámetros. Se entiende en estos casos que el modelo de AF ha sido aplicado en su vertiente confirmatoria (AFC).

Ya que no existe una fórmula para hacer ciencia, el progreso de ésta, en cualquier dominio, requiere que la investigación le otorgue una naturaleza acumulativa, aprendiendo de los errores de los que precedieron para evolucionar desde etapas exploratorias a otras de confirmación de resultados.

Durante muchos años las ciencias sociales, biológicas, y de la conducta se han servido de «modelos estadísticos» para recorrer esta vía inductiva hacia la generación de teorías. Si bien llegar a ordenar las ideas en forma de modelos puede estimular el curso de éstas, los modelos no pueden ser mejores que las ideas que se quieren expresar mediante ellos. Debe perderse la ingenua esperanza de adquirir una técnica tal, que al ser aplicada mecánicamente a un conjunto de datos, proporcione investigación científica de forma automática.

Duncan (1975) dice: «...así el anhelo de psicología instantánea, la superstición de que ésta pueda lograrse por la mera complicación o perfeccionamiento del aparato formal (entiéndase métodos estadísticos) y el instinto de suponer que cualquier conjunto de datos torturados de acuerdo con el ritual prescrito producirá interesantes descubrimientos científicos, son hábitos patológicos del pensamiento que se desarrollan en la falacia de la inducción».

Bentler (1980: 419-456), resume los párrafos anteriores en una frase: «entender, es lo que se requiere para modelar».

Estas consideraciones no pretenden invalidar el uso de modelos estadísticos en etapas exploratorias, por el contrario, al denunciar la errónea utilización que en ocasiones se ha hecho de ellos, se quiere ensalzar su carácter orientativo y/o descriptivo. En cuanto al poder explicativo de los modelos, la adecuada interacción entre teoría y datos que el análisis de las causas supone, implica tanto exploración como confirmación.

Lo deseable desde la ciencia, así como para la estadística, sería que toda hipótesis sugerida sólo por procedimientos exploratorios fuera posteriormente contrastada con nuevos datos a los que debería someterse a controles estadísticos más rigurosos (en Jöreskog y Sörbom [1978] se ilustra en parte la afirmación anterior con un ejemplo, en el que la muestra original se divide en dos mitades, utilizando una para generar hipótesis y la otra para validarlas), con la conciencia de que la confirmación de teorías debe pasar necesariamente por filtros externos a la estadística, propios del campo específico de que se trate. Pues aun errando en las técnicas de análisis, las conclusiones teóricas podrían ser aceptables. Por el contrario, el olvido de una variable relevante o la invalidez en la medida de un concepto clave, aun asegurando que el posterior tratamiento estadístico se llevara a cabo de manera impecable, conducirá inevitablemente a pésimos resultados teóricos.

2.4.3. *El modelo de Análisis Factorial con restricciones*

Consideremos de nuevo las seis v.o. del ejemplo anterior, y supongamos que a través de éstas se quiere confirmar la existencia de dos factores: uno que reflejaría la habilidad verbal de los alumnos, y otro, su habilidad matemática. Se asume además: que estas habilidades posiblemente estarán intercorrelacionadas: que los factores úni-

cos están incorrelacionados entre sí y con los factores comunes. Todo ello puede trasladarse a un diagrama causal como el de la fig. 5; en él se ha utilizado la notación consuetudinaria al planteamiento LISREL, programa que más adelante proporciona las estimaciones de los parámetros del modelo.

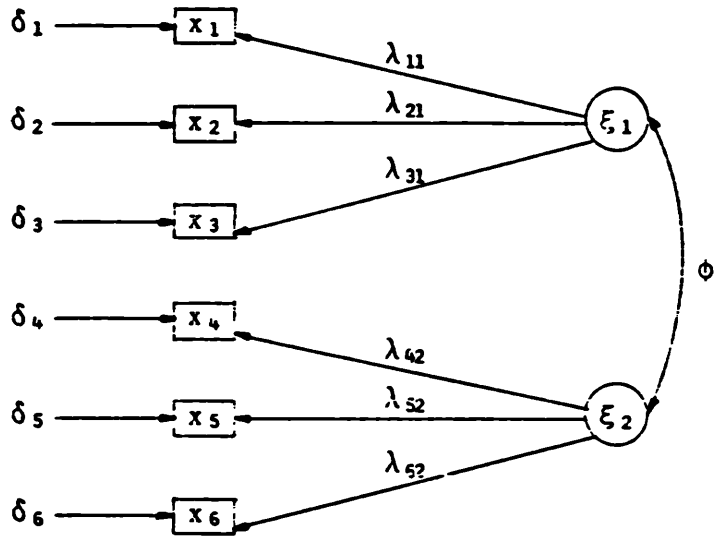


FIGURA 5. Diagrama causal del modelo de AF, supuesto generador de las correlaciones descritas en la Tabla 11.

El diagrama de la fig. 5, puede traducirse en un modelo de AF en el que se han introducido algunas restricciones, concretamente, se suponen nulas algunas saturaciones,

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \\ \delta_6 \end{bmatrix} \qquad [54]$$

$$\underline{x} \qquad \underline{\underline{\Lambda}} \qquad \underline{\xi} \qquad \underline{\delta}$$

$$E(\xi\delta') = 0 \qquad [55]$$

$$E(\delta\delta') = \theta = \text{diagonal} \qquad [56]$$

$$E(\xi_1\xi_2') = \phi = \begin{bmatrix} 1 \\ \phi_{21} 1 \end{bmatrix} \qquad [57]$$

Los parámetros del modelo se estiman mediante el programa LISREL IV (Jöreskog y Sörbom, 1978). En esencia el proceso de estimación consiste en ajustar la matriz de correlaciones reproducida, $\hat{R}$, por los parámetros del modelo, a la de correlaciones observadas, R, según el criterio de la máxima verosimilitud.

Consideremos a título ilustrativo la correlación entre los vectores muestrales x_1 y x_4. Este valor puede efectivamente establecerse en función de los elementos de las matrices paramétricas Λ, Φ y θ incluidas todas ellas en el modelo (54; 55; 56; 57), así

$$\begin{aligned} \hat{r}_{14} &= E(x_1 x_4') = E((\lambda_{11}\xi_1 + \delta_1)(\lambda_{42}\xi_2 + \delta_2)') = \\ &= E((\lambda_{11}\xi_1\xi_2'\lambda_{42}') + (\lambda_{11}\xi_1\delta_2') + (\delta_1\xi_2'\lambda_{42}') + \delta_1\delta_2')) \end{aligned} \quad [58]$$

expresión que se simplifica, al aplicar las propiedades del operador Esperanza Matemática junto con los supuestos [55; 56; 57] del modelo especificado

$$\hat{r}_{14} = \lambda_{11}\phi_{12}\lambda_{42} + 0 + 0 + \theta_{14} = \lambda_{11}\lambda_{42}\phi_{12} \quad [59]$$

Disponer de $\frac{6 \times 7}{2} = 21$, correlaciones observadas, r_{ij}, permite plantear otras tantas ecuaciones análogas a [59] cada una de las cuales expresaría las correlaciones entre v.o. en función de los parámetros del modelo; así, si recogemos estos parámetros en un vectos genérico π que los resuma, puede plantearse la ecuación general

$$\hat{r}_{ij} = f(\pi) \quad [60]$$

y ya que únicamente debemos estimar 13 parámetros (seis λ_{ik}, un ϕ_{12}, y seis θ_i), precisaremos al menos de un número igual de ecuaciones [60]. Las $\nu = 8$ ecuaciones restantes, se conocen como número de grados de libertad. Este valor, ν, categoriza a los modelos en sobredeterminados *(overidentified)*, $\nu > 0$; determinados *(exactily identified)*, $\nu = 0$; e indeterminados *(underidentified)*, $\nu < 0$. Tres categorías que no son simétricas, pues mientras la analogía con la resolución de los sistemas algebraicos de ecuaciones puede ayudar a entender que en el último caso no existe información suficiente para estimar los parámetros, que sí podrían estimarse en cambio en el segundo, no aclara que en los modelos sobredeterminados, a diferencia de lo que ocurre en los sistemas algebraicos, no sólo es posible la estimación de los parámetros, sino que además, como se verá, las relaciones excedentes permiten comprobar la significación del modelo, lo cual resulta inviable en cualquier otro caso, $\nu \leq 0$, pues si el modelo fuese correcto estas ν ecuaciones deberían ser redundantes, es decir, compatibles con las estimaciones obtenidas a partir de las primeras.

Una vez resuelto el problema de la identificación surge el de la estimación de los parámetros identificables. En el desarrollo anterior se ha ignorado por simplicidad, al esbozar la identificación del modelo, la diferenciación existente entre la matriz de correlaciones Σ de la población y la muestral R. No obstante, si ahora consideramos esta distinción debemos enfrentarnos con el problema de la inferencia estadística.

Si suponemos a las v.o. con distribución multivariable normal, la población está perfectamente caracterizada por los momentos de primer y segundo orden. Ya que en

general no se imponen restricciones al vector de Esperanzas Matemáticas, su estimador máximo verosímil coincidirá con el vector de medias muestrales. Todo ello reduce la estimación al ajuste de la matriz de correlaciones observadas, R, por la reproducida, $\hat{R}$, desde el modelo, según el procedimiento de estimación de la máxima verosimilitud con restricciones (Jöreskog, 1978; 1969: 183-202; 1977). Los resultados obtenidos en nuestro caso son

TABLA 27. Estimaciones máximo verosímiles de los parámetros del modelo de la fig. 5.

AFC1 (HVM)
LISREL ESTIMATES

LAMBDA Y

	ETA 1	ETA 2
FR	0.687	0.000
AN	0.672	0.000
HS	0.533	0.000
AR	0.000	0.767
AL	0.000	0.768
GM	0.000	0.616

(a)

PSI

	EQ. 1	EQ. 2
EQ. 1	1.000	
EQ. 2	0.597	1.000

(b)

THETA EPS

FR	AN	HS	AR	AL	GM
0.528	0.548	0.716	0.412	0.410	0.621

Una de las principales ventajas de la estimación máximo versosímil, razón por la que en determinados campos ha prevalecido sobre las técnicas mínimo cuadráticas clásicas, radica en que permite verificar la idoneidad del modelo propuesto como generador de las correlaciones observadas, mediante el ratio de funciones de verosimilitud, λ. En efecto, siempre y cuando $\nu > 0$, este ratio corregido, $-2\ ln\lambda$, refiere en qué medida las ν condiciones excedentes violan los momentos muestrales, pues coincide con el mínimo de la función criterio que se utiliza en el proceso de estimación

$$-2\ ln\lambda = (N-1)F_m(\Lambda\ \theta) \qquad [61]$$

siendo

$$F_m(\Lambda\ \theta) = ln|\Sigma| - tr|R\Sigma^{-1}| - ln|R| - p \qquad [62]$$

la función a minimizar mediante el algoritmo de Fletcher y Powell (1963: 163-168).

Este mínimo [62], se distribuye aproximadamente según una ley de χ^2 con ν gra-

dos de libertad, siempre que las hipótesis hechas sobre multinormalidad[4] de las v.o., número de factores, etc., puedan considerarse correctas. (Véase, para más información, Jöreskog, 1977; 1978: 443-477).

Si resumimos el proceso seguido, tenemos que: por un lado, el modelo de la fig. 5 se ha formulado introduciendo restricciones en los posibles valores de algunos elementos de Λ, θ y Φ. Por otro lado, tras comprobar la identificabilidad del modelo, se han recogido los datos apropiados, generalmente varianzas y covarianzas, para estimar los parámetros desconocidos, de manera que varianzas y covarianzas reproducidas por el modelo (54 ÷ 57) sean en algún sentido —criterio de la máxima verosimilitud— próximas a las observadas.

En estas condiciones, se ha contrastado el modelo propuesto con los datos recogidos en la tabla 22, pues si las restricciones especificadas en el modelo del AF fueran

TABLA 28. (a) Verificación del modelo de la fig. 5: Valor del estadístico χ^2 y nivel de significación; (b) Matriz, $\hat{R}$, reproducida a partir del modelo; (c) Matriz de residuos, $R - \hat{R}$.

TEST OF GOODNESS OF FIT
CHI SQUARE WITH 8 DEGREES OF FREEDOM IS 7,9533
PROBABILITY LEVEL = 0,9916

(a)

SIGMA

	FR	AN	HS	AR	AL	GM
FR	1,000					
AN	0,462	1,000				
HS	0,366	0,358	1,000			
AR	0,314	0,308	0,244	1,000		
AL	0,315	0,308	0,244	0,589	1,000	
GM	0,252	0,247	0,196	0,472	0,473	1,000

(b)

RESIDUALS: S − SIGMA

	FR	AN	HS	AR	AL	GM
FR	0,000					
AN	−0,023	0,000				
HS	0,044	−0,007	0,000			
AR	−0,026	0,046	−0,080	0,000		
AL	0,014	0,012	−0,054	0,006	0,000	
GM	−0,004	0,082	−0,015	−0,002	−0,009	0,000

(c)

[4] Poco se ha estudiado sobre la robustez de los estimadores máximo verosímiles, frente a la violación del supuesto de normalidad, sin embargo merecen destacarse los trabajos de Boosma (1981) y Muthen (1978: 551-560).

correctas, el ajuste del modelo a los datos será aceptable; por el contrario incorrectas restricciones, independientemente del criterio de estimación utilizado, redundarían en el empobrecimiento del ajuste, rechazándose si procediera este modelo como plausible representación de los datos. Es decir, hemos verificado la validez estadística del modelo global evaluando la bondad del ajuste.

En el caso que nos ocupa, aun reconociendo que otros modelos podrían ajustar igualmente nuestros datos, no puede rechazarse el modelo de la fig. 5 con una significación $\alpha = .0084$, según se desprende de la tabla 28. Esta precisión en el ajuste de las correlaciones observadas se traduce en residuos de pequeña magnitud, que se detallan a continuación.

Concretamente, la correlación reproducida $\hat{r}_{14}$ es:

$$\hat{r}_{14} = \lambda_{11}\lambda_{42}\phi_{12} = (.687)(.767)(.597) = .3145 \qquad [63]$$

y el residuo correspondiente

$$r_{14} - \hat{r}_{14} = .288 - .314 = -.026 \qquad [64]$$

Por último el criterio de estimación de la máxima verosimilitud proporciona, junto a este test global, la posibilidad de verificar individualmente cada uno de los parámetros estimados. En el proceso de minimización se obtiene como resultado marginal la inversa de la matriz de información (Jöreskog, 1977), de la que se derivan las varianzas de las estimaciones de los parámetros, que nos permitirán, mediante el estadístico de Student, su contrastación.

La tabla adjunta refiere para cada uno de los parámetros la mentada verificación

TABLA 29. (a) Desviaciones tipo de los parámetros estimados; (b) Valores del estadístico [66], para cada una de las estimaciones de la tabla 27.

LAMBDA X	STANDARD ERRORS		T-VALUES	
FR	0.076	0.000	9.075	0.000
AN	0.076	0.000	8.899	0.000
HS	0.076	0.000	7.044	0.000
AR	0.000	0.067	0.000	11.380
AL	0.000	0.067	0.000	11.410
GM	0.000	0.069	0.000	8.945

(a) (b)

PHI	EQ. 1 (a)	EQ. 2 (a)	EQ. 1 (b)	EQ. 2 (b)
EQ. 1	0.000		0.000	
EQ. 2	0.072	0.000	8.313	0.000

(a)

THETA DELTA	FR	AN	HS	AR	AL	GM
1	0.082	0.082	0.082	0.068	0.068	0.071

(a)

	FR	AN	HS	AR	AL	GM
1	6.422	6.707	8.754	6.057	6.011	8.692

(b)

Fijémonos, por ejemplo, en la saturación λ_{42}; la consiguiente prueba de hipótesis que su contrastación supone es

$$\begin{array}{l}\text{Hipótesis nula: } H_0 : E(\hat{\lambda}_{42}) = \lambda_{42} = 0 \\ \text{Hipótesis alternativa: } H_1 ; E(\hat{\lambda}_{42}) \neq 0\end{array} \qquad [65]$$

Si la hipótesis nula fuese cierta, el estadístico

$$\frac{\hat{\lambda}_{42} - \lambda_{42}}{SE(\hat{\lambda}_{42})} \qquad [66]$$

debe distribuirse según la *t*-Student, no obstante

$$\frac{.767 - 0}{.067} = 11.38 \qquad [67]$$

la magnitud del resultado [67], relativa al correspondiente valor de Student ($\cong 2$), implica el rechazo de la hipótesis nula[5]. En la práctica, este procedimiento se abrevia comprobando simplemente si el intérvalo de confianza:

$$\hat{\lambda}_{ik} \pm 2\ SE(\hat{\lambda}_{ik}) \qquad [68]$$

equivalente al estadístico [66], incluye o no el cero.

Para finalizar, se presentan los coeficientes de las v.o. en la combinación lineal que determina cada uno de los factores comunes. El programa LISREL IV obtiene estos valores por el método de regresión del apart. 2.2.4.

TABLA 30. Coeficientes de las seis v.o. en la determinación de los factores de la fig. 5, proporcionados por el programa LISREL IV.

FACTORES SCORES REGRESSIONS

	ETA					
	FR	AN	HS	AR	AL	GM
ETA 1	0,372	0,351	0,213	0,098	0,099	0,052
ETA 2	0,069	0,065	0,039	0,387	0,391	0,207

[5] No obstante, este proceso de contrastación debe llevarse a cabo una vez el modelo «ajusta» y con ciertas precauciones: por un lado, Pijper y Saris (1979), aclaran que la selección de las restricciones en el modelo no es independiente respecto de los tests estadísticos, como el de Student; por otro lado, en Muthen (1978: 551-560) se advierte que la estandarización de las variables originales influye considerablemente en los valores de las desviaciones tipo de los parámetros estimados (SE).

Resultados que asimismo difieren de los obtenidos mediante el programa de AF exploratorio del SPSS, en el que se permitió rotación oblicua, $\phi = .517$, a los factores de la tabla 26.

TABLA 31. Estimaciones de: (a) los coeficientes para la obtención de los factores; (b) la matriz factorial; (c) la estructura factorial, referida en nota a pie de página del apartado 2.2.3, obtenidas con el programa Factor Analysis del SPSS.

FACTOR SCORE COEFFICIENTS

	Factor 1	Factor 2
FR	0,06382	0,40152
AN	0,10045	0,28060
HS	0,00838	0,29875
AR	0,42705	0,03774
AL	0,35521	0,06678
GM	0,19427	0,06456

FACTOR PATTERN

	Factor 1	Factor 2
FR	0,05420	0,66804
AN	0,19234	0,51832
HS	−0,08681	0,63728
AR	0,80889	−0,04619
AL	0,74844	0,01111
GM	0,57921	0,05996

(b)

FACTOR STRUCTURE

	Factor 1	Factor 2
FR	0,39939	0,69605
AN	0,46017	0,61770
HS	0,24249	0,59243
AR	0,78503	0,37179
AL	0,75418	0,39785
GM	0,61019	0,35926

(c)

2.4.4. *Generalidad del modelo de Análisis Factorial*

En el apartado anterior, el modelo de AFC ha sido útil para estructurar los datos según una cierta teoría o para confirmar los resultados de estudios exploratorios previos.

En otras situaciones, en las que los datos provienen de respuestas a los items de un test, cuya confección se llevó a cabo asignando los items a categorías de una clasificación factorial o jerárquica de acuerdo a características de contenido o formato, el análisis de los resultados obedece a metodologías propias de los diseños factoriales (Bock, 1960: 151-163). No obstante, la posibilidad de introducir restricciones en el modelo de AF permite considerar como casos particulares un gran número de modelos (Alwin y Jackson, 1980), entre los cuales se encuentran: los modelos para el análisis de las componentes de la varianza y covarianza de Bock y Bargamn (1966: 507-534); las

matrices multirasgo-multimétodo (MMM) de Campbell y Fiske (1959: 81-105); los modelos de la teoría clásica de los tests de Lord y Novick (1968): Congenéricos, Tau-equivalentes y Paralelos; y los modelos Simplex y Circumplexos de Guttman (1954). Es decir, todos aquellos instrumentos de medida que obedecen a un diseño o estructura predeterminada.

Algunos ejemplos que ilustran la especificación de modelos a partir del de AFC o del modelo más general para el análisis estructural de las matrices de varianzas-covarianzas (ACOVS), pueden encontrarse en Jöreskog (1970: 239-251; 1970; 1971: 109-133) y Batista (1983). Se han seleccionado a continuación dos matrices Multirasgo-Multimétodo para ilustrar la aplicación del AFC a la Psicología y Sociología.

1. En un estudio sobre validez convergente y discriminante, Kelley y Fiske (1959) evaluaron a 124 estudiantes de Psicología con respecto a varios rasgos de personalidad, según tres métodos distintos: Autoevaluación (S); Mediana de las puntuaciones asignadas por tres de sus compañeros de clase (C); y combinación de las puntuaciones otorgadas por tres miembros del profesorado (E).

Basándose en este estudio, Campbell y Fiske (1959: 81-105) escogieron las características: Afirmativo; Alegre; Serio; Equilibrado; De intereses amplios, para introducir las MMM en la comprobación de la validez de los métodos de medida utilizados para medir los rasgos.

Si en primer lugar se consideran los tres métodos como un conjunto de medidas congenéricas, ignorando en principio el incremento que redundaría de su diferen-

TABLA 32. Matriz de correlaciones entre cinco rasgos de personalidad, evaluados por tres métodos diferentes (E, C, S).

MATRIX TO BE ANALYZED

	AFIR(E)	ALEG(E)	SERI(E)	EQUI(E)	INTE(E)	AFIR(C)	ALEG(C)	SERI(C)	EQUI(C)	INTE(C)
AFIR(E)	1.000									
ALEG(E)	0.370	1.000								
SERI(E)	-0.240	-0.140	1.000							
EQUI(E)	0.250	0.460	0.080	1.000						
INTE(E)	0.350	0.190	0.090	0.310	1.000					
AFIR(C)	0.710	0.350	-0.180	0.260	0.410	1.000				
ALEG(C)	0.390	0.530	-0.150	0.380	0.290	0.370	1.000			
SERI(C)	-0.270	-0.310	0.430	-0.060	0.030	-0.150	-0.190	1.000		
EQUI(C)	0.030	-0.050	0.030	0.200	0.070	0.110	0.230	0.190	1.000	
INTE(C)	0.190	0.050	0.040	0.290	0.470	0.330	0.220	0.190	0.290	1.000
AFIR(S)	0.480	0.310	-0.220	0.190	0.120	0.460	0.360	-0.150	0.120	0.230
ALEG(S)	0.170	0.420	-0.100	0.100	-0.050	0.090	0.240	-0.250	-0.110	-0.030
SERI(S)	-0.040	-0.130	0.220	-0.130	-0.050	-0.040	-0.110	0.310	0.060	0.060
EQUI(S)	0.130	0.270	-0.050	0.220	-0.040	0.100	0.150	0.000	0.140	-0.030
INTE(S)	0.370	0.150	-0.220	0.090	0.260	0.270	0.120	-0.070	0.050	0.350

MATRIX TO BE ANALYZED

	AFIR(S)	ALEG(S)	SERI(S)	EQUI(S)	INTE(S)
AFIR(E)	1.000				
ALEG(E)	0.230	1.000			
SERI(E)	-0.050	-0.120	1.000		
EQUI(E)	0.160	0.260	0.110	1.000	
INTE(E)	0.210	0.150	0.170	0.310	1.000

ciación en la variabilidad explicada de los datos de la tabla 32, se obtiene un modelo de cinco factores-rasgos que comporta un valor de $\chi^2_8 = 140.46$. Ya que $E(\chi^2) = \nu$, el excesivo valor del estadístico χ^2 con respecto al número de grados de libertad conduce al rechazo del modelo de cinco factores.

Con el objetivo de mejorar el ajuste de los datos, se añaden los métodos de medida (S, C y E) como factores comunes entre las v.o., esto conduce a la siguiente especificación

$$x = \Lambda\xi + \delta \qquad [69]$$

donde

$$\Lambda = \begin{bmatrix} \lambda_{11} & o & o & o & o & \lambda_{16} & o & o \\ o & \lambda_{22} & o & o & o & \lambda_{26} & o & o \\ o & o & \lambda_{33} & o & o & \lambda_{36} & o & o \\ o & o & o & \lambda_{44} & o & \lambda_{46} & o & o \\ o & o & o & o & \lambda_{55} & \lambda_{56} & o & o \\ \lambda_{61} & o & o & o & o & o & \lambda_{67} & o \\ o & \lambda_{72} & o & o & o & o & \lambda_{77} & o \\ o & o & \lambda_{83} & o & o & o & \lambda_{87} & o \\ o & o & o & \lambda_{94} & o & o & \lambda_{97} & o \\ o & o & o & o & \lambda_{105} & o & \lambda_{107} & o \\ \lambda_{111} & o & o & o & o & o & o & \lambda_{118} \\ o & \lambda_{122} & o & o & o & o & o & \lambda_{128} \\ o & o & \lambda_{133} & o & o & o & o & \lambda_{36} \\ o & o & o & \lambda_{144} & o & o & o & \lambda_{46} \\ o & o & o & o & \lambda_{155} & o & o & \lambda_{58} \end{bmatrix} \qquad [70]$$

$$E(x) = E(\xi) = E(\delta) = 0 \qquad [71]$$

$$E(\xi\delta') = 0 \qquad [72]$$

$$E(\delta\delta') = \theta \text{ , diagonal} \qquad [73]$$

adicionalmente se asume incorrelación entre métodos y rasgos, lo que induce cierta forma característica en la matriz de correlaciones entre factores

$$\Phi = \begin{bmatrix} 1 & & & & & & & \\ \phi_{12} & 1 & & & & & & \\ \phi_{13} & \phi_{23} & 1 & & & & & \\ \phi_{14} & \phi_{24} & \phi_{34} & 1 & & & & \\ \phi_{15} & \phi_{25} & \phi_{35} & \phi_{45} & 1 & & & \\ o & o & o & o & o & 1 & & \\ o & o & o & o & o & \phi_{67} & 1 & \\ o & o & o & o & o & \phi_{68} & \phi_{78} & 1 \end{bmatrix} \qquad [74]$$

Aunque este modelo proporciona un valor de $\chi^2_{n-62} = 57.46$, el criterio de parquedad, siempre presente en el modelado, al ser $\phi_{68} = 1$ aconseja considerar como congenéricos los métodos-factores de Autoevaluación (ξ_8) y el del equipo de profesores (ξ_6), integrándolos en un único factor (ξ_6). Los resultados obtenidos con el programa LISREL IV, son

TABLA 33. Estimaciones de: (a) la matriz factorial; (b) las entre rasgos-métodos; (c) las varianzas residuales del modelo propuesto.

LISREL ESTIMATES

LAMBDA X

	KSI 1	KSI 2	KSI 3	KSI 4	KSI 5	KSI 6	KSI 7
AFIR(E)	0.871	0.000	0.000	0.000	0.000	0.107	0.000
ALEG(E)	0.000	0.836	0.000	0.000	0.000	0.017	0.000
SERI(E)	0.000	0.000	0.573	0.000	0.000	-0.296	0.000
EQUI(E)	0.000	0.000	0.000	0.781	0.000	-0.253	0.000
INTE(E)	0.000	0.000	0.000	0.000	0.689	-0.337	0.000
AFIR(C)	0.829	0.000	0.000	0.000	0.000	0.000	0.162
ALEG(C)	0.000	0.696	0.000	0.000	0.000	0.000	0.294
SERI(C)	0.000	0.000	0.722	0.000	0.000	0.000	0.322
EQUI(C)	0.000	0.000	0.000	0.213	0.000	0.000	0.532
INTE(C)	0.000	0.000	0.000	0.000	0.599	0.000	0.440
AFIR(S)	0.552	0.000	0.000	0.000	0.000	0.110	0.000
ALEG(S)	0.000	0.454	0.000	0.000	0.000	0.221	0.000
SERI(S)	0.000	0.000	0.427	0.000	0.000	0.227	0.000
EQUI(S)	0.000	0.000	0.000	0.429	0.000	0.380	0.000
INTE(S)	0.000	0.000	0.000	0.000	0.697	0.622	0.000

(a)

PHI

	KSI 1	KSI 2	KSI 3	KSI 4	KSI 5	KSI 6	KSI 7
KSI 1	1.000						
KSI 2	0.559	1.000					
KSI 3	-0.371	-0.438	1.000				
KSI 4	0.380	0.662	-0.081	1.000			
KSI 5	0.548	0.291	-0.023	0.430	1.000		
KSI 6	0.000	0.000	0.000	0.000	0.000	1.000	
KSI 7	0.000	0.000	0.000	0.000	0.000	-0.208	1.00

(b)

THETA DELTA

AFIR(E)	ALEG(E)	SERI(E)	EQUI(E)	INTE(E)
0.239	0.302	0.583	0.318	0.391
AFIR(S)	**ALEG(S)**	**SERI(S)**	**EQUI(S)**	**INTE(S)**
0.689	0.751	0.767	0.678	0.166
AFIR(C)	**ALEG(C)**	**SERI(C)**	**EQUI(C)**	**INTE(C)**
0.291	0.468	0.349	0.676	0.428

(c)

2. Saris y otros, basándose en los resultados de una encuesta sobre comportamiento electoral, quieren determinar mediante MMM: a) Si el encuestado podía proporcionar más información de la que realmente se recoge con el procedimiento clásico de la escala de categorías. b) Si los resultados difieren sustancialmente de los obtenidos utilizando otros métodos de medida[6].

[6] En Psicofisiología, se admite generalmente que el sujeto almacena sus sensaciones como puntuaciones en un continuo, y por tanto es capaz de reproducirlas en una escala cuantitativa. Esto sugiere que puede obtenerse mayor información del sujeto que la recogida por una escala ordinal como la de categorías, sirviéndonos, por ejemplo, de los llamados procedimientos «de comparación» de Stevens (1966: 530-541; 1975).

En una muestra de 60 individuos se seleccionaron tres variables, típicas en estudios «del voto»: Identificación con un partido, ξ_1; Confianza en un candidato, ξ_2; y Preferencia por un candidato, ξ_3. Cada una de ellas se midió de tres formas distintas: en una escala de categorías (c); asignando un número (n); o según la longitud de una línea (l), que el encuestado trazaba.

Independientemente del modelo causal en el que se relacionasen las variables «del voto», puede considerarse el modelo de medida de la fig. 6, en el que se explicita el posible efecto de los distintos métodos de medida en los patrones de relación entre v.o.

El modelo matemático que traduce el diagrama de la fig. 6, consta de análogas ecuaciones que las referidas en el ejemplo anterior, si exceptuamos aquellas [70], [74], que corresponden a la matriz de ponderaciones, y a la de correlaciones entre los factores, cuya especificación puede encontrarse en el programa de la tabla 33.

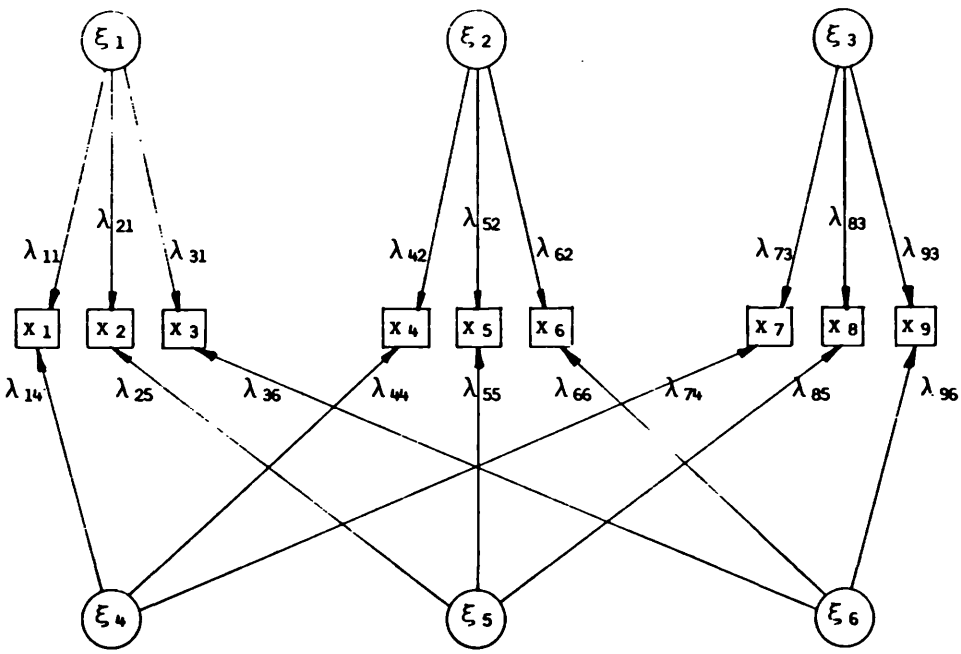


FIGURA 6. Modelo de medida de las tres variables del voto evaluadas por tres métodos (ξ_4, ξ_5, ξ_6).

El conjunto de sentencias requeridas por el programa LISREL IV, para la especificación del modelo confirmatorio de AF que se corresponde con el diagrama causal de la fig. 6, son extraordinariamente sencillas. La secuencia de tablas que se presentan a continuación detallan dicho programa, junto con las estimaciones de los parámetros y algunas de las habituales comprobaciones del modelo que el listado proporciona.

TABLA 33. Especificación del programa correspondiente al modelo de la fig. 6, para su análisis con el LISREL IV.

```
AFC.MMM.1(V)
DA NI=9 NO=60 MA=KM
KM FU UN=2 FO
*
LABELS
(AB)
PIC
CCC
CPC
PIL
CCL
CPL
PIN
CCN
CPN
SELECT
1 4 7 2 5 8 3 6 9/
MODEL NX=9 NK=6 TD=DI,FR
PA LX
*
1 0 0 1 0 0
1 0 0 0 1 0
1 0 0 0 0 1
0 1 0 1 0 0
0 1 0 0 1 0
0 1 0 0 0 1
0 0 1 1 0 0
0 0 1 0 1 0
0 0 1 0 0 1
PA PH
*
0
1 0
1 1 0
0 0 0 0
0 0 0 1 0
0 0 0 1 1 0
START     .7 ALL
START     .7 PH(1,2) PH(1,3) PH(1,4) PH(1,5) PH(1,6) C
             PH(2,3) PH(2,4) PH(2,5) PH(2,6) C
             PH(3,4) PH(3,5) PH(3,6) C
             PH(4,5) PH(4,6) C
             PH(5,6)
START     1  PH(1,1) PH(2,2) PH(3,3) PH(4,4) PH(5,5) PH(6,6)
OU        PM MR SE FD FS TV TO
```

TABLA 34. Estimación de los parámetros y verificación del modelo referido en el programa de la tabla anterior.

AFC. MMM. 1(V)
LISREL ESTIMATES

LAMBDA X

	KSI 1	KSI 2	KSI 3	KSI 4	KSI 5	KSI 6
PIc	0.781	0.000	0.000	0.445	0.000	0.000
PIl	0.945	0.000	0.000	0.000	0.218	0.000
PIn	0.950	0.000	0.000	0.000	0.000	0.131
CCc	0.000	0.861	0.000	0.310	0.000	0.000
CCl	0.000	0.905	0.000	0.000	0.358	0.000
CCn	0.000	0.884	0.000	0.000	0.000	0.434
CPc	0.000	0.000	0.751	0.651	0.000	0.000
CPl	0.000	0.000	0.792	0.000	0.639	0.000
CPn	0.000	0.000	0.785	0.000	0.000	0.575

PHI

	KSI 1	KSI 2	KSI 3	KSI 4	KSI 5	KSI 6
KSI 1	1.000					
KSI 2	0.833	1.000				
KSI 3	0.932	0.930	1.000			
KSI 4	0.000	0.000	0.000	1.000		
KSI 5	0.000	0.000	0.000	0.467	1.000	
KSI 6	0.000	0.000	0.000	0.378	0.922	1.000

THETA DELTA

	PIc	PIl	PIn	CCc	CCl	CCn
1	0.169	0.065	0.084	0.152	0.062	0.032

THETA DELTA

	CPc	CPl	CPn
1	-0.017	-0.022	0.062

TEST OF GOODNESS OF FIT
CHI SQUARE WITH 12 DEGREES OF FREEDOM IS 6.8596
PROBABILITY LEVEL = 1.0000

Saris y otros (1982), llegan a la conclusión de que realmente puede obtenerse más información de la que se obtiene en una escala de categorías utilizando métodos psicofísicos de comparación. El interesado puede consultar Pipjer y Saris (1979), para mayor información sobre el sesgo en las estimaciones de los parámetros estructurales, que relacionan las variables PI (ξ_1); CC (ξ_2); y CP (ξ_3), al utilizar la escala de categorías.

Desde esta perspectiva confirmatoria, Jöreskog ha planteado modelos cada vez más generales: Análisis de la estructura de las matrices de Variancia-Covariancias, ACOVS (Jöreskog, 1981: 65-92), y las recientes versiones del planteamiento LISREL, para analizar cualquier tipo de relaciones de estructura lineal (Jöreskog y Sorbom, 1978; 1981). Véase también para mayor información Saris (1980: 205-224; 1978), Batista (1982) y Batista y Cuadras (1983).

TABLA 35. Matriz reproducida (SIGMA = $\hat{R}$), según las estimaciones de la tabla 34. Evaluación del desajuste a través de los residuos, $R - \hat{R}$.

SIGMA

	PIc	PIl	PIn	CCc	CCl	CCn
PIc	0.978					
PIl	0.784	1.005				
PIn	0.764	0.924	1.003			
CCc	0.698	0.709	0.696	0.989		
CCl	0.663	0.790	0.759	0.831	1.009	
CCn	0.648	0.783	0.756	0.811	0.943	1.001
CPc	0.837	0.728	0.697	0.803	0.741	0.724
CPl	0.710	0.837	0.778	0.726	0.896	0.907
CPn	0.668	0.807	0.770	0.695	0.850	0.895

SIGMA

	CPc	CPl	CPn
CPc	0.971		
CPl	0.789	1.014	
CPn	0.731	0.961	1.008

RESIDUALS : S - SIGMA

	PIc	PIl	PIn	CCc	CCl	CCn
PIc	0.022					
PIl	0.016	-0.005				
PIn	0.006	-0.004	-0.003			
CCc	0.042	0.011	0.034	0.011		
CCl	0.027	-0.030	-0.019	0.009	-0.009	
CCn	0.022	-0.013	0.004	-0.001	-0.003	-0.001
CPc	0.033	0.012	0.023	0.027	0.009	0.006
CPl	0.020	-0.017	-0.008	0.014	-0.016	-0.007
CPn	0.032	-0.027	-0.010	0.035	0.000	0.005

RESIDUALS : S - SIGMA

	CPc	CPl	CPn
CPc	0.029		
CPl	0.011	-0.014	
CPn	0.019	-0.011	-0.008

TABLA 36. Valores del estadístico [66], para comprobar la significación individual de las estimaciones de cada uno de los parámetros del modelo.

AFC. MMM. 1(V)
T-VALUES

LAMBDA X

	KSI 1	KSI 2	KSI 3	KSI 4	KSI 5	KSI 6
PIc	6.918	0.000	0.000	4.247	0.000	0.000
PIl	9.168	0.000	0.000	0.000	1.524	0.000
PIn	9.467	0.000	0.000	0.000	0.000	0.877
CCc	0.000	8.121	0.000	3.010	0.000	0.000
CCl	0.000	8.045	0.000	0.000	2.634	0.000
CCn	0.000	7.550	0.000	0.000	0.000	3.464
CPc	0.000	0.000	5.451	5.054	0.000	0.000
CPl	0.000	0.000	5.156	0.000	4.382	0.000
CPn	0.000	0.000	5.186	0.000	0.000	3.504

PHI

	KSI 1	KSI 2	KSI 3	KSI 4	KSI 5	KSI 6
KSI 1	0.000					
KSI 2	16.739	0.000				
KSI 3	18.478	29.268	0.000			
KSI 4	0.000	0.000	0.000	0.000		
KSI 5	0.000	0.000	0.000	2.129	0.000	
KSI 6	0.000	0.000	0.000	1.440	15.497	0.000

THETA DELTA

	PIc	PIl	PIn	CCc	CCl	CCn
1	3.116	2.186	2.228	3.921	3.153	1.778

THETA DELTA

	CPc	CPl	CPn
1	-0.279	-0.718	3.015

3. Análisis Factorial de Correspondencias

Por *José Miguel García Santesmases*

3.1. Introducción

La rápida generalización del uso de esta técnica de análisis de datos se debe fundamentalmente a lo adecuada que resulta su utilización para atacar problemas de análisis donde juegan un papel importante las variables nominales.

El impulso inicial fue dado por la escuela francesa de análisis de datos y a uno de sus fundadores, J. P. Benzecri, puede citarse como el primer autor de esta técnica en lo que se refiere a sus aspectos descriptivos y algebraicos; es decir, al ataque de tablas de contingencia fuera del marco de la estadística inferencial clásica.

La utilización de esta técnica responde a la necesidad de profundizar en las relaciones de dependencia que se establecen entre dos variables categóricas observadas en una misma población, incidiendo sobre todo en explicar cómo los distintos valores o categorías de ambas variables se relcionan unos con otros.

Las consideraciones que se hacen para llegar a los resultados son de carácter geométrico y están dentro de las técnicas descriptivas de la estadística, no pudiendo debido precisamente a esto extender en principio ninguna de las conclusiones que se obtengan más allá del conjunto colectivo observado.

El tipo de resultados a los que se llega tiene por tanto un carácter meramente descriptivo del colectivo estudiado, siendo especialmente idóneo para aplicarlo a situaciones donde sean pocas o ninguna las hipótesis previas de trabajo y se requiera un análisis exploratorio de la situación a tratar (a través de una muestra del colectivo en estudio) con el fin de establecer los puntos de partida de análisis posteriores.

El objeto del análisis de correspondencias son las tablas de contingencia o tablas que cruzan dos variables categóricas. Existe una generalización del método a más de dos variables categóricas, llamado análisis de correspondencias múltiples, que lo hace especialmente útil en situaciones multivariables categóricas.

Vamos a desarrollar primero el método simple o el análisis de una tabla de contingencia.

3.2. Correspondencias simples

Sean X e Y las variables que forman la tabla de contingencia, con «p» y «q» valores o categorías respectivamente.

Denotamos la tabla en la siguiente forma:

X \ Y	1	2	j	q
1	n_{11}	n_{12}	n_{1j}	n_{1q}
2	n_{21}	n_{22}	n_{2j}	n_{2q}
P	n_{p1}	n_{p2}	n_{pj}	n_{pq}

donde n_{ij} designa el número de unidades observadas que toman el valor i en la variable X y el valor j en la variable Y.

$$N = \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij}$$

Sea

$$f_{ij} = \frac{n_{ij}}{N} \qquad \begin{matrix} i = 1, \ldots, p \\ j = 1, \ldots, q \end{matrix}$$

$$f_{i\cdot} = \sum_{j=1}^{q} f_{ij}$$

$$f_{\cdot j} = \sum_{i=1}^{p} f_{ij}$$

Esta técnica también se puede aplicar a cuadros o tablas de números positivos que no sean necesariamente tablas frecuentistas o de recuento. Así podemos considerar la distribución del Producto Nacional Bruto entre los sectores Agrícola, Industrial y de Servicios observada en un determinado conjunto de países, tal y como aparece en la tabla 1.

Las preguntas que nos hacemos a la hora de establecer el tipo de dependencia que se da entre las variables X e Y pueden reducirse a:

1. ¿Qué categorías de la variable X se comportan de forma similar respecto de la variable Y? (Análogamente para las categorías de la variable Y respecto de X).
2. ¿Cuáles son los valores de Y que no influyen en una determinada categoría de X?
3. ¿Qué tipo de representaciones gráficas pueden visualizar los resultados obteni-

TABLA 1. Matriz de Datos.

	AGRI	INDU	SERV	
$f_{\cdot j}$	400.	1055.	1545.	3000.
ARGE	13.	46.	41.	100.
BOLI	17.	29.	54.	100.
BRAS	11.	38.	51.	100.
CHIL	8.	37.	55.	100.
COLO	29.	28.	43.	100.
COST	19.	26.	55.	100.
ECUA	15.	37.	48.	100.
SALV	28.	22.	50.	100.
GUAT	26.	20.	54.	100.
HOND	32.	26.	42.	100.
MEJI	10.	38.	52.	100.
NICA	29.	28.	43.	100.
PANA	23.	21.	56.	100.
PARA	31.	24.	45.	100.
PERU	10.	43.	47.	100.
REDO	19.	26.	55.	100.
URUG	13.	37.	50.	100.
VENE	6.	47.	47.	100.
EEUU	3.	34.	63.	100.
CANA	4.	33.	63.	100.
ALEM	2.	49.	49.	100.
BELG	2.	37.	61.	100.
DINA	5.	59.	36.	100.
ESPA	9.	31.	60.	100.
FRAN	5.	34.	61.	100.
ITAL	7.	43.	50.	100.
PBAJ	4.	37.	59.	100.
PORT	13.	47.	40.	100.
INGL	2.	36.	62.	100.
JAPO	5.	42.	53.	100.

dos anteriormente, de forma que la alteración producida por estas representaciones en los datos observados sea mínima?

Para contestar a estas preguntas procedemos a la construcción y análisis de las nubes de puntos generados por ambas variables.

3.2.1. *Construcción de las nubes de puntos*

- Nube de puntos en R^p

Consideramos cada una de las categorías de la variable Y como un punto en un espacio de p dimensiones que corresponde a cada uno de los valores que toma respecto de la variable X-normalizado respecto del número total de efectivos.

Si nos fijamos en la categoría j de Y, que denominamos punto columna j, representamos este punto en R^p a través del vector

$$\left(\frac{f_{1j}}{f_{\cdot j}}, \frac{f_{2j}}{f_{\cdot j}}, \ldots, \frac{f_{pj}}{f_{\cdot j}}\right)$$

que corresponde en términos frecuentistas a la distribución de X condicionada al valor j de Y.

Así, en la tabla 1 podemos construir 3 puntos columnas que corresponden a las variables AGRI, INDU, SERV, situados en un espacio de treinta dimensiones.

Por ejemplo:

$$\text{AGRI} = \left(\frac{13}{400}, \frac{17}{400}, \frac{11}{400}, \ldots, \frac{5}{400}\right)$$

- Nube de puntos en R^q

De la misma forma en que construimos la nube de puntos en R^p construimos una nube de puntos en R^q, formada por los perfiles de las filas.

Tendremos p puntos filas correspondientes a las p categorías de la variable X; así, la categoría i, que denominaremos punto fila i, vendrá dada por el vector:

$$\left(\frac{f_{i1}}{f_{i\cdot}}, \frac{f_{i2}}{f_{i\cdot}}, \ldots, \frac{f_{iq}}{f_{i\cdot}}\right)$$

En la tabla 1 tendremos 30 puntos fila correspondientes a los perfiles de los 30 países considerados.

Por ejemplo Brasil está representado por:

$$\text{BRAS} = \left(\frac{13}{100}, \frac{46}{100}, \frac{41}{100}\right)$$

Tanto en R^p como en R^q, en las nubes de puntos consideradas las proximidades entre los puntos pueden interpretarse como proximidades entre las categorías que representan; así PBAJ y BELG están más cerca que PBAJ y BRAS.

3.2.2. *Distancia de Benzecri*

Con el fin de dar una idea más precisa de las proximidades entre los puntos en cada uno de los espacios considerados, les dotamos de una distancia que nos da dichas proximidades. Tomaremos para ello la distancia de Benzecri, basada en la Gi cuadrado.

Distancia en R^p.

Sean j y j' dos puntos columna que vienen representados respectivamente por:

$$\left(\frac{f_{1j}}{f_{\cdot j}}, \frac{f_{2j}}{f_{\cdot j}}, \ldots, \frac{f_{pj}}{f_{\cdot j}}\right) \text{ y } \left(\frac{f_{1j'}}{f_{\cdot j'}}, \frac{f_{2j'}}{f_{\cdot j'}}, \ldots, \frac{f_{pj'}}{f_{\cdot j'}}\right)$$

La distancia al cuadrado entre los puntos j y j' viene dada por:

$$d^2(j, j') = \sum_{i=1}^{p} \frac{1}{f_{i\cdot}} \left(\frac{f_{ij}}{f_{\cdot j}} - \frac{f_{ij'}}{f_{\cdot j'}}\right)^2$$

que a diferencia de la métrica euclidea pondera cada sumando con los inversos de los tanto por uno de los efectivos en cada fila.

Análogamente definimos la distancia entre filas en R^q. Así, sean i e i' dos puntos fila en R^q

$$d^2(i, i') = \sum_{j=1}^{q} \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{i'j}}{f_{i'\cdot}}\right)^2$$

La justificación de la elección de dicha distancia se basa en que cumple con la propiedad de invariancia, que se puede expresar del siguiente modo:

«Si agregamos o juntamos dos columnas (filas) para formar una sola, con tal de que sus perfiles sean idénticos la distancia entre filas (columnas) queda inalterada».

Es decir, si se diese en nuestro ejemplo el caso de que dos países presentasen la misma distribución del P.N.B. en los tres sectores considerados, se podría considerar una nueva tabla con las mismas filas que la antigua, pero suprimiendo las de los dos países con la misma distribución e incluyendo una nueva que recogiese la suma de los efectivos de dichos países; en la nueva tabla, las distancias entre las columnas o sectores permanecen invariables.

Esta propiedad hace adecuada su utilización en el problema que nos ocupa, pues permite identificar aquellas filas (columnas) que sean semejantes sin que varíen sensiblemente las posiciones de las columnas (filas).

A los puntos que definen la nube en R^q o puntos filas, los vamos a ponderar con sus marginales; es decir, el punto fila i vendrá ponderado por $f_{i\cdot}$.

De la misma forma para los puntos columnas en R^p. Así, el punto columna j vendrá ponderado por $f_{\cdot j}$.

En nuestro ejemplo, todas las filas tienen la misma ponderación o peso y vale 100/3000.

Por el contrario, las columnas están ponderadas de forma distinta: AGRI tiene una ponderación de 400/3000, INDU tiene una ponderación de 1055/3000 y SERV tiene una ponderación de 1545/3000.

3.2.3. *Notación*

En el desarrollo que sigue emplearemos la siguiente notación matricial.
Si llamamos F a la matriz de frecuencias relativa de la tabla de contingencia

$$F = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1q} \\ f_{21} & f_{22} & \cdots & f_{2q} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ f_{p1} & f_{p2} & \cdots & f_{pq} \end{bmatrix}$$

Sea D_p la matriz diagonal

$$D_p = \begin{bmatrix} f_{1.} & o & \cdots & o \\ o & f_{2.} & \cdots & o \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ o & o & \cdots & f_{p.} \end{bmatrix} \quad ; \quad D_p^{-1} = \begin{bmatrix} 1/f_{1.} & o & \cdots & o \\ o & 1/f_{2.} & \cdots & o \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ o & o & \cdots & 1/f_{p.} \end{bmatrix}$$

que recoge las marginales de las filas.

Sea D_q la matriz diagonal

$$D_q = \begin{bmatrix} f_{.1} & o & \cdots & o \\ o & f_{.2} & \cdots & o \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ o & o & \cdots & f_{.q} \end{bmatrix} \quad ; \quad D_q^{-1} = \begin{bmatrix} 1/f_{.1} & o & \cdots & o \\ o & 1/f_{.2} & \cdots & o \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ o & o & \cdots & 1/f_{.q} \end{bmatrix}$$

que recoge las marginales de las columnas.

Se tiene:

$$D_p^{-1}F = \begin{bmatrix} \frac{f_{11}}{f_{1.}} & \frac{f_{12}}{f_{1.}} & \cdots & \frac{f_{1q}}{f_{1.}} \\ \\ \frac{f_{21}}{f_{2.}} & \frac{f_{22}}{f_{2.}} & \cdots & \frac{f_{2q}}{f_{2.}} \\ \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \\ \frac{f_{p1}}{f_{p.}} & \frac{f_{p2}}{f_{p.}} & \cdots & \frac{f_{pq}}{f_{p.}} \end{bmatrix}$$

que recoge en sus filas los puntos de la nube definida en R^q.

Análogamente,

$$FD_q^{-1} = \begin{bmatrix} \frac{f_{11}}{f_{.1}} & \frac{f_{12}}{f_{.2}} & \cdots & \frac{f_{1q}}{f_{.q}} \\ \\ \frac{f_{21}}{f_{.1}} & \frac{f_{22}}{f_{.2}} & \cdots & \frac{f_{2q}}{f_{.q}} \\ \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \\ \frac{f_{q1}}{f_{.1}} & \frac{f_{p2}}{f_{.2}} & \cdots & \frac{f_{pq}}{f_{.q}} \end{bmatrix}$$

recoge en sus columnas los puntos de la nube definida en R^p; luego las filas de la traspuesta de dicha matriz, $D_q^{-1}F^t$, representan a estos puntos.

De la misma forma utilizaremos la notación matricial para representar a los pesos o ponderaciones de los puntos de cada nube, así como las distancias entre ellos.

D_p matriz con los pesos de los p puntos en R^p.
D_q matriz con los pesos de los q puntos en R^q.
D_p^{-1} matriz que define la distancia en R^p.
D_q^{-1} matriz que define la distancia en R^q.

tendremos que la distancia entre los puntos i e i' de R^q, en términos de D_q^{-1}, vendrá dada por:

$$d^2(i, i') = \left(\frac{f_{i1}}{f_{i.}} - \frac{f_{i'1}}{f_{i'.}}, \frac{f_{i2}}{f_{i.}} - \frac{f_{i'2}}{f_{i'.}}, \ldots, \frac{f_{iq}}{f_{i.}} - \frac{f_{i'q}}{f_{i'.}} \right) D_q^{-1} \begin{pmatrix} \frac{f_{i1}}{f_{i.}} - \frac{f_{i'1}}{f_{i'.}} \\ \frac{f_{i2}}{f_{i.}} - \frac{f_{i'2}}{f_{i'.}} \\ \vdots \\ \frac{f_{iq}}{f_{i.}} - \frac{f_{i'q}}{f_{i'.}} \end{pmatrix}$$

análogamente para cualquier par de puntos j, j' de R^p.

Dado que el objetivo del análisis de correspondencias es estudiar las relaciones que existen entre las filas y columnas de una tabla de contingencia y debido a que las representaciones geométricas de las filas (columnas) están en espacios de dimensiones grandes $q(p)$, el método que emplearemos para estudiar dichas relaciones será el de obtener subespacios de $R^q(R^p)$ de dimensiones pequeñas que mejor se ajustan a dichas nubes de puntos, estudiando las relaciones que existen entre dichos subespacios.

3.2.4. *Ajuste de un subespacio a una nube de puntos*

Sea X, matriz de orden (p, q) que representa las q coordenadas de p puntos en R^q; sea M, matriz simétrica definida positiva que define la métrica en R^q; sea P, matriz diagonal de orden (p, p) cuyos elementos son los pesos de los p puntos.

Para encontrar el subespacio de R^q de dimensión $l < q$ que mejor se ajusta a la nube de puntos, comenzaremos por buscar el subespacio de dimensión 1 que mejor se ajuste. El criterio de ajuste que emplearemos es el de los mínimos cuadrados, que en nuestro caso se traduce en la siguiente expresión a optimizar.

Sea u_1 vector de norma 1 que engendra el subespacio de dimensión 1 a localizar.

XMu representan las proyecciones de los p puntos sobre el eje definido por u.

Se trata de maximizar la suma ponderada de los cuadrados de las proyecciones, condicionado a que u sea de norma 1; es decir, u_1 vendrá dado como solución al problema

$$\operatorname*{Max}_{u_1} \; u_1' M X' P X M u_1 \;/\; u_1' M u_1 = 1$$

La solución de dicho problema nos da que u_1 es el autovector de la matriz $X'PXM$ asociado al mayor autovalor de dicha matriz.

Luego:

$$X^tPXMu_1 = \lambda_1 u_1$$

a u_1 le llamamos primer eje factorial.

A $Mu_1 = \varphi_1$, operador proyección, le llamaremos primer factor; este factor es de norma 1 definida por la métrica M^{-1}; es decir:

$$\varphi_1^t M^{-1} \varphi_1 = u_1^t M M^{-1} M u_1 = u_1^t M u_1 = 1$$

Para calcular el subespacio de dimensión 2 que mejor se ajusta a la nube de puntos, basta con calcular el subespacio de dimensión 1 que mejor se ajusta a la nube entre todos los ortogonales al u_1 previamente calculado, ya que el «mejor» subespacio de dimensión 2 contiene al mejor subespacio de dimensión 1.

Sea u_2 el vector de norma 1 que engendra este subespacio, u_2 nos viene dado como aquel que resuelve el problema de máximo:

$$\underset{u_2}{\text{Max}}\ u_2^t M X^t P X M u_2 \ /\ u_2^t M u_2 = 1\ ;\ u_2^t M u_1 = 0$$

y este es el autovector asociado al segundo mayor autovalor de la matriz X^tPXM.

Denominaremos:

u_2 segundo eje factorial.

$Mu_2 = \varphi_2$ segundo factor.

En general tendremos:

El subespacio de dimensión $l < q$ que mejor se ajusta a la nube de puntos en R^q, definida más arriba, está engendrado por los l autovectores de la matriz X^tPXM asociados a los l mayores autovalores de dicha matriz.

Al α-ésimo autovector de X^tPXM, ordenado según el valor de su autovalor, lo llamaremos α-éximo eje factorial.

u_α α-ésimo eje factorial.

$\varphi_\alpha = Mu_\alpha$ α-ésimo factor.

Cálculo de los factores en R^q

Aplicando los resultados anteriores a la situación definida por:

$$X = D_p^{-1}F$$
$$M = D_q^{-1}$$
$$N = D_p$$

tendremos:

El eje factorial α-ésimo u_α de la nube definida en R^q viene dado por el autovector asociado al α-ésimo autovalor de

$$S = F^t D_p^{-1} F D_q^{-1}$$

siendo el elemento (j, j') de S:

$$S_{ij} = \sum_{i=1}^{p} \frac{f_{ij} f_{ij'}}{f_{i\cdot} f_{\cdot j'}}$$

$\varphi_\alpha = D_q^{-1} u_\alpha$ es el α-ésimo factor.

Las proyecciones de los p puntos de la nube R^q sobre el eje vienen dadas por

$$D_p^{-1} F D_q^{-1} u_\alpha = D_p^{-1} F \varphi_\alpha$$

Cálculo de los factores en R^p

Análogamente al caso en R^q el α-ésimo eje factorial es el autovector asociado al α-ésimo autovalor μ_α de la matriz:

$$F D_q^{-1} F' D_p^{-1}$$

Designamos por ψ_α al α-ésimo factor que viene dado por:

$$\psi_\alpha = D_p^{-1} v_\alpha$$

Las proyecciones de los q puntos de R^p *sobre* v_α vienen dados por:

$$D_q^{-1} F' \varphi_\alpha$$

Relaciones entre los resultados obtenidos para R^q y R^p

Por los anteriores resultados sabemos que se cumple:

$$F D_q^{-1} F' D_p^{-1} v_1 = \lambda_1 v_1$$

$$(2) \qquad F' D_p^{-1} F D_q^{-1} u_1 = \mu_1 u_1$$

Como las matrices $F D_q^{-1} F' D_p^{-1}$ y $F' D_p^{-1} F D_q^{-1}$ tienen los mismos autovalores no nulos se tiene que

$$\lambda_\alpha = \mu_\alpha$$

premultiplicando (2) por $F D_q^{-1}$ tenemos:

$$F D_q^{-1} F' D_p^{-1} (F D_q^{-1} u_1) = \lambda_1 (F D_q^{-1} u_1)$$

luego, $FD_q^{-1}u_1$ es el autovector asociado a λ_1 y, por tanto, proporcional a v_1. Como v_1 tiene norma 1 y además se cumple que

$$u_1' D_q^{-1} F' D_p^{-1} F D_q^{-1} u_1 = \lambda_1$$

ya que es la función objetivo a maximizar y tomaba su máximo para u_1 con valor λ_1, se tiene:

$$v_1 = \frac{1}{\sqrt{\lambda_1}} F D_q^{-1} u_1$$

En general:

$$v_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} F D_q^{-1} u_\alpha$$

$$u_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} F' D_p^{-1} v_\alpha$$

Estas relaciones se extienden a los factores, teniendo:

$$\varphi_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} D_q^{-1} F \psi_\alpha$$

$$\psi_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} D_p^{-1} F \varphi_\alpha$$

Si designamos por $\hat{\psi}_\alpha$, $\hat{\varphi}_\alpha$ los vectores de las proyecciones de las nubes de puntos sobre los α-ésimos ejes factoriales de R^q y R^p, respectivamente, se tiene:

$\hat{\varphi}_\alpha = \sqrt{\lambda_\alpha}\ \varphi_\alpha$ proyecciones de los puntos de R^p.

$\hat{\psi}_\alpha = \sqrt{\lambda_\alpha}\ \psi_\alpha$ proyecciones de los puntos de R^q.

y se establecen las siguientes relaciones baricéntricas entre las coordenadas de las proyecciones

$$\hat{\varphi}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} D_q^{-1} F \hat{\psi}_\alpha$$

$$\hat{\psi}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} D_q^{-1} F \hat{\psi}_\alpha$$

o lo que es igual:

$$\hat{\psi}_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^{q} \frac{f_{ij}}{f_{i\cdot}} \hat{\varphi}_{\alpha j}$$

$$\hat{\varphi}_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^{p} \frac{f_{ij}}{f_{\cdot j}} \hat{\psi}_{\alpha i}$$

3.3. Ejemplo de aplicación de las correspondencias simples

A la hora de llevar a la práctica esta técnica, y debido a los resultados anteriores, será suficiente con resolver el problema de la representación de una de las nubes de puntos, pues en función de las relaciones baricéntricas a las que hemos llegado siempre es posible resolver una nube a partir de la otra.

Se tomará la nube de puntos definida en el espacio de dimensión más pequeña, sea éste R^q; se extraerán como máximo los q mayores autovalores, pues debido a las relaciones que ligan a los puntos de la nube

$$\sum_{j=1}^{q} \frac{f_{ij}}{f_{i\cdot}} = 1$$

éstos se encuentran en un subespacio de dimensión $q - 1$.

Para la tabla 1 tomaremos la nube de puntos fila o países, pues estando definidos en R^3, solamente tendremos que extraer dos ejes factoriales que corresponden a los autovalores

$$u_1 = 0.09317$$

$$u_2 = 0.019432$$

y que representan respectivamente 81.062% y 18.93% de la variabilidad total.

Estos valores nos permiten establecer la importancia relativa de cada eje factorial y por tanto la de los planos que formemos combinando estos ejes.

En nuestro caso el plano factorial formado por los dos primeros ejes recoge un 100% de la variabilidad, teniendo definidos correctamente los puntos que representamos en él.

No será en general este el caso en aquellas situaciones donde tanto p como q san mayores que 3, debiendo tener cuidado en las interpretaciones que se hagan a la vista de los planos observados, puesto que los puntos en ellos representados sufrirán deformaciones tanto más fuertes cuanto menor sean los porcentajes de variabilidad que recogen estos ejes.

Las proyecciones de los puntos sobre el plano definido por los dos ejes extraídos para la tabla 1, vienen dadas en las figuras 1 y 2 que corresponden a las tablas 2 y 3.

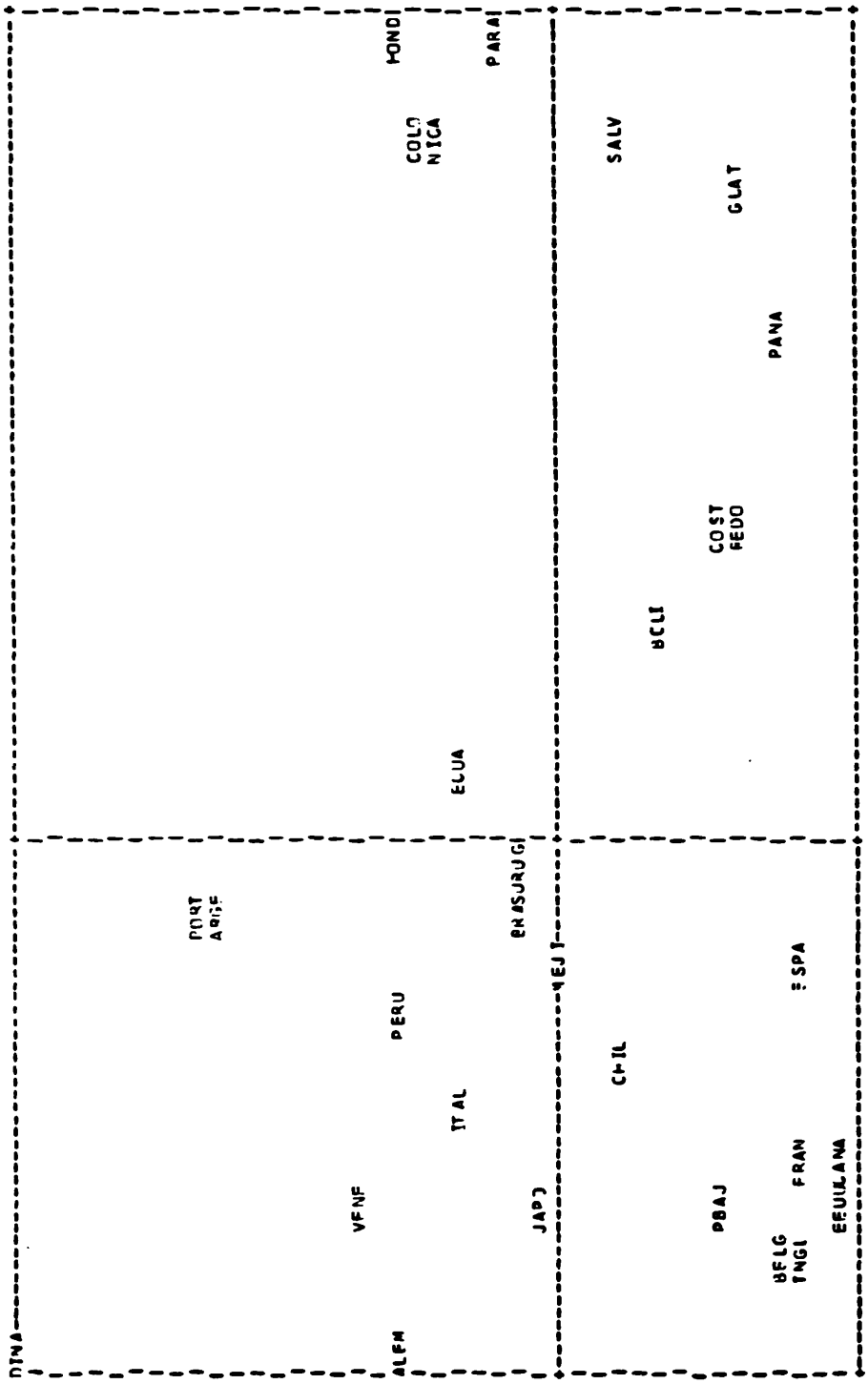


FIGURA 1. Proyecciones de los puntos-filas sobre el plano definido por los ejes 1 (horizontal) y 2 (vertical).

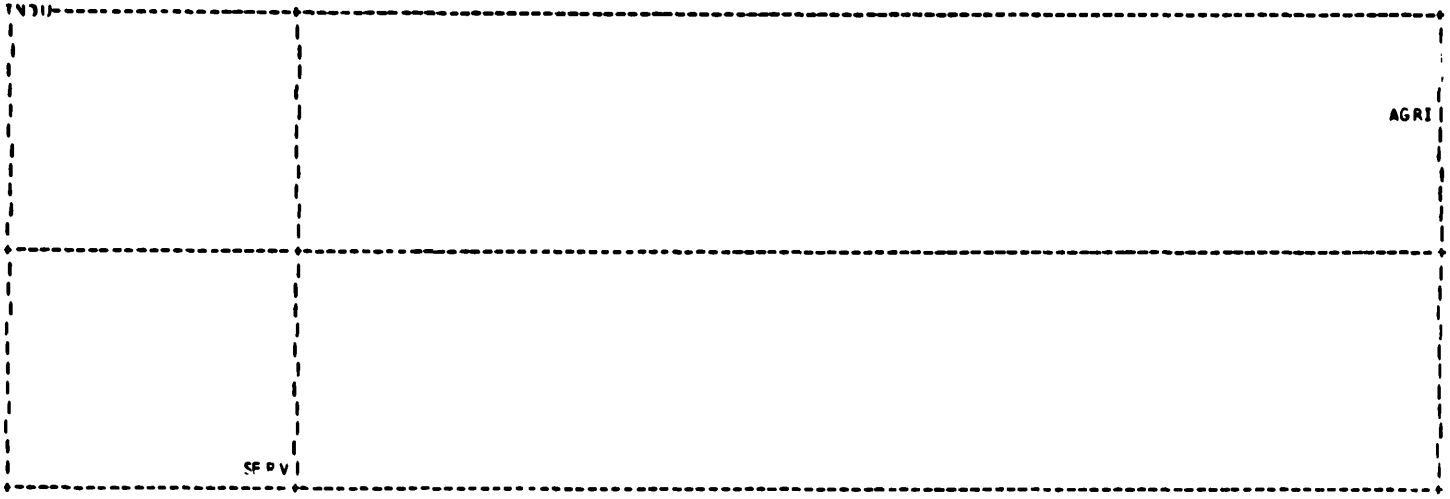


FIGURA 2. Proyecciones de los puntos-columna sobre el plano definido por los ejes 1 (horizontal) y 2 (vertical).

TABLA 2. Coordenadas de los puntos-fila en los 2 primeros ejes.

	FACTOR 1	FACTOR 2
ARGE	− 67	224
BOLI	130	− 71
BRAS	− 76	23
CHIL	−149	− 45
COLO	451	98
COST	199	−103
ECUA	34	65
SALV	457	− 44
GUAT	416	−118
HOND	540	104
MEJI	−102	7
NICA	451	98
PANA	331	−145
PARA	525	46
PERU	−129	113
REDO	199	−103
URUG	− 18	34
VENE	−256	134
EEUU	−264	−189
CANA	−232	−194
ALEM	−372	112
BELG	−307	−141
DINA	−348	372
ESPA	− 90	−156
FRAN	−211	−157
ITAL	−208	65
PBAJ	−254	−109
PORT	− 73	245
INGL	−301	−162
JAPO	−255	12

TABLA 3. Coordenadas de los puntos-columna sobre los 2 primeros ejes.

	FACTOR 1	FACTOR 2
AGRI	712	89
INDU	−202	162
SERV	− 44	−132

A veces resulta útil superponer las proyecciones de los punos fila y los puntos columna para formar un único plano donde podamos situar filas y columnas. Las posibles interpretaciones que puedan hacerse de esta representación conjunta se justifican por las relaciones baricéntricas.

En éstas, la coordenada de un punto fila (columna) sobre un eje se puede expresar como la medida ponderada de las coordenadas de los puntos columna (filas) sobre ese mismo eje, salvo el factor multiplicativo $1/\sqrt{\lambda_\alpha}$.

En la figura 3 aparece esta representación conjunta, a partir de la que podemos hacer las siguientes interpretaciones.

Si tomamos un punto fila, por ejemplo FRAN que representa Francia, observamos que los puntos columnas que más influyen en su posición son: SERV, INDU y AGRI, en este orden. Este tipo de análisis lo podemos hacer para cada punto fila o país observando que hay un conjunto de países que se comportan de forma similar a Francia y que se oponen horizontalmente a los países con predominio de un PNB en Agricultura (Honduras, Paraguay, etc...) y verticalmente a aquellos donde predomina la Industria respecto de los Servicios (Venezuela, Alemania, etc.).

Estas interpretaciones son válidas para la tabla que hemos tratado, pues tal como se comentó más arriba, en la representación bidimensional estudiada no se pierde ninguna información.

Para otras tablas de datos, de dimensiones mayores, en las representaciones planas que se construyen combinando los ejes extraídos habrá que tener más cuidado con las interpretaciones, pues dependerá de lo bien o mal que los puntos filas o columnas estén representados en el plano elegido para dar validez a las anteriores interpretaciones.

Además de la importancia relativa de cada eje, que viene dada por el porcentaje de variabilidad que se lleva, y que representa el autovalor asociado a dicho eje cuyo valor viene dado por la expresión

$$\lambda_\alpha = \sum_j f_{.j}\varphi_{\alpha j}^2 = \sum_j f_{.j}(\sqrt{\lambda_\alpha}\ \varphi_{\alpha j})^2$$

consideraremos expresiones que nos dan: 1) la importancia que tiene cada punto en la definición de un eje y, 2) la calidad de la representación de un punto sobre un eje, que vienen dadas respectivamente por las contribuciones absolutas y relativas.

Contribución absoluta del punto j para definir el eje.

$$C_a(j, \alpha) = \frac{f_{.j}(\sqrt{\lambda_\alpha}\ \varphi_{\alpha j})^2}{\lambda_\alpha}$$

representa el porcentaje de la inercia o variabilidad del eje definido por el punto j.

Se cumple:

$$\sum_j C_a(\alpha, j) = 1$$

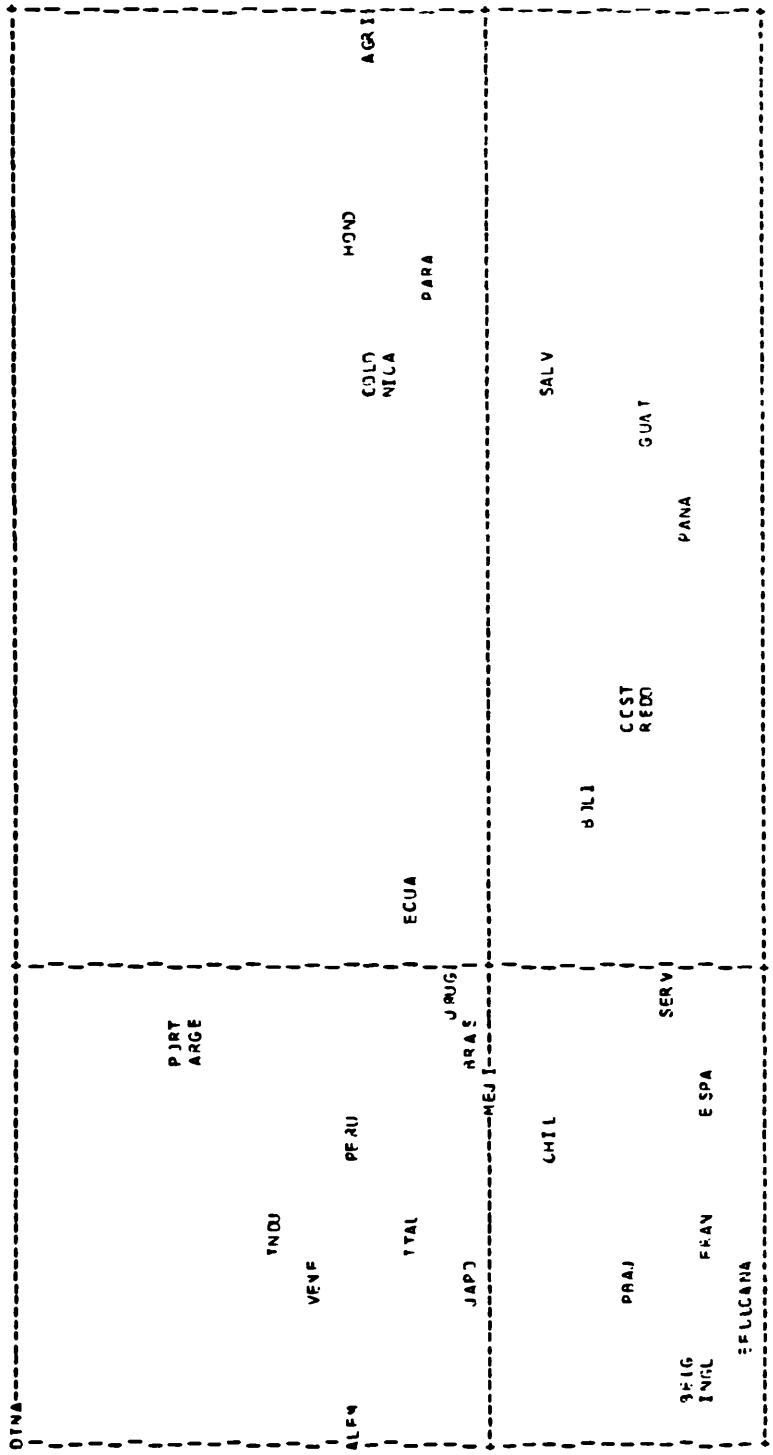


FIGURA 3. Proyecciones de los puntos fila-columna sobre el plano definido por los ejes 1 (horizontal) y 2 (vertical).

Calculando las contribuciones absolutas para cada punto fila (columna) podremos establecer qué puntos fila (columna) con los que más contribuyen en la definición de un eje y que al mismo tiempo sugieren una posible interpretación del mismo.

En las tablas 4 y 5 aparecen las contribuciones absolutas de los puntos filas y columnas de los dos ejes extraídos sobre los datos de la tabla 1.

TABLA 4. Contribuciones absolutas de los puntos-fila a los 2 primeros ejes.

	FACTOR 1	FACTOR 2
ARGE	2	86
BOLI	7	9
BRAS	2	1
CHIL	9	4
COLO	81	17
COST	16	18
ECUA	0	7
SALV	84	3
GUAT	69	24
HOND	117	19
MEJI	4	0
NICA	81	17
PANA	44	36
PARA	110	4
PERU	7	22
REDO	16	18
URUG	0	2
VENE	27	31
EEUU	28	62
CANA	22	65
ALEM	56	22
BELG	38	35
DINA	49	237
ESPA	3	43
FRAN	18	43
ITAL	18	7
PBAJ	26	21
PORT	2	103
INGL	37	46
JAPO	26	0

Si nos fijamos en las contribuciones de los puntos columnas a los dos ejes considerados, observamos que el 1.er eje está definido por AGRI con una contribución absoluta de 812 sobre 1.000, mientras que el 2.º eje está definido por INDU y SERV con contribuciones de 473 y 472 respectivamente, en este último podemos observar que existe una oposición entre estos dos puntos pues sus coordenadas tienen signos opuestos, 162 y −132 respectivamente (véase Tabla 3).

TABLA 5. Contribuciones absolutas de los puntos columna a los 2 primeros factores.

	FACTOR 1	FACTOR 2
AGRI	812	54
INDU	175	473
SERV	13	472

La interpretación que sugieren estos valores puede ser la siguiente:

Los puntos se sitúan a lo largo del eje horizontal o primer eje, de acuerdo con la importancia que tiene la agricultura en el P.N.B. Dado que en este eje existe una oposición de signo entre AGRI e INDU y SERV, podríamos pensar que en este eje agrícola se oponen los países indutrializados situados a la izquierda y los países no industrializados situados a la derecha. Un país se situará tanto más a la izquierda cuanto menos importancia tenga la agricultura en su PNB.

Además este eje recoge un 81.062%, lo que indica que se lleva casi toda la variabilidad de la nube de puntos, pudiendo considerarlo como un buen factor a la hora de clasificar los países. La oposición entre INDU positivio y SERV negativo en el 2.° eje, nos va a situar a los puntos fila o países tanto más arriba cuanto más importancia tenga en el P.N.B. la actividad Industrial frente a los Servicios. Así vemos que Bélgica y Alemania, con una situación muy similar sobre el eje horizontal, se oponen en el eje vertical con lo que siendo países con una importancia relativa semejante en cuanto a su componente agrícola, en uno de ellos priman más los Servicios y en el otro la Industria.

Muchas veces no es suficiente con dar un interpretación a los ejes que definen un plano para establecer una acertada clasificación de los puntos que se sitúan sobre el mismo.

Aunque éste no es nuestro caso por tratarse de una representación que recoge el 100% de la inercia, en otras muchas situaciones los planos que interpretamos no recogen toda la variabilidad y una vez hecha la interpretación de los ejes a partir de las contribuciones absolutas de los puntos a cada eje es necesario preguntarse por lo bien o mal que éstos definen a los puntos situados sobre ellos. Para medir la calidad de la representación de los puntos situados sobre un eje consideraremos las contribuciones relativas de un eje en la definición de un punto.

La contribución relativa de un eje en la explicación de un punto j, la definimos como la proporción de la inercia total del punto definida por el eje. Si $d^2(j, G)$ es la inercia del punto j, donde G es el centro de la nube, y $d^2_\alpha(j, G)$ es la inercia del punto j sobre el eje α-ésimo, definimos como contribución relativa del eje α-ésimo al punto j.

$$C_r(j, \alpha) = \frac{d^2_\alpha(j, G)}{d^2(j, G)}$$

Evidentemente se cumple:

$$\sum_\alpha C_r(j, \alpha) = 1$$

En las tablas 6 y 7 aparecen las contribuciones relativas a los dos ejes extraídos a los puntos fila y columna, respectivamente.

TABLA 6. Contribuciones relativas de los 2 primeros ejes a los puntos-fila.

	FACTOR 1	FACTOR 2
ARGE	85	915
BOLI	765	235
BRAS	919	81
CHIL	913	87
COLO	955	45
COST	787	213
ECUA	210	790
SALV	991	9
GUAT	924	76
HOND	964	36
MEJI	996	4
NICA	995	45
PANA	838	162
PARA	993	7
PERU	573	427
REDO	787	213
URUG	240	761
VENE	788	212
EEUU	661	339
CANA	589	411
ALEM	917	83
BELG	824	176
DINA	469	531
ESPA	250	750
FRAN	644	356
ITAL	913	87
PBAJ	843	157
PORT	83	917
INGL	774	226
JAPO	998	2

TABLA 7. Contribuciones relativas de los 2 primeros ejes a los puntos-columna.

	FACTOR 1	FACTOR 2
AGRI	985	15
INDU	613	387
SERV	103	897

Es importante considerar estos valores cuando se tratan de establecer grupos de puntos con comportamiento similar respecto de un par de ejes interpretados previamente. Estos (los ejes) vienen definidos por aquellos puntos que se sitúan en el mismo o, dicho de otra manera, que la suma de las contribuciones relativas de los dos ejes al punto es suficientemente grande. En este ejemplo, dado que los puntos se pueden representar en dos ejes, la suma de la contribución relativa de los dos ejes y para todos los puntos es 1.000; por lo tanto, todos están bien representados y no hay problema de interpretación. No será este el caso cuando se necesiten más de dos ejes para la representación (véase ejemplo de aplicación de las correspondencias múltiples).

3.4. Correspondencias múltiples

El método que hemos desarrollado para el estudio del comportamiento conjunto de dos variables categóricas se puede generalizar al estudio de más de dos variables.

Esta generalización se puede hacer basándonos en el hecho de que el estudio de la tabla F a través de un análisis de correspondencias es equivalente al análisis de correspondencia realizado sobre la tabla de BURT generada por ésta.

La tabla de BURT, construida a partir de las variables categóricas X e Y, toma la siguiente forma:

$$B = \begin{bmatrix} n_{1.} & 0 & \cdots & 0 & n_{11} & n_{12} & \cdots & n_{1q} \\ 0 & n_{2.} & & & n_{21} & n_{22} & \cdots & n_{2q} \\ \vdots & & \ddots & & \vdots & & & \\ 0 & \cdots & \cdots & n_{p.} & n_{p1} & n_{p2} & \cdots & n_{pq} \\ n_{11} & n_{21} & \cdots & n_{p1} & n_{.1} & 0 & \cdots & 0 \\ \vdots & & & & 0 & n_{.2} & \cdots & 0 \\ \vdots & & & & \vdots & \vdots & & \vdots \\ n_{1q} & n_{2q} & \cdots & n_{pq} & 0 & 0 & \cdots & n_{.q} \end{bmatrix}$$

Que es una matriz de orden $(p + q, p + q)$ y simétrica.

Donde:

$$n_{i.} = \sum_{j=1}^{q} n_{ij} \qquad i = 1, \ldots, p$$

$$n_{.j} = \sum_{i=1}^{p} n_{ij} \qquad j = 1, \ldots, q$$

Al realizar un análisis de correspondencias sobre B y por ser simétrica la matriz, las dos nubes de puntos filas y columnas son idénticas. Por tanto será suficiente con estudiar una sola nube de puntos.

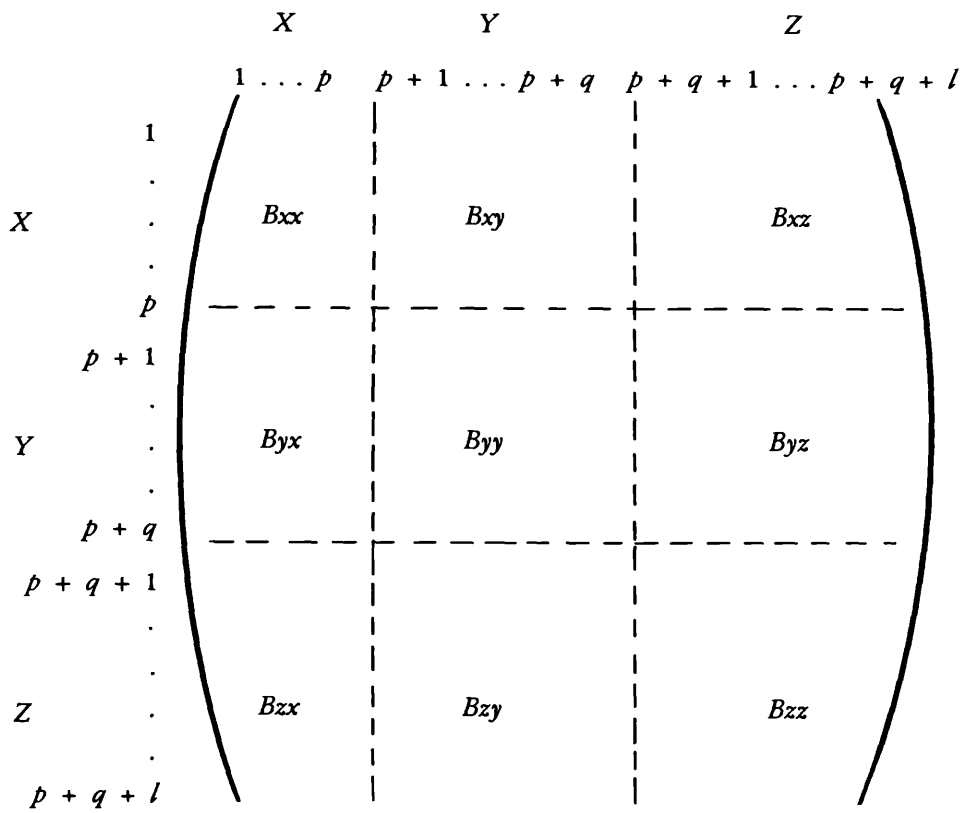
Se cumple el siguiente resultado:

Para todo par de factores $(\varphi_\alpha, \psi_\alpha)$ asociados a un mismo valor propio obtenido de efetuar un análisis de correspondencias sobre las variables X e Y, le corresponde un factor $\phi_\alpha = \begin{pmatrix} \varphi_\alpha \\ \psi_\alpha \end{pmatrix}$ obtenido de efectuar un análisis de correspondencias sobre la matriz B.

La anterior equivalencia nos conduce a la generalización del método cuando queremos estudiar el comportamiento conjunto de más de dos variables. Para ello se procederá a formar la tabla de BURT generada por dichas variables y a realizar un análisis de correspondencia sobre ella. Esta, análogamente construida a partir de dos variables, estará formada por bloques que correspondan a las tablas de contingencia entre los posibles pares de variables.

Por ejemplo, consideremos la tabla de BURT formada por las variables X, Y y Z con p, q y l categorías respectivamente.

La matriz de BURT asociada viene dada por:



donde los bloques diagonales *Bxx*, *Byy*, *Bzz* corresponden a matrices diagonales que contienen los efectivos existentes en cada categoría de las variables *x*, *y*, *z*, respectivamente.

Los bloques no diagonales recogen las tablas de contingencia; así el bloque *Bxy* es la tabla de contingencia formada a partir de *X* e *Y*.

A la hora de interpretar los factores extraídos en un análisis de correspondencias múltiples son válidas las relaciones a las que llegábamos y siguen manteniendo su sentido las contribuciones relativas y absolutas como ayuda a la interpretación de los factores.

3.5. Ejemplo de aplicación de las correspondencias múltiples

Con el fin de ilustrar la capacidad descriptiva que esta técnica tiene en situaciones multivariables, vamos a comentar los resultados que se obtienen de aplicarla a la siguiente situación[1]. Se ha recogido información sobre un total de 400 refugiados políticos residentes en Madrid y Barcelona (tabla 8).

TABLA 8. Variables y categorías del análisis.

• Lugar de residencia	BARC	(Barcelona)
	MADR	(Madrid)
• Edad	20-25	(Entre 20 y 25 años)
	26-30	(Entre 26 y 30 años)
	31-35	(Entre 31 y 35 años)
	36-50	(Entre 36 y 50 años)
	+ 50	(más de 50 años)
• Nacionalidad	ARGE	(Argentina)
	CUBA	(Cuba)
	CHIL	(Chile)
	URUG	(Uruguay)
	OTRO	(Otros)
• Situación jurídica	TURI	(Turismo)
	PERM	(Permanencia)
	RESI	(Residente)
	REFE	(Refugiado español)
	ACNU	(ACNUR)
	NACI	(Nacionalidad)
• Actividad	TRAB	(Trabajo)
	DESE	(Desempleo)
	PENS	(Pensionista, retirado)
	ESTU	(Estudiante)
	AMCA	(Ama de casa)

[1] Los datos que vamos a utilizar provienen de un estudio sobre la emigración latinoamericana a España, realizado por Miguel Roig y Olga Lutz, quienes amablemente nos han permitido hacer uso de sus datos.

TABLA 9. Tabla de BURT asociada a las variables del análisis.

	BARC	MADR	2025	2630	3135	3650	+50	ARGE	CUBA	CHIL	URUG	OTRO	TURI	PERM	RESI	REFU	ACNU	NACI	TRAB	DESE	PENS	RETI	ESTU
BARC	125.	0.	24.	42.	24.	29.	4.	54.	16.	36.	16.	2.	28.	14.	39.	1.	6.	33.	94.	12.	1.	11.	4.
MADR	0.	268.	38.	59.	62.	101.	6.	77.	119.	47.	19.	5.	19.	39.	90.	39.	10.	67.	150.	52.	3.	26.	21.
2025	24.	38.	62.	0.	0.	0.	0.	22.	20.	13.	5.	2.	12.	9.	18.	11.	2.	8.	25.	12.	0.	20.	2.
2630	42.	59.	0.	104.	0.	0.	0.	41.	24.	20.	15.	4.	16.	19.	31.	11.	4.	21.	60.	22.	1.	10.	4.
3135	24.	62.	0.	0.	89.	0.	0.	32.	30.	20.	6.	1.	8.	8.	36.	8.	2.	27.	62.	10.	0.	4.	9.
3650	29.	101.	0.	0.	0.	131.	0.	36.	61.	25.	9.	0.	11.	15.	41.	13.	8.	41.	93.	20.	2.	2.	10.
+50	4.	6.	0.	0.	0.	0.	10.	3.	1.	5.	0.	1.	1.	2.	5.	0.	0.	2.	6.	1.	1.	1.	1.
ARGE	54.	77.	22.	41.	32.	36.	3.	136.	0.	0.	0.	0.	22.	11.	42.	14.	4.	40.	106.	8.	0.	11.	3.
CUBA	16.	119.	20.	24.	30.	61.	1.	0.	136.	0.	0.	0.	7.	20.	44.	27.	4.	33.	66.	40.	1.	8.	19.
CHIL	36.	47.	13.	20.	20.	25.	5.	0.	0.	83.	0.	0.	12.	15.	33.	1.	3.	17.	48.	11.	1.	12.	4.
URUG	16.	19.	5.	15.	6.	9.	0.	0.	0.	0.	35.	0.	6.	5.	10.	1.	2.	11.	25.	4.	1.	4.	0.
OTRO	2.	5.	2.	4.	1.	0.	1.	0.	0.	0.	0.	8.	1.	2.	2.	0.	3.	0.	3.	2.	1.	2.	0.
TURI	28.	19.	12.	16.	8.	11.	1.	22.	7.	12.	6.	1.	48.	0.	0.	0.	0.	0.	29.	10.	1.	5.	1.
PERM	14.	39.	9.	19.	8.	15.	2.	11.	20.	15.	5.	2.	0.	53.	0.	0.	0.	0.	21.	12.	0.	13.	4.
RESI	39.	90.	18.	31.	36.	41.	5.	42.	44.	33.	10.	2.	0.	0.	131.	0.	0.	0.	89.	16.	2.	9.	9.
REFE	1.	39.	11.	11.	8.	13.	0.	14.	27.	1.	1.	0.	0.	0.	0.	43.	0.	0.	15.	17.	0.	2.	8.
ACNU	6.	10.	2.	4.	2.	8.	0.	4.	4.	3.	2.	3.	0.	0.	0.	0.	16.	0.	7.	8.	0.	0.	0.
NACI	33.	67.	8.	21.	27.	41.	2.	40.	33.	17.	11.	0.	0.	0.	0.	0.	0.	101.	85.	2.	1.	6.	3.
TRAB	94.	150.	25.	60.	62.	93.	6.	106.	66.	48.	25.	3.	29.	21.	89.	15.	7.	85.	248.	0.	0.	0.	0.
DESE	12.	32.	12.	22.	10.	20.	1.	8.	40.	11.	4.	2.	10.	12.	16.	17.	8.	2.	0.	66.	0.	0.	0.
PENS	1.	3.	0.	1.	0.	2.	1.	0.	1.	1.	1.	1.	1.	0.	2.	0.	0.	1.	0.	0.	4.	0.	0.
RETI	11.	26.	20.	10.	4.	2.	1.	11.	8.	12.	4.	2.	5.	13.	9.	2.	0.	6.	0.	0.	0.	37.	0.
ESTU	4.	21.	2.	4.	9.	10.	1.	3.	19.	4.	0.	0.	1.	4.	9.	8.	0.	3.	0.	0.	0.	0.	26.

Con este ejemplo se pone de manifiesto la capacidad descriptiva del método, sobre todo en su aspecto asociador de categorías de diferentes variables cuando éstas son suficientes en número para hacer difícil un estudio global de ellas.

Siguiendo los pasos desarrollados en el apartado de generalización del método, en primer lugar construimos la tabla de BURT asociada a las variables tratadas (ésta aparece en la tabla 9), sobre la cual se procede a realizar un análisis de correspondencias.

El segundo paso, diagonalización de la matriz asociada, nos da los autovalores asociados a cada uno de los factores, que interpretábamos como la importancia que cada uno de ellos tiene respecto de la información que recoge tanto respecto del total como en relación a los otros. En la tabla 10 aprecen los autovalores asociados a los factores, ordenados de mayor a menor juntamente con el porcentaje respecto de la suma, o inercia total y el porcentaje acumulado.

TABLA 10. Autovalores asociados a los factores.

AUTOVALORES	PROCENTAJES	ACUMULADOS
0.14245939	17.069	17.069
0.10945886	13.115	30.184
0.07598025	9.104	39.287
0.07019061	8.410	47.697
0.05333618	6.390	54.087
0.05006146	5.998	60.085
0.04787207	5.736	65.821
0.04522061	5.418	71.239
0.03957317	4.741	75.981
0.03640264	4.362	80.342
0.03229317	3.869	84.211
0.02864569	3.432	87.644
0.02508798	3.006	90.650
0.02048729	2.455	93.104
0.01962502	2.351	95.456
0.01703500	2.041	97.497
0.01119899	1.342	98.838
0.00961920	1.153	99.991
0.00005734	0.007	99.998
0.00001375	0.002	99.999
0.00000394	0.000	100.000
0.00000015	0.000	100.000

A la vista de estos resultados se puede afirmar:

1. No existe ningún factor que explique un porcentaje alto de la inercia; el primero sólo explica un 17% de la misma.

2. Para obtener una explicación de un 95% de la inercia se necesitarán 15 factores, lo cual hace pensar en la imposibilidad de hacer una reducción efectiva de la dimensionalidad del problema, aunque como resultado subsidiario, pero importante, podemos considerar cuantificado nuestro problema descrito en términos de variables cualitativas a un problema descrito en función de variables cuantitativas o numéricas que serían los factores.

Todo lo anterior nos hace ir con las mayores reservas al aspecto descriptivo de las asociaciones entre las diferentes categorías, ya que la metodología que empleamos es la observación de los mapas o representaciones gráficas de las proyecciones de las categorías sobre parejas de factores con porcentajes de la inercia total altos.

A partir de ahora vamos a fijarnos en las representaciones que se obtienen de considerar las parejas formadas por el 1.° y 2.° factor, 1.° y 3.° factor y 2.° y 3.° factor que aparecen en las figuras 4, 5 y 6, respectivamente, conjuntamente con la información que aparece en la tabla 11 que corresponde a los valores de las coordenadas, contribuciones relativas y absolutas de las categorías de las distintas variables sobre estos tres primeros ejes.

TABLA 11. (a) Coordenadas, (b) contribuciones relativas y (c) absolutas de las categorías en los tres primeros factores.

	(a)	(b)	(c)	(a)	(b)	(c)	(a)	(b)	(c)
	1#F	COR	CTR	2#F	COR	CTR	3#F	COR	CTR
1 BARC	586	622	153	-60	7	2	3	0	0
2 MADR	-267	605	69	27	6	1	3	0	0
3 2025	50	2	1	-673	355	131	547	234	124
4 2630	219	76	18	-239	91	28	-127	26	11
5 3135	-6	0	0	335	152	47	120	20	9
6 3650	-227	114	24	297	193	54	-183	75	30
7 +50	441	24	7	-342	15	6	-664	56	30
8 ARGE	327	233	51	173	65	19	132	38	16
9 CUBA	-617	762	188	61	7	2	15	0	0
10 CHIL	293	101	25	-176	37	12	40	2	1
11 URUG	447	93	25	-110	6	2	-207	20	10
12 OTRO	197	3	1	-1660	240	102	-1922	322	197
13 TURI	559	194	54	-337	71	26	75	4	2
14 PERM	-119	10	3	-587	244	86	185	24	12
15 RESI	33	3	1	123	36	9	-23	1	1
16 REFE	-970	466	144	-125	8	3	267	35	21
17 ACNU	-165	5	2	-612	70	28	-1474	408	234
18 NACI	187	53	13	474	339	106	5	0	0
19 TRAB	215	319	41	265	484	82	-33	8	2
20 DESE	-620	311	91	-486	192	73	-273	61	33
21 PENS	314	5	1	-564	16	6	-2111	219	122
22 RETI	239	25	8	-993	424	171	722	223	130
23 ESTU	-924	266	80	211	14	5	303	29	16

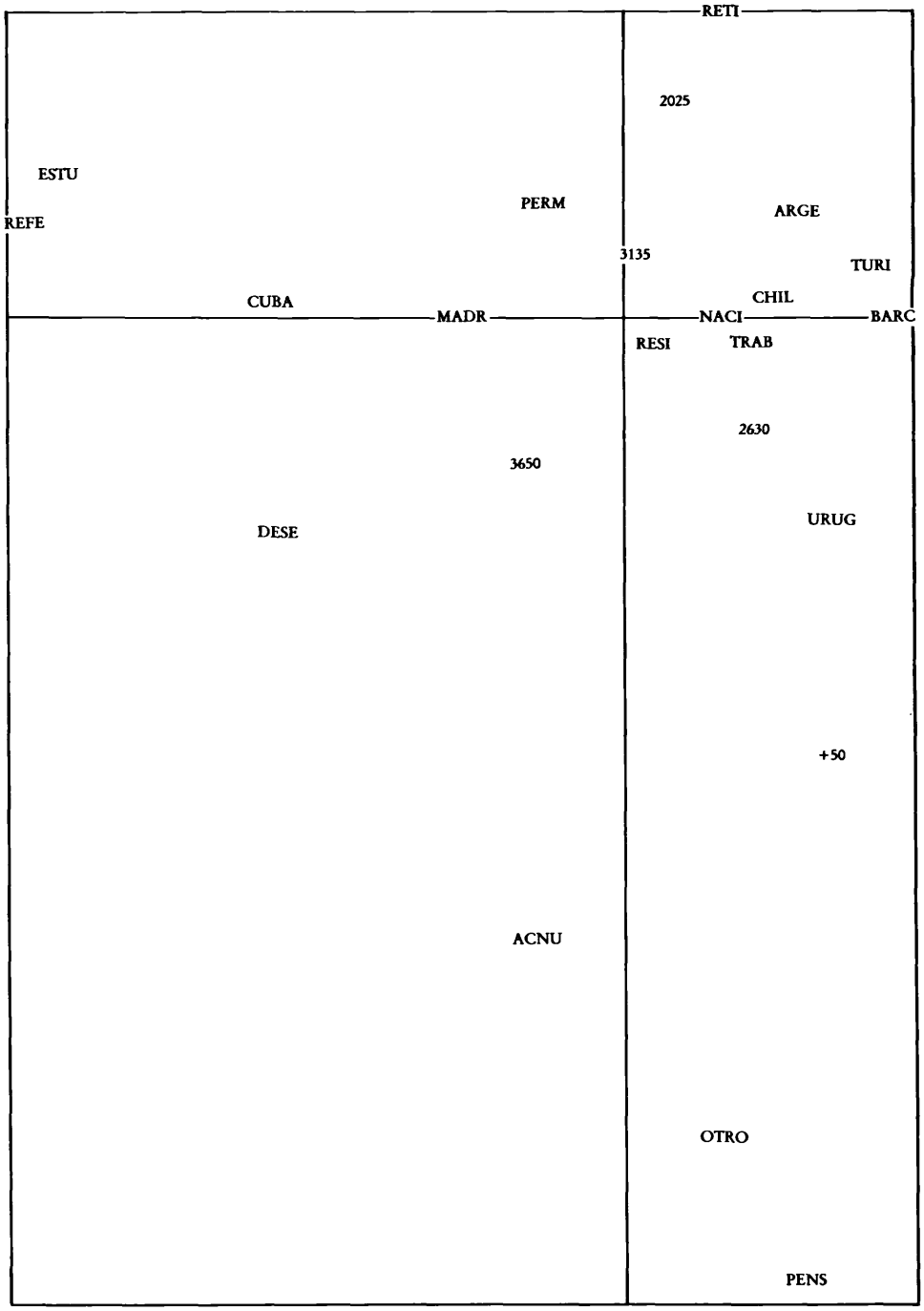


FIGURA 4. Proyección de las categorías sobre los dos primeros ejes: 1.er factor (eje horizontal) y 2.º factor (eje vertical).

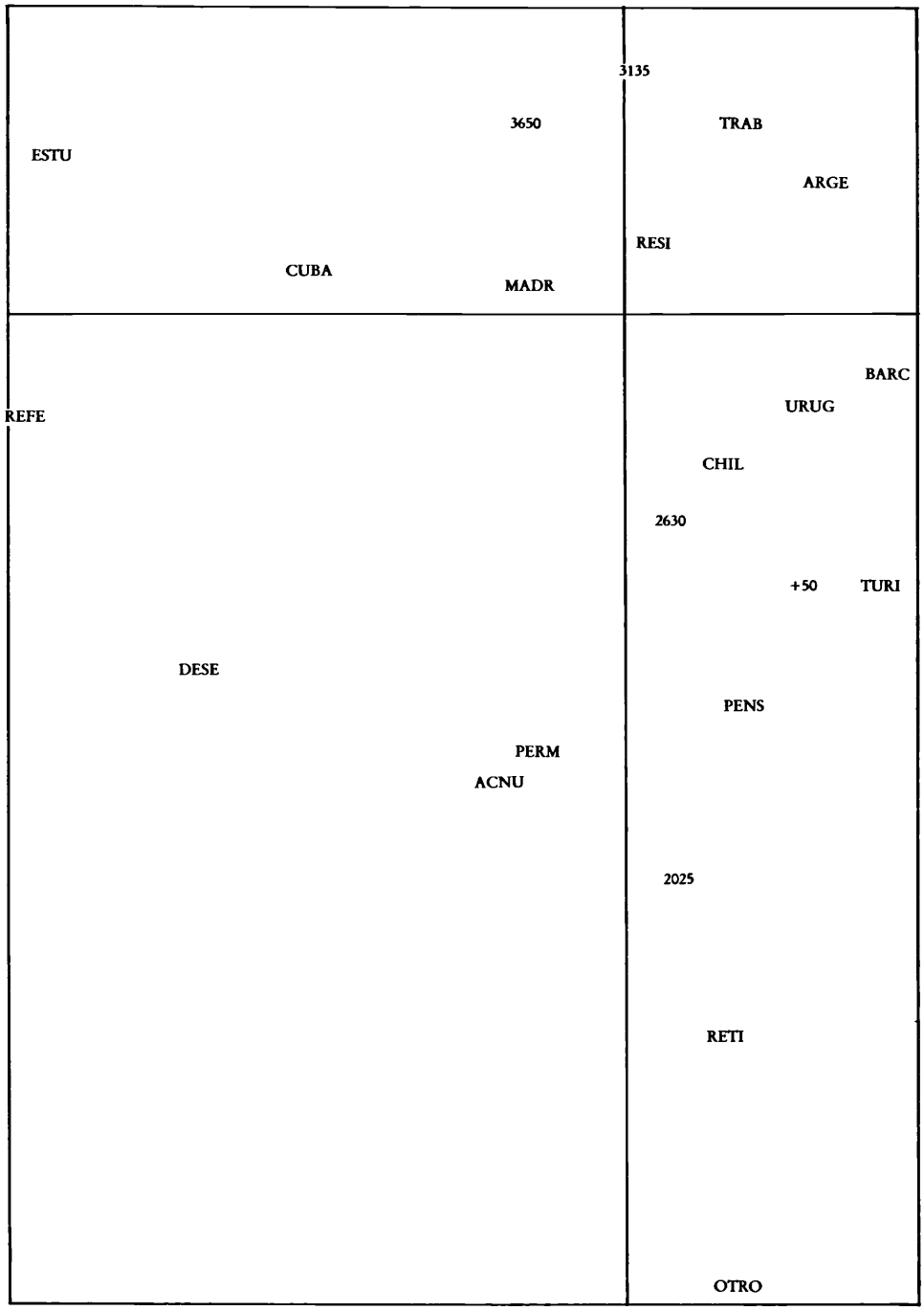


FIGURA 5. Representación de las categorías en el plano definido por el 1.er factor (eje horizontal) y el 3.er factor (eje vertical).

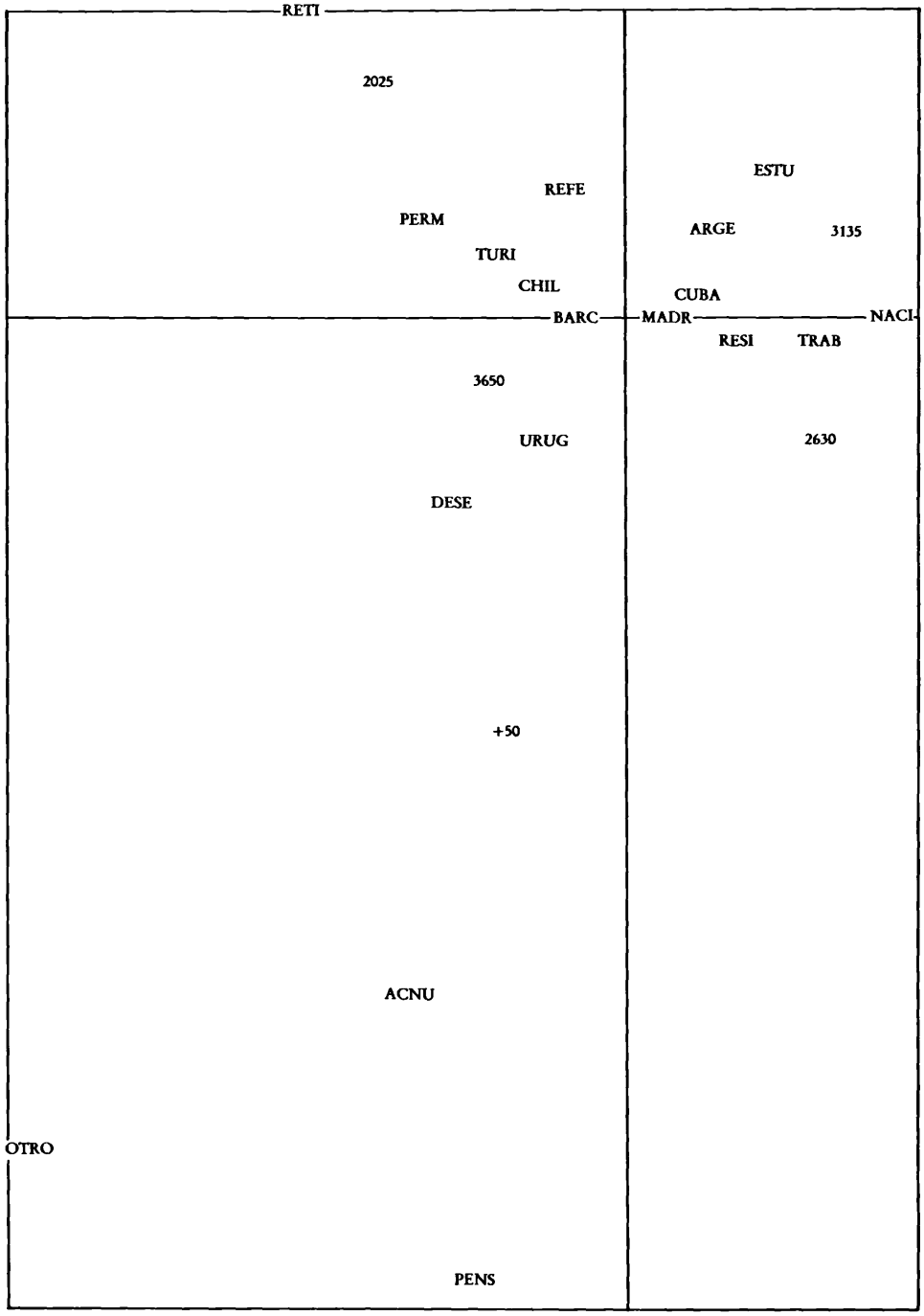


FIGURA 6. Representación de las categorías en el plano definido por el 2.º factor (eje horizontal) y 3.er factor (eje vertical).

Antes de proceder a realizar cualquier interpretación, o establecer cualquier hipótesis, tendremos que tener presente que el poder explicativo de los 3 factores es de un 39.28%, siendo el de cada uno de ellos un 17%, 13.11% y 9.10%, respectivamente. Con lo cual la deformación de las representaciones planas de consideramos son muy altas respecto de las representaciones de puntos en los espacios de partida.

Fijaremos nuestra atención en la figura 4, correspondiente a la representación de las categorías a partir de sus valores en los dos primeros factores, así como en sus correspondientes valores numéricos que aparecen reflejados en la tabla 11.

Algunas de las conclusiones a las que llegamos son:

1. Dado que la distancia utilizada pondera inversamente la frecuencia absoluta, es conveniente eliminar aquellas categorías que tienen frecuencias bajas, pues su inercia se hace muy grande y provoca deformaciones en los gráficos al situar siempre las categorías en los extremos de las representaciones.

En nuestro ejemplo será conveniente eliminar las categorías +50, OTRO, ACNU y PENS pues sus frecuencias absolutas son de 10, 8, 16 y 4, respectivamente (ver diagonal en la tabla 9).

2. Si nos fijamos en el comportamiento de dos categorías como, por ejemplo, ARGE y CUBA, que tienen el mismo peso (sus frecuencias son 136), vemos que CUBA está bastante bien representada en el 1.er factor: tiene una contribución relativa de 762 sobre 1.000 lo que hace que su deformación sea pequeña.

Las categorías que mejor definen a este eje vendrán dadas como resultado de analizar la columna correspondiente a las contribuciones absolutas. Podemos establecer la siguiente oposición entre grupo de categorías con el fin de dar una posible interpretación a este factor (véase figura 4 y tabla 11):

		MADR	
BARC	ARGE		REFE
	TURI		DESE
	TRAB		ESTU
		CUBA	

Podríamos decir que, por ejemplo, se establece una oposición entre MADRid y BARCelona TURI y REFE.

Estando los cubanos situados en la parte izquierda, podríamos decir que los cubanos están descritos en términos de refugiados (REFE), siendo estudiantes (ESTU) o desempleados (DESE), habitando preferentemente en Madrid (MADR) y estando muy pocos de turismo (TURI).

Esta primera aproximación a la definición de exiliado político cubano es exploratoria y se trataría de dar una confirmación posterior a través de técnicas más específicas (por ejemplo, el ajuste de algún modelo).

Si intentásemos hacer lo mismo con la categoría ARGE, tropezaríamos con las siguientes dificultades:

La calidad de representación que tiene este punto sobre el eje 1.° es de 233 (frente a 762 para CUBA), con lo cual nos hace ser más precavidos a la hora de intentar definirlo a partir de las categorías que definen el 1.er eje. Tenemos que recurrir a más

de un eje (o a otro eje) para su definición; no siendo ni tan siquiera suficiente con los dos primeros ejes para obtener una buena representación, pues ésta sería solamente de 298.

No ocurre lo mismo con las categorías MADR y BARC de la 1.ª variable, pues ambas tienen una alta calidad de representación sobre el primer eje (605 y 622, respectivamente), teniendo que:

BARC	está atraída por	ARGE
		TURI
		TRAB
MADR	está atraída por	REFE
		DESE
		ESTU
		CUBA

siendo ésta una hipótesis a comprobar en una etapa posterior del análisis y con métodos más apropiados.

De esta forma vamos separando el comportamiento de las categorías que seleccionemos, bien con un interés predefinido bien valiéndonos de sus posiciones en los gráficos. Una vez seleccionados los puntos o categorías consideraremos aquellos ejes que mejor les representan (aquellos donde los puntos tienen contribuciones relativas altas) y explicaremos las categorías en función de los puntos que mejor definen dichos ejes (aquellos que tienen contribuciones absolutas altas).

4. Escalas Multidimensionales *

Por *A. P. M .Coxon y C. L. Jones*
Traducción: *J. L. Muñoz Yanguas*

4.1. Introducción

Nuestro propósito en este capítulo es triple:

1. Mostrar que una representación geométrica proporciona un marco de trabajo útil y natural para analizar los datos de las ciencias sociales en general, y los de las encuestas en particular.
2. Proporcionar una visión panorámica de los modelos de escalonamiento multidimensional (a partir de aquí denominados EMD), destacando los modos en que se ha generalizado el enfoque original para enfrentarse con una amplia variedad de modelos y tipos de datos.
3. Ilustrar sobre la aplicabilidad de tales modelos a problemas de análisis que los investigadores de encuestas afrontan rutinariamente, mediante el estudio de las respuestas de una muestra grande y heterogénea a 21 ítems sacados de un cuestionario sobre actitudes hacia el trabajo, con ayuda del modelo INDSCAL de Carroll y Chang.

4.1.1. *Los datos de encuesta*

Los métodos de escalonamiento se conciben a menudo como parte del arsenal de técnicas descriptivas del analista de encuestas. El escalonamiento multidimensional (EMD) puede contemplarse como una extensión de métodos ya familiares en la investigación por encuesta, como el análisis del escalograma de Guttman, el principo de desdoblamiento de Coombs *(unfolding principle)*, y el análisis factorial. Los problemas fundamentales del escalonamiento se derivan de la construcción y comprobación

* *Nota de los autores:* Este capítulo es la traducción de un trabajo de los autores escrito en 1975 y publicado en 1977. Desde entonces ambos autores han trabajado en los métodos de las Escalas Multidimensionales con el doble objetivo de ilustrar la utilidad del enfoque en el análisis de los datos sociales, y de mostrar que las EMD no son sólo un método exploratorio sino que también puede ser confirmatorio, en el sentido de que puede contrastar hipótesis. Tanto el editor como los autores coinciden en que este artículo proporciona una introducción amplia y no técnica que resulta satisfactoria. Algunos trabajos más recientes de los autores pueden ser de interés Coxon, A. P. M. y C. L. Jones, «Multidimensional scaling» en D. McKay, N. Schofield y P. Whitley (Eds.). *Data analysis and the Social Sciences.* Londres, France Pinter, pp. 171-225, 1983. Coxon, A. P. M. y C. L. Jones, *The images of occupational prestige,* Londres, MacMillan, 1978. Coxon, A. P. M., *The user's guide to multidimensional scaling.* Londres, Heinemann, 1982.

de los modelos, aunque los que se tratan aquí son diferentes de los «modelos causales» actualmente de moda en el análisis centrado en variables. Sin embargo tenemos que hacer una advertencia previa. Las propiedades muestrales de casi todos los modelos y procedimientos EMD son poco conocidas, aunque se han realizado avances sobre el problema (Barlow *et al.*, 1972). Por lo tanto, los procedimientos EMD deben usarse con la precaución reservada para cualquier herramienta potente pero poco familiar.

Después de llevar a cabo una encuesta sobre personas, viviendas, firmas comerciales, u otras unidades, el investigador a menudo deseará describir las características del individuo «típico» o «medio» (o vivienda, firma, etc.). También deseará dar a sus lectores alguna información sobre el margen de variación alrededor de esta «media». Mientras se ocupe de una sola variable, como el tamaño de la familia o el valor de los activos líquidos de una empresa, le será posible describir el margen y el tipo de variación por medio de los índices habituales de tendencia central y dispersión, junto con métodos gráficos para mostrar las formas de las distribuciones de frecuencias. Pero así las variables se presentan de una en una, de manera independiente, y, puesto que los científicos sociales se ocupan de relaciones estructuradas entre variables, no siempre encontrarán útil esta clase de procedimiento. A menudo el principal problema es hacer enunciados significativos sobre la forma en que los sujetos de una encuesta se parecen los unos a los otros (se agrupan) e inversamente sobre los modos en que difieren, y esto con respecto a conjuntos de variables y no a una variable cada vez. Los economistas hacen referencia al problema como «la cuestión de la agregación», y los psicólogos lo llaman «problema de las diferencias individuales». En los dos casos es lo mismo: si los sujetos difieren entre sí sistemáticamente en términos de procesos que subyacen en los datos, no será posible recuperar esta información después de haber sido sumariamente agregados. En los estudios exploratorios y no experimentales, donde lo que importa es la forma general de los procesos subyacentes más que los detalles precisos, debemos prestar especialmente atención a la forma de presentarse las «diferencias individuales» y a su alcance, ya que probablemente afecten a nuestras deducciones. No todos los modelos de escalonamiento multidimensional se preocupan por presentar las diferencias individuales, pero creemos que aquellos que puedan hacerlo se utilizarán progresivamente por los investigadores de encuestas, ya que proporcionan un marco de trabajo sobre el problema de la agregación en función del cual las pautas existentes en los datos individuales (o de empresas, etc.) permanecen representadas en descripciones suscintas. Los enfoques en el escalonamiento de las diferencias individuales deberían afinar estas descripciones, haciéndolas más sensibles y al mismo tiempo más precisas, distinguiendo los atributos que tengan en común todos los sujetos de la encuesta de aquellos en los que difieran de un modo estructurado, y a su vez también de la variación aleatoria.

4.1.2. *Escalas Multidimensionales*

Las Escalas Multidimensionales (EMD) tienen una historia más corta que el análisis factorial, pero en su versión original eran igualmente restrictivas. Para el «EMD clásico» (expuesto en detalle en Torgerson, 1958) los datos básicos normalmente consistían en juicios sobre la semejanza o diferencia entre todas y cada una de las pa-

rejas posibles sacadas de un conjunto de estímulos, y se obtenían sobre la percepción de colores, donde las respuestas se obtenían por el llamado «método completo de las tríadas». Se presentaban las 84 tríadas que se pueden formar con 9 estímulos a cada sujeto, y se le pedía que dijera cuál de los dos colores en cuestión era más parecido al tercero. Esto es una muestra de una de las principales desventajas del «EMD clásico»: la gran cantidad de datos que se recogen sistemáticamente de cada sujeto. En el estudio de Torgerson sobre la percepción de los colores se recogieron datos completos para 38 sujetos, y se agregaron de forma que para cada color había una estimación empírica de la proporción de veces que era apreciado como más similar a un color *i* que a uno *j*. Asumiendo que sea útil tratar de representar tales datos con un modelo de distancia, estas proporciones se interpretan como si estuvieran relacionadas con distancias en un espacio de una, dos, tres o más dimensiones. La analogía más obvia aquí es la de un mapa de carreteras ordinario, donde Blackpool está en millas más cerca de Londres que de Edimburgo, lo cual se puede descubrir comparando la distancia entre Blackpool y Londres con la distancia entre Londres y Edimburgo, o sea, observando una diferencia entre distancias. Así el «EMD clásico» estima las distancias absolutas entre todos los pares de estímulos de los datos originales por métodos de mínimos cuadrados que ahora no nos interesan. Volviendo a la analogía de los mapas de carreteras, el resultado era equivalente a esa útil tabla que muestra la distancia en millas entre cualesquiera dos ciudades. El «EMD clásico» procede desde aquí convirtiendo la matriz de distancias absolutas en una matriz de productos escalares a los que se trata como covarianzas, sometiéndolos a un análisis factorial (con todos los problemas habituales sobre las dimensiones, etc.) para obtener las proyecciones de los estímulos sobre tantos ejes ortogonales como el usuario desee. En la analogía del mapa de carreteras, los datos sobre distancias entre parejas de ciudades se convierten, sin pérdida de información, en una «solución» que consiste en la posición de cada ciudad con respecto a dos ejes de referencia. En el estudio de Torgerson, los nueve colores usados como estímulos eran todos del mismo tinte rojo, pero diferentes en términos de brillo y saturación. Los juicios de similaridad de los sujetos reflejaron este hecho, ya que fue posible rotar la solución de dos factores de forma que las proyecciones de los estímulos sobre los dos factores reflejasen los valores (conocidos de antemano) de los estímulos en brillo y saturación, respectivamente.

Debería quedar ya claro que los métodos del «EMD clásico» acababan sobradamente con la paciencia de los sujetos experimentales, y también que implicaban grandes cantidades de cálculo tedioso. Los desarrollos en este área habrían permanecido en el puro interés académico de no haber sido por dos hechos. Primero, que en los primeros años 60 se produjeron un cambio metodológico de enfoque y una ruptura técnica, que convirtieron a los procedimientos EMD en aptos para analizar grandes cantidades de datos de cualquier nivel de medida; y segundo, el modelo básico se amplió rápidamente para tratar los tipos de datos comúnmente encontrados por los analistas de encuestas. El punto de vista metodológico, asociado en lo principal con Coombs, se desarrolla en detalle en *Una teoría de lo datos* (*A Theory of data,* 1964) y su defensa de las aproximaciones no-métricas a la medida puede resumirse como sigue:

1. Los supuestos que se hacen sobre el nivel de medida de los datos y los que se contienen en los modelos de escalonamiento aplicados para analizarlos no son meras

ficciones «ad hoc», sino que implican hipótesis sustantivas sobre el comportamiento humano.

2. Al atribuir propiedades métricas a los datos de las ciencias sociales, es mejor errar del lado conservador, utilizando escalas de medidas más débiles para representarlos. También es preferible un acercamiento cauteloso cuando se establezcan supuestos sobre las distribuciones de la población en los datos multivariables.

3. Ya que la mayoría de los datos en ciencias sociales se extraen en situaciones no experimentales, y se refieren a menudo a poblaciones «diversificadas» o heterogéneas, es bueno ser especialmente sensibles a las diferencias individuales, ya que pueden ser cruciales en la interpretación de nuestros datos.

La ruptura técnica se produjo cuando Shepard (1962, 1966) consiguió demostrar que cuando se imponen en número suficiente constricciones simplemente no-métricas, esto es, informaciones ordinales en los datos, se ponen límites verdaderamente muy estrechos a las soluciones posibles. («Solución» significa en este contexto las proyecciones de los puntos-estímulo sobre un conjunto de ejes de referencia: por ejemplo, las ordenadas y abscisas de los puntos). Los límites en la solución, impuestos por existir un número tan elevado de constreñimientos ordinales, son tan estrictos que resulta posible identificar la solución métrica óptima (proyecciones de los puntos-estímulo sobre los ejes de referencia en un nivel de medida interval).

4.2. Escalonamiento Multidimensional no métrico

La mayor ventaja del EMD no métrico es que el mismo procedimiento básico se puede extender sin dificultad a muy distintos tipos de datos y a diferentes modelos (además del usual de distancias). Se puede calibrar la popularidad de esta extensión por el hecho de que pronto aparecieran aplicaciones en disciplinas tan diversas como la Electrónica (Kruskal, 1966) y la Historia (Hodson *et al.*, 1971), aunque continúan prevalenciendo las aplicaciones en ciencias sociales.

A diferencia de los modelos estadísticos multivariables convencionales, una ventaja más del EMD no métrico es que las soluciones son invariantes al orden (Sibson, 1972). Es decir, que para todo conjunto de datos con la misma ordenación siempre se produce la misma solución, y los procesos no dependerán en absoluto de la medida de similaridad específicamente usada. Los datos podrán representarse con un modelo de distancia siempre que de alguna manera aporten información sobre la proximidad relativa entre puntos (que es precisamente la forma en que Coombs interpreta la mayoría de los tipos de datos). Ni siquiera se necesita que sean medidas complejas, como las covarianzas y las correlaciones del análisis factorial; son igualmente legítimas otras como los coeficientes de contingencia entre atributos, las co-ocurrencias de palabras en un texto, o las confusions entre sonidos en una audición. Las numerosas maneras de construir índices de semejanza entre individuos o entre variables han sido documentadas por Cormack (1971: p. 324 y sigs.), Jardine y Sibson (1971: parte 1) y Sokal y Sneath (1963: cap. 6). Además se han desarrollado varios modelos EMD que conservan intactos los datos individuales, y esto representa un avance significativo sobre el análisis factorial.

Por ejemplo, si un individuo puntua u ordena un cierto número de estímulos en términos de preferencia (o analogía) sus datos se conservan juntos en el modelo MD-PREF de Carroll y Chang (1964) (véase más adelante un ejemplo). De igual modo, si un individuo emite juicios sobre la semajanza entre cada pareja de un conjunto de estímulos, en el modelo INDSCAL de Carroll y Chang (1970) sus datos se conservan juntos (véase también un ejemplo más adelante). («Conservarse juntos» quiere aquí decir que los modelos de escalas contienen parámetros para representar a cada individuo y a cada punto estímulo).

¿Cómo funciona el EMD no-métrico? Shepard (1966) comienza a explicar los fundamentos considerando el hecho de que la representación numérica de un conjunto de objetos está relativamente indeterminada si sólo se conoce la ordenación de los mismos. Esta indeterminación deriva de que la representación de los puntos en el espacio-solución puede fluctuar bastante (es decir, puede tomar valores numéricos en un rango amplio), sin dejar de satisfacer los constreñimientos puramente ordinales que los datos representan. (Aún no hemos decidido si estos valores van a resultar unidimensionales o no). Sin embargo, toda vez que la localización de los puntos en el espacio solución debe cumplir constricciones suplementarias que se refieren al orden de las distancias entre puntos, entonces el rango de posibles valores se reduce mucho.

«Si se imponen constreñimientos no métricos en cantidad suficiente, estos comienzan a actuar como si fuesen métricos... Cuando se obliga a que los puntos satisfagan más y más desigualdades en las distancias, su colocación se restringe hasta el extremo de que cualquier modificación, por pequeña que sea, normalmente terminará violando una o más desigualdades.» (p. 288).

La noción de que las relaciones de orden entre las distancias imponen severas constricciones a la unicidad de la representación numérica es hoy día un lugar común, pero para que esto se demostrara fehacientemente hubo que esperar al desarrollo de un algoritmo iterativo diseñado para producir una solución que cumpliese el conjunto de condiciones derivadas de los datos.

El fundamento teórico del algoritmo del EMD no-métrico lo proporcionó J. B. Kruskal (1964) y ha constituido la base de casi todo el trabajo posterior en este área. El objetivo de cualquiera procedimiento no-métrico (al menos en los modelos de distancia) es encontrar un conjunto de puntos en un espacio de dimensión mínima, tales que los datos de disimilaridad sean una función monótona (ordenada) de las distancias en este espacio, es decir, que cuando la disimilaridad entre los estímulos i y j sea menor que la disimilaridad entre k y l, la distancia entre i y j será menor o igual que la distancia entre k y l: es decir,

$$\text{cuando} \quad \delta_{ij} < \delta_{kl}, \quad \text{entonces} \quad d_{ij} \leq d_{kL}$$

(lo que se conoce como criterio de orden débil *[weak monotonicity]*), donde δ_{ij} y d_{ij} son respectivamente la disimilaridad y la distancia entre el punto i y el punto j. Una configuración de puntos en un espacio r-dimensional que cumpla este criterio para un conjunto de datos recibe el nombre de solución r-dimensional.

Una clara aplicación del EMD no métrico, en un contexto sociológico, ha sido presentada por McDonald (1972), como parte de una investigación sobre las dimensiones subyacentes a la clasificación de las ocupaciones que hace el censo británico de 1951

en trece grupos socioeconómicos. Ya que estaba interesado en la movilidad ocupacional entre padres e hijos, McDonald basó su índice de disimilaridad (entre cada dos de los trece grupos socioeconómicos) en el cruce de los grupos socioeconómicos de los hijos con los de sus padres. Una tabla de esta clase había sido publicada por Benjamin (1958), y parte de ella se muestra en la tabla 1. La primera fila de la tabla contiene la proporción de hijos de trabajadores agrícolas que acabaron en los diferentes grupos socioeconómicos: el 3.4% de ellos terminaron como granjeros (grupo socioeconómico 1 (a partir de aquí, GSE 1)); el 40.0% siguieron los pasos de sus padres; el 0.6% se convirtieron en altos administrativos (GSE 3); etc. Cada fila de la tabla contiene información sobre las proporciones de hijos que ingresan en cada GSE, siempre que sus padres estuvieran empleados en un grupo dado.

TABLA 1. Porcentaje de hijos que ingresan en trece grupos ocupacionales, para tres grupos de sus padres.

Grupo socioeconómico del padre	Grupo socioeconómico del hijo													Total
	1	2	3	4	5	6	7	8	9	10	11	12	13	100%
2. Trabajadores agrícolas	3.4	40.0	0.6	2.2	2.8	0.0	1.7	2.2	1.7	26.4	5.1	13.5	0.5	178
5. Comerciantes	2.5	2.5	5.7	14.7	27.9	7.4	9.0	0.8	1.6	17.2	3.3	6.6	0.8	122
10. Obreros especialistas	0.3	2.8	2.4	7.2	4.1	5.7	2.5	1.6	3.8	47.0	12.0	9.9	0.6	996

La medida de disimilaridad entre grupos socioeconómicos tal como se usa para clasificar las ocupaciones de los padres es simple, e incluso tosca. Para cada par de filas de la tabla 1, se calculan los valores absolutos de las diferencias entre las respectivas casillas de la tabla. La suma de estos valores absolutos es un índice de disimilaridad entre cualesquiera pares de grupos socioeconómicos de padres, correspondientes a los pares de filas de la tabla 1. Por ejemplo, si tomamos la segunda y la tercera filas de esta tabla (GSE 5 y GSE 10), el índice de disimilaridad sería:

$$|2.5 - 0.3| + |2.8 - 2.5| + |5.7 - 2.4| + |14.7 - 7.2| +$$
$$+ |27.9 - 4.1| + |7.4 - 5.7| + |9.0 - 2.5| + |0.8 - 1.6| +$$
$$+ |1.6 - 3.8| + |17.2 - 47.0| + |3.3 - 12.0| + |6.6 - 9.9| +$$
$$+ |0.6 - 0.8| = 90.2$$

Claramente, si los destinos ocupacionales de los hijos respecto de los padres en el GSE 10 y en el GSE 5 hubieran sido los mismos (en un sentido probabilístico), las dos filas correspondientes de la tabla 1 hubieran tenido elementos idénticos, y el índice

de disimilaridad habría tomado su valor mínimo de cero. Si no hubiera habido ninguna coincidencia parcial se habría obtenido una disimilaridad máxima entre los destinos ocupacionales del GSE 5 y el GSE 10. Cuando se divide por dos, este índice de disimilaridad se puede interpretar como el porcentaje de hijos procedente de una categoría de padres (o de la otra) que habrían tenido que cambiar de destino para que el índice alcanzara su valor mínimo (cero). Dividiendo por dos da 45.1 como índice entre el GSE 5 y el GSE 10; es decir, que el 45.1% de los hijos de comerciantes habrían tenido que cambiar de destino laboral real hacia otros GSE, de forma que así los hijos de comerciantes se equiparan con los de obreros especialistas en términos del destino ocupacional de sus hijos. Este índice de disimilaridad tiene entonces una interpretación: es simétrico; no puede ser menor que cero; solamente vale cero cuando las dos filas son idénticas; y para tres filas cualesquiera se cumple la desigualdad triangular. Por consiguiente, se comportaría como una distancia entre puntos en un espacio métrico. Una vez que todos los pares de disimilaridades hubieran sido ensamblados, podría usarse un procedimiento EMD métrico al modo clásico para obtener una solución de los trece GSE de ocupaciones de los padres, en dos o quizás tres dimensiones. Las diferencias entre las versiones métrica y no métrica del EMD son aproximadamente como sigue. En la primera (EMD métrico), las distancias entre pares de puntos en el espacio solución se relacionan *linealmente* con las disimilaridades entre los correspondientes puntos en los datos. En la segunda (no-métrica) las distancias entre los pares de puntos en el espacio-solución están relacionadas ordinalmente con las disimilaridades entre las correspondientes parejas de puntos en los datos. Como hemos dicho anteriormente, los numerosos constreñimientos ordinales parecen dejar tan poca incertidumbre sobre las posiciones relativas de los puntos como los métricos. McDonald prefirió analizar la matriz de coeficientes de disimilaridad entre grupos socioeconómicos con un procedimiento EMD no-métrico (aunque señala que el análisis con un EMD métrico habría producido resultados muy similares). Las coordenadas de los trece grupos socioeconómicos se muestran en la tabla 2 (solución en dos dimensiones rotadas en ejes principales) y están representadas en la figura 1. Al igual que en el análisis factorial, los ejes de coordenadas (dimensiones) de cualquier solución se pueden rotar rígidamente sin pérdida de información, y tal transformación notacional puede ser de considerable ayuda a la hora de interpretar el patrón general de los puntos en un espacio (como es nuestro caso) bidimensional.

Otro paralelismo entre el análisis factorial y el EMD se refiere al problema de «interpretar» sustantivamente las dimensiones en términos de las proyecciones de los puntos sobre ellas. McDonald trata de interpretar el primer eje principal (dibujado como dimensión horizontal en la figura 1) como la base educativa que los padres de una determinada ocupación proporcionan a sus hijos, y parece claro que la segunda dimensión (vertical en la figura 1) sirve principalmente para separar a los granjeros (GSE 1) del resto de los grupos socioeconómicos. Cuanto más cercanos están dos GSE en el «mapa» de la figura 1, más parecidos son los destinos ocupacionales de los hijos de padres de esos grupos. A la luz de esta interpretación, el modelo de los GSE que se ofrece en la figura 1 cobra sentido, y es en efecto una representación de la distancia social entre grupos en el sentido de la movilidad padres-hijos.

TABLA 2. Coordenadas de trece grupos socioeconómicos (GSE) en una solución bidimensional (rotada en ejes principales). (Reproducido de K. I. McDonald, *The analysis of social mobility: Methods and Approaches*, p. 210, Oxford, Clarendon Press).

Grupos socioeconómicos	Dimensión 1	Dimensión 2
1. Granjeros	−0.496	−1.944
2. Trabajadores agrícolas	−1.123	−0.453
3. Altos administradores, etc.	1.708	0.117
4. Otros administradores, etc.	0.508	0.235
5. Tenderos	0.766	−0.516
6. Oficinistas	0.761	0.276
7. Dependientes de comercio	0.297	0.220
8. Servicio doméstico	−0.120	−0.270
9. Capataces	−0.115	0.519
10. Obreros especialistas	−0.352	0.550
11. Obreros semi-cualificados	−0.592	0.547
12. Obreros sin cualificación	−0.691	0.283
13. Fuerzas armadas	−0.461	0.534

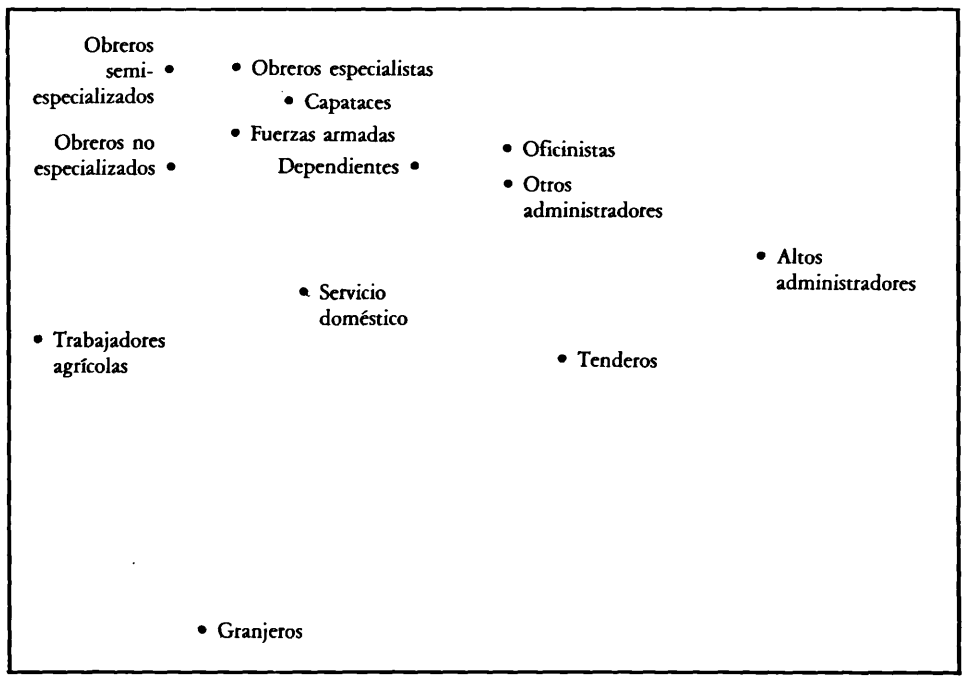


FIGURA 1. Gráfico de la solución bidimensional para trece grupos socieconómicos tal como aparece en la Tabla 2 (Reproducido de K. I. McDonald: *The analysis of social mobility*, p. 210. Clarendon Press, Oxford).

4.3. Ampliaciones del modelo multidimensional no-métrico de distancias

El paradigma original del EMD no-métrico se aplicaba al análisis de:

1. matrices *cuadradas.*
2. y *simétricas* de coeficientes de semejanza,
3. por medio de un modelo *euclídeo*
4. de *distancias*
5. que requiere una *transformación monótona* de los datos.

(Por comparación, el análisis en componentes principales puede considerarse como análisis de una matriz *cuadrada y simétrica,* por medio de un modelo *vectorial* que requiere una *transformación lineal* de los datos). Históricamente hablando, ha habido además otra restricción importante:

6. *El análisis tradicionalmente se limitaba a matrices de datos de doble entrada.*

Todas y cada una de estas restricciones han sido suprimidas, permitiendo la ampliación del EMD a una extensa variedad de modelos y tipos de datos, que abarcan muchas situaciones que se presentan en la investigación por encuestas. Examinemos ahora estas generalizaciones del paradigma:

(1) y (2): Quizá la generalización más útil provino de la extensión de las matrices cuadradas y simétricas a datos de similaridad condicionada, en los que las constricciones ordinales se satisfacen condicionalmente, esto es, sólo entre las filas o sólo entre las columnas. Los datos relativos a dos conjuntos distintos de objetos (p. ej., individuos e ítems de un cuestionario) pueden ser así analizados en el marco del EMD, al igual que se puede hacer con matrices de datos cuadradas y *asimétricas.* Es muy común que los investigadores recojan datos del primer tipo. Por ejemplo, se puede pedir a los sujetos que valoren u ordenen ocupaciones según su utilidad social, detergentes según su poder limpiador, aspiraciones de status en términos de convivencia, etc. Un ejemplo del segundo tipo (matrices cuadradas asimétricas) podría ser el de los estudios de sociometría, donde cada miembro del grupo juzga a todos los otros en función de un criterio como puede ser la amistad.

Un modelo comúnmente usado para escalonar tales datos es el modelo de desdoblamiento *(unfolding)* de Coombs, que localiza en un espacio común tanto al conjunto de datos que representan a los estímulos (los objetos elegidos o preferidos) como al conjunto de puntos que representan los sujetos (donde cada punto representa la localización preferida por un sujeto o su «ideal»). Para ello sigue la regla de que el orden de las preferencias de un individuo se corresponde con el auténtico orden de las distancias entre su punto ideal (fijado) y el conjunto de estímulos. Se pueden encontrar varios de estos programas de «desdoblamiento multidimensional». Una aplicación interesante del modelo de desdoblamiento ha sido documentada por Levine (1972), en un estudio sobre conexiones entre los miembros de los consejos de administración de 14 bancos y los de 96 corporaciones industriales. Estableciendo una analogía con el modelo corriente que estudia los juicios de preferencia de unos individuos acerca de un conjunto de estímulos, Levine decidió considerar a los bancos como «individuos»,

mientras que las corporaciones industriales serían los «estímulos». Un banco se consideraba el más próximo a una corporación cuando tenían el máximo número de directivos comunes, y se consideraba lejano respecto de ella cuando compartían un número de directores pequeño. Los datos consistían en 14 clasificaciones, todas hechas con las mismas 96 corporaciones industriales. En cada una de las clasificaciones había un gran número de empates (esto ocurría cuando un banco manifestaba tener el mismo número de directores comunes con la corporación A que con la B). La solución proporcionó un esclarecedor mapa de ambos, bancos y corporaciones industriales, representados como puntos en un espacio tridimensional común a los dos grupos.

(3) Hasta la fecha, la mayoría de los estudios del EMD han usado el modelo de distancia *euclídea,* ya sea por conveniencia, porque se creyera en su robustez *(robustness),* o debido a sus implicaciones sustantivas (Shepard, 1964; Sherman, 1972). Sin embargo, se han incorporado otros tipos de distancia al paradigma del EMD (Carroll y Wish, 1975) aplicándolos en las ciencias sociales (Arnold, 1971).
(4) El modelo de distancia se ha ampliado a otros modelos. Lingoes (1972) y otros han desarrollado analogías no métricas del análisis factorial y de la regresión múltiple, buscando una solución que sea acorde con los supuestos del modelo lineal. Una vez más, el objetivo es proporcionar la mejor solución a la transformación monótona de la matriz simétrica de los datos de (di-)similitudes. Un ejemplo de modelo vectorial del EMD para el análisis de preferencias individuales (análogo al análisis de desdoblamiento) es el modelo MDPREF de Carroll y Chang (1964) (véase también en Kruskal y Shepard, 1974), en el cual los estímulos se representan como puntos en un espacio multidimensional, mientras que cada conjunto de preferencias del sujeto se representa en el espacio de los estímulos como un *vector* o línea dirigida hacia su región de máxima preferencia. El orden de las proyecciones de los puntos-estímulo sobre este vector representa el orden de las preferencias del sujeto. Tanto si se parte de un conjunto de valores escalonados obtenidos de matrices de comparaciones por parejas, como si se hace de ordenaciones (y de la dimensión del espacio especificada previamente por el usuario), el modelo utiliza el teorema de Eckart-Young (1936) para dar lugar a una solución donde el orden de las preferencias de los sujetos se corresponde de forma óptima con el orden de las proyecciones de los estímulos sobre el vector que representa al sujeto.

En un estudio sobre aspectos subjetivos de la estratificación social hemos pedido a una serie de sujetos que valoraran un conjunto de ocupaciones en términos de lo que ellos consideraran que era su grado de *utilidad social.* El modelo MDPREF presentado aquí proporciona un esquema de análisis aplicable a datos como los que se recogen con regularidad en encuestas. La solución bidimensional del MDPREF ilustra sobre la aplicabilidad del modelo, y se presenta en la figura 2.

Para conseguir que los puntos-estímulo se distinguiesen nítidamente, cada sujeto se representa sólo por la *punta de flecha de su vector,* cuyo origen es el centro de coordenadas del espacio).

No sería apropiado aquí un análisis detallado, pero sí son oportunos algunos comentarios. La configuración de los estímulos se interpreta con facilidad. Se estará de acuerdo en que la mayoría de las preferencias se orientan en la dirección del eje horizontal de la figura, y en que el abanico de preferencias abarca sujetos con ordena-

ciones casi diametralmente opuestas. Es importante tener presente que este modelo proporciona una representación gráfica de las diferencias individuales en los ordenamientos de las preferencias fácilmente comprensible.

(5) Los primeros modelos EMD estaban diseñados para conservar en los datos información métrica, suponiendo que las disimilaridades empíricas eran una función *lineal* de las distancias del modelo. Las versiones no métricas del EMD sólo buscan conservar información ordinal, y de aquí que se permita usar la más amplia gama de funciones monótonas. En los procedimientos EMD se distingue entre los dos tipos de función según el método de regresión que se emplea para obtener una solución (en el caso métrico se utiliza la regresión lineal sobre los datos de distancias, mientras que en el no métrico se usa la regresión monótona (Kruskal, 1964: Barlow *et al.*, 1974)).

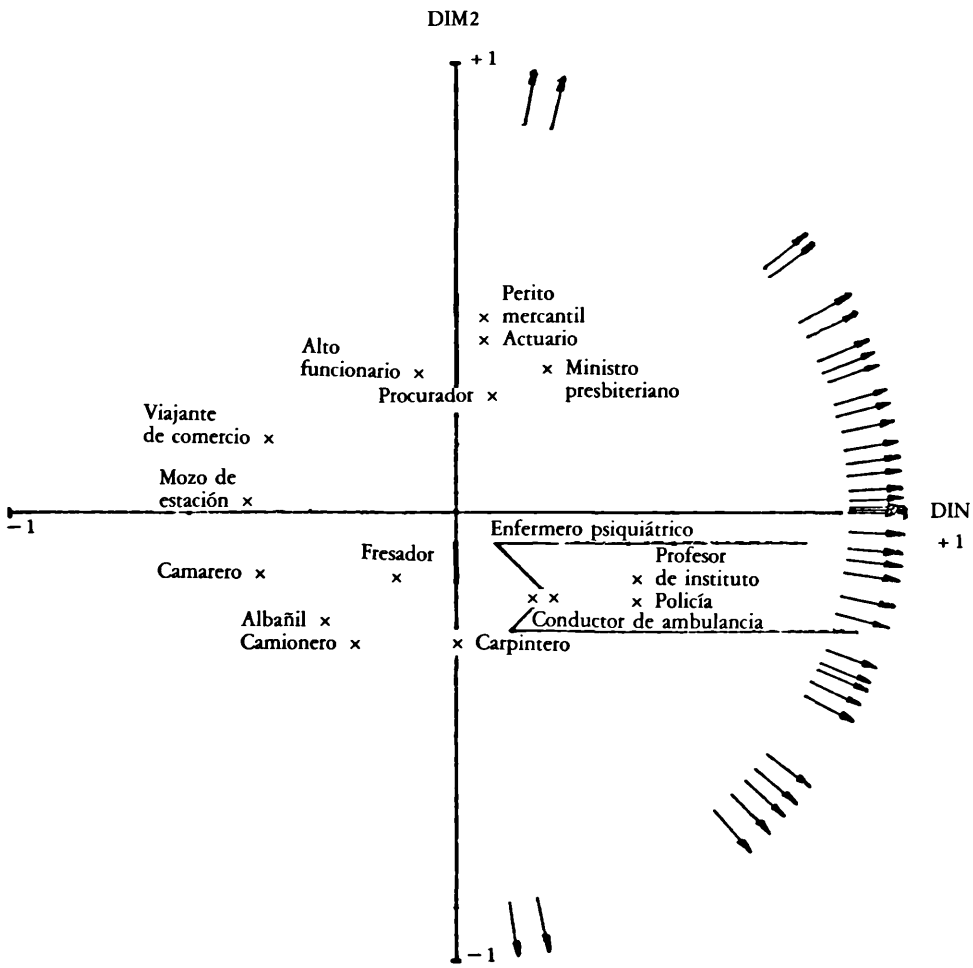


FIGURA 2. Ordenaciones de la utilidad social de 16 ocupaciones. Espacio conjunto (solución MDPREF bidimensional).

(6) *Ampliación del enfoque del EMD a matrices de más de dos entradas.*

Volvamos ahora a la discusión sobre la manera en que el enfoque del EMD se ha ampliado a matrices de más de dos entradas (en realidad se entiende por matrices de más de dos entradas el tratamiento simultáneo de una serie de matrices de doble entrada (filas y columnas)). Estos datos aparecen, por ejemplo, en los estudios sobre el diferencial semántico, donde la matriz básica de datos tiene tres entradas: individuos, conceptos y escalas.

En el modelo de distancia no métrico usual los datos (disimilaridades) se contemplan como una función monótona de las distancias (euclídeas) entre los puntos que representan los estímulos:

$$\delta_{jk} = M(d_{jk}) = \left\{ \sum_{a=1}^{r} (x_{ja} - x_{ka})^2 \right\}^{1/2} \qquad [1]$$

donde δ_{jk} es la disimilaridad entre los puntos j y k; d_{jk} es la correspondiente distancia representada en un espacio euclídeo r-dimensional, cuyos ejes de coordenadas definen un espacio x; y M es la función monótona que relaciona las disimilaridades con las distancias.

En el contexto de las encuestas sociales, a menudo nos encontramos con muestras que incluyen subgrupos de características completamente dispares, no sólo en sus medias y varianzas, sino también en sus pautas de covariación. Si los subgrupos son suficientemente parecidos es razonable agregar todos los datos y analizarlos según el modelo de distancias. Pero si los subgrupos son completamente diferentes y simplemente agregamos sus matrices de disimilaridad, entonces este procedimiento probablemente borrará las diferencias sistemáticas entre los subgrupos. Una alternativa podría ser escalonar los datos de cada subgrupo separadamente, pero quedaría sin respuesta la pregunta de cómo puedan compararse unos con otros o referirse a algún grupo común de coordenadas.

En el ámbito de la psicología, este problema aparece bajo el epígrafe de «diferencias individuales» de la estructura cognitiva, y gran parte del pensamiento metodológico en este área ha sido estimulado por este problema. Como vía de aproximación al tema de la agregación, Horan (1969) inventó un «espacio de atributos normales», o «espacio de los estímulos del grupo», definido por todos los atributos o dimensiones que un conjunto de individuos emplea en la percepción, y que les diferencian entre sí. Cualquier espacio cognitivo «privado» de un individuo (o de un subgrupo) puede ser concebido como un subespacio de este «espacio de atributos del grupo», y aquellas dimensiones que un sujeto no utilice podrían entenderse como ponderadas en su caso por un coeficiente igual a cero. Una extensión sencilla de esto es pensar en los sujetos o subgrupos como si estuvieran dotados de una «importancia», «prominencia» o «peso» diferenciales con respecto a cada dimensión en el espacio. Dicho de forma diferente, los datos de cada sujeto se pueden ver como el resultado de la aplicación de la «métrica subjetiva» de cada cual (dilatación o encogimiento sistemáticos) a las dimensiones del «espacio del grupo».

En tal modelo, cada individuo se representa en un «espacio del sujeto» (que tiene las mismas dimensiones que el «espacio del grupo») como un punto localizado por los

componentes en cada dimensión. Entonces, aplicando el conjunto de pesos del individuo al «espacio del grupo» (esto es, reconstruyendo el espacio a escala según la propia métrica subjetiva del entrevistado) se obtiene el «espacio privado» propio del individuo. Este modelo de escalonamiento de las diferencias individuales fue desarrollado por Carroll y Chang (1970) (Individual Differences Scaling - INDSCAL, a partir de aquí).

La principal diferencia entre el modelo INDSCAL y el modelo simple de distancias consiste en que el juicio de disimilaridad del sujeto se representa como una distancia ponderada en el «espacio del grupo», y en este sentido el modelo INDSCAL es una generalización sencilla del de distancias. La disimilaridad entre los estímulos *j* y *k* para un individuo *i* aún se supone función (lineal o monótona, dependiendo del modelo elegido por el investigador) de las distancias en su propio «espacio privado», que podemos definir por la matriz *Y:*

$$\delta_{jk}^{(i)} = F(d_{jk}^{(i)}) \qquad [2]$$

en donde:

$$d_{jk}^{(i)} = \left\{ \sum_{a=1} (y_{ja}^{(i)} - y_{ka}^{(i)})^2 \right\}^{1/2} \qquad [3]$$

pero estas distancias «privadas» pueden entenderse como equivalentes a unas distancias modificadas, en un espacio común (del grupo), *X:*

Las coordenadas del grupo (x_{ja}) y las «privadas» (y_{ja}) se relacionan por estas ecuaciones:

$$y_{ja}^{(i)} = w_{ia}^{1/2} x_{ja} \qquad [4]$$

$$\delta_{jk}^{(i)} = F \left\{ \sum_{a=1}^{r} w_{ia}(x_{ja} - x_{ka})^2 \right\}^{1/2} \qquad [5]$$

La ecuación [2] establece que para el sujeto *i*-ésimo las disimilaridades entre pares de estímulos son función (lineal o puramente monótona) de las distancias (d_{jk}) entre los correspondientes pares de estímulos en el espacio solución del sujeto, *Y.* Los puntos de los estímulos están localizados según sus proyecciones sobre un conjunto de ejes y_1, y_2, y_3, etc., y como muestra la ecuación [3], el modelo asume que las distancias entre los puntos de los estímulos, en lo que concierne al sujeto *i*-ésimo, se obtienen aplicando la función de la distancia euclídea a las proyecciones de los correspondientes pares de estímulos sobre los ejes de referencia. El modelo INDSCAL fue diseñado originalmente para representar las matrices de coeficientes de disimilaridad de una serie de individuos que juzgan un mismo conjunto de estímulos. La ecuación [4] muestra cómo se relacionan los ejes de coordenadas de los espacios «privados», *Y,* con todos los del «espacio del grupo», *X.*

Los pesos individuales, w_{ia}, pueden ser interpretados psicológicamente como las «importancias» o «prominencias» de la dimensión *a* para el individuo *i.* Con mayor generalidad, la incorporación de estos pesos al modelo permite que cada una de las

matrices individuales de distancias entre puntos tenga en una dimensión dada distancias sistemáticamente mayores o menores, comparadas con las del resto de las matrices. La ecuación [5] resume las anteriores, y muestra que el juicio de disimilaridad entre el estímulo *j* y el *k* para el invidiuo *i*-ésimo es una función *(F)* de la distancia euclídea individualmente ponderada según sus proyecciones sobre los ejes x_1, x_2, x_3, etc., en el espacio común del grupo *X*.

4.4. Modos de aplicación del modelo INDSCAL

La aplicación más directa de INDSCAL es el análisis de los juicios de similaridad por parejas emitidos por un conjunto de sujetos. Tales datos no se recogen normalmente en las encuestas sociales convencionales, aunque hay pocas razones para que esto no se haga (véase Wish *et al.*, 1970, y Coxon y Jones, 1974, para encontrar ejemplos). Sin embargo, es posible concebir que los datos de similaridad vengan de fuentes que no sean individuos; para cubrir esta ampliación del concepto se utiliza a menudo el término de «pseudo-sujeto». En el ejemplo que damos más adelante, los grupos de individuos sirven de «pseudo-sujetos» y las distancias se derivan de coeficientes de correlación.

Los analistas de encuestas a menudo hacen exámenes preliminares de sus datos inspeccionando las matrices de coeficientes de asociación, simétricas y cuadradas, entre todos los pares de variables que consideren relevantes. Con ayuda de esta inspección, toman decisiones sobre qué variables combinarán en índices compuestos de una u otra clase. Es muy común que los analistas empleen alguna técnica como el método de las componentes principales, el análiis factorial, o el análisis de conglomerados (clusters), para facilitar esta inspección. El EMD no métrico puede contemplarse como una herramienta más de este juego de reducción de los datos, y tiene la ventaja de que sólo se necesita asumir que el nivel de medida de la asociación entre las variables es ordinal. Por ejemplo, Napior (1972) ha explicado un análisis de una matriz de coeficientes gamma de Goodman-Kruskal entre dieciséis variables usando tres modelos nométricos: un algoritmo de agrupamiento jerárquico, el MDSCAL de Kruskal y el análisis factorial no métrico de Guttman-Lingoes. Estos análisis se resolvieron en la identificación de varios conjuntos de ítems unidimensionales que entonces se escalonaron independientemente utilizando un procedimiento simple de escalas de Guttman. McRae (1970) ha trabajado de una manera parecida con datos de elecciones legislativas. Calculó la *Q* de Yule como índice de similaridad entre cada par de listas por las que los representantes podían votar a favor o en contra, y obtuvo una representación en dos dimensiones de la matriz de similaridades resultante utilizando el EMD no métrico

Pero cualquiera que sea la técnica que se emplee en ayuda del ojo clínico del investigador, mientras que éste siga examinando sus datos en forma agregada corre el riesgo de reunir subgrupos de casos con diferentes pautas de relación entre las variables. Los psicólogos han expresado esto sugiriendo que diferentes grupos de personas pueden mostrar diferentes estructuras factoriales sobre la misma batería de tests. El siguiente ejemplo muestra cómo el modelo INDSCAL de Carrol y Chang proporciona al investigador de encuestas algunas palancas para afrontar el problema, y ade-

más le ofrece una representación visual simplificada de una masa de datos numéricos complejos.

Los datos para este ejemplo consistían en respuestas a un corto inventario de 21 ítems relativos a los valores que la gente espera satisfacer en el trabajo. Los sujetos que rellenaron el inventario fueron una muestra representativa de alumnos escoceses que dejaron la escuela en 1970 (para los detalles del muestreo, véase Jones y McPherson, 1972).

Se pensó *a priori* que las diferencias de comportamiento estadístico de las relaciones entre los ítems referidos a valores ocupacionales bien podían estar ligadas con las diferencias entre los perfiles medios de tales valores entre grupos diversos. Por consiguiente se definieron seis subgrupos del total de la muestra como probablemente portadores de perfiles de valores marcadamente distintos. Estos grupos eran los siguientes:

1. Hombres que ingresaron en cursos universitario de ingeniería o escuelas técnicas (n = 93).
2. Hombres que ingresaron en cursos de ciencias (n = 165).
3. Hombres que accedieron a un empleo a tiempo completo, sin seguir simultáneamente cursos de educación o adiestramiento a tiempo parcial (n = 175).
4. Mujeres que ingresaron en cursos de lengua o literatura en la Universidad (n = 115).
5. Mujeres que ingresaron en el curso de magisterio (n = 224).
6. Mujeres que accedieron a un empleo a tiempo completo, sin tomar clases a tiempo parcial (n = 195).

Los grupos de cursos universitarios se definieron de acuerdo con las normas del Consejo Central de Admisiones a la Universidad. Se calcularon las correlaciones producto-momento entre cada dos de los 21 ítems de valoración, para cada uno de los seis subgrupos.

Las seis matrices de correlaciones se podrían interpretar quizá como matrices de coeficientes de similaridad entre valores. Cuanto más grande sea la correlación entre dos variables, mayor será la similaridad entre ellas, y estarán más próximas en un espacio en el que las variables estén representadas como puntos. Sin embargo, una vez que hemos «producido» información al suponer que las relaciones eran lineales, lo cual es necesario para calcular los coeficientes de correlación, ya no tiene sentido pretender que no la tenemos. Es fácil mostrar que las distancias están relacionadas con las covarianzas y las correlaciones de Pearson (productos escalares) (véase Torgerson, 1958, o van de Geer, 1971) y pueden transformarse fácilmente. Para las covarianzas:

$$d_{ij} = (s_i^2 + s_j^2 - 2s_i s_j r_{ij})^{1/2}$$

donde s_i^2 es la varianza de la variable i y r_{ij} es la correlación producto-momento entre las variables i y j.

En el caso de que las variables estén normalizadas *(standard)*, la ecuación se reduce a:

$$d_{ij} = (2 - 2r_{ij})^{1/2} = 2(1 - r_{ij})^{1/2}$$

TABLA 3. Items de autovaloración que tienen que ver con «valores ocupacionales» y que se presentaron a una muestra representativa de personas que terminaban la escuela en Escocia.

1.	He pensado mucho en el nivel de ingresos que espero obtener antes de los treinta años (INGRESO). (Escala de nueve puntos, desde «fuertemente de acuerdo» a «fuertemente en desacuerdo»).
2.	(Mi trabajo ideal debería darme...) oportunidad de trabajar más con gente que con cosas (GENTE). (Escala de nueve puntos, desde «absolutamente esencial» hasta «no deseado en absoluto»).
3.	...oportunidad de tener una elevada posición cuando tenga treinta años (POSICION).
4.	...oportunidad de seguir intensamente mis intereses académicos e intelectuales (ACADEMIA).
5.	...permitirme mirar hacia delante y ver un futuro estable y seguro (SEGURIDAD).
6.	...oportunidad para ayudar a los jóvenes (AYUDAR).
7.	...quedar libre de la supervisión de otras personas (LIBERTAD).
8.	...pedirme una actividad original y creativa (CREACION).
9.	...proporcionarme aventura y excitación (EXCITACION).
10. (A)	...absorber mis energías e intereses tanto dentro como fuera del trabajo (ABSORBER).
11. (B)	...darme una oportunidad de ejercer un liderazgo moral (LIDERAZGO).
12. (C)	...incluir viajes al Reino Unido y al extranjero (VIAJES).
13. (D)	...operar intensamente con teorías e ideas (IDEAS).
14. (E)	...darme la ocasión de trabajar con niños (NIÑOS).
15. (F)	...darme poder para tomar decisiones importantes (PODER).
16. (G)	...tener una actividad laboral bien delimitada (DEFINICION).
17. (H)	...posibilitarme ascender en la vida (ASCENDER).
18. (I)	...permitirme estar junto a mis padres y mi familia (PADRES).
19. (J)	...no comprometerme con un determinado tipo de trabajo durante mucho tiempo (NO COMPROMISO).
20. (K)	...permitirme ser autónomo o mi propio jefe (AUTONOMIA).
21. (L)	¿Qué importancia crees que tiene el que un trabajo proporcione la oportunidad de ganar mucho dinero? (DINERO). (Escala de nueve puntos desde «extremadamente importante» a «en absoluto importante»).

y las distancias entre ítems pueden estimarse a partir de las correlaciones entre los mismos, tanto si tenemos en cuenta las desviaciones típicas de los ítems en cuestión como si no las tenemos. Claramente, el investigador debe tomar su propia decisión sobre qué método utilizará en un caso como éste. Dado que no es práctica normal estandarizar las variables *dentro* de cada subgrupo antes de formar las escalas o analizar los datos, nosotros aconsejamos estimar las distancias utilizando las covarianzas (y las desviaciones típicas), siendo este el análisis al que nos referimos aquí. Las seis matrices de distancias entre ítems fueron ajustadas al modelo INDSCAL que hemos descrito anteriormente. Se empleó la versión métrica del INDSCAL, ya que deseábamos que las distancias en el espacio solución fuesen una función lineal de (las «distancias» en) los datos.

El escalonamiento se hizo en tres y en dos dimensiones, y las coordenadas de los valores en las configuraciones de los espacios de los grupos se muestran en la tabla 4. Se debería tener presente que las dimensiones (correlacionadas) del espacio del grupo de una solución INDSCAL están «fijas», en el sentido de que cualquier rotación destruiría las propiedades estadísticas (mínimos cuadrados) de la solución. Por lo tanto tiene pleno sentido tratar de interpretar estos ejes, que pueden recibir un nombre en función de los ítems que tengan proyecciones máximas sobre ellos.

TABLA 4. Coordenadas tridimensionales y bidimensionales de la solución INDSCAL (solución sucesiva, IOY = 1), de las matrices de correlaciones producto-momento entre 21 ítems de un cuestionario (interpretadas como índices de similaridad entre ítems).

		Coordenadas INDSCAL (IOY = 1)				
		Solución 3-D			Solución 2-D	
	Cuestionario abreviado	I	II	III	I	II
1.	INGRESO	0.283	0.054	−0.062	0.283	−0.053
2.	GENTE	−0.081	−0.283	0.202	−0.081	0.283
3.	POSICION	0.189	0.176	−0.132	0.189	−0.176
4.	ACADEMIA	−0.054	−0.022	−0.330	−0.054	0.022
5.	SEGURIDAD	0.375	−0.128	−0.038	0.375	0.128
6.	AYUDAR	−0.093	−0.369	0.176	−0.093	0.369
7.	LIBERTAD	−0.125	0.096	0.015	−0.125	−0.091
8.	CREACION	−0.288	0.027	−0.240	−0.288	−0.027
9.	EXCITACION	−0.268	0.229	0.075	−0.268	−0.229
10. (A)	ABSORBER	−0.276	−0.134	−0.187	−0.276	0.134
11. (B)	LIDERAZGO	−0.077	−0.210	−0.214	−0.077	0.210
12. (C)	VIAJES	−0.168	0.338	0.019	−0.167	−0.338
13. (D)	IDEAS	−0.149	−0.069	−0.318	−0.149	0.069
14. (E)	NIÑOS	−0.158	−0.376	0.397	−0.158	0.376
15. (F)	PODER	0.039	0.078	−0.244	0.039	−0.078
16. (G)	DEFINICION	0.342	−0.218	0.078	0.342	0.219
17. (H)	ASCENDER	0.264	0.222	−0.142	0.264	−0.222
18. (I)	PADRES	0.229	−0.179	0.253	0.229	0.179
19. (J)	NO COMPROMISO	−0.077	0.303	0.435	−0.077	−0.303
20. (K)	AUTONOMIA	−0.225	0.306	0.230	−0.224	−0.306
21. (L)	DINERO	0.319	0.166	0.078	0.319	−0.166

Tomando sólo la solución bidimensional (ya que el ajuste para este modelo es tan solo un poco peor que para el tridimensional) la primera dimensión separa por un lado los ítems relacionados con el bienestar económico y las recompensas externas de un trabajo, y por otro aquéllos que implican satisfacciones intrínsecas al mismo (INGRESO, DINERO, DEFINICION, SEGURIDAD, ASCENSO, frente a CREACION, ABSORBER); la segunda dimensión, que podría denominarse «localismo *versus* valores cosmopolitas», distingue los ítems que incluyen una cierta orientación de cuidado o

alimentación de los niños, de aquellos que pueden describir los empleos que dan gran libertad de movimientos (AYUDAR, NIÑOS, DEFINICION, PADRE, LIDERAZGO, versus EXCITACION, VIAJES, NO COMPROMISO, AUTONOMIA).

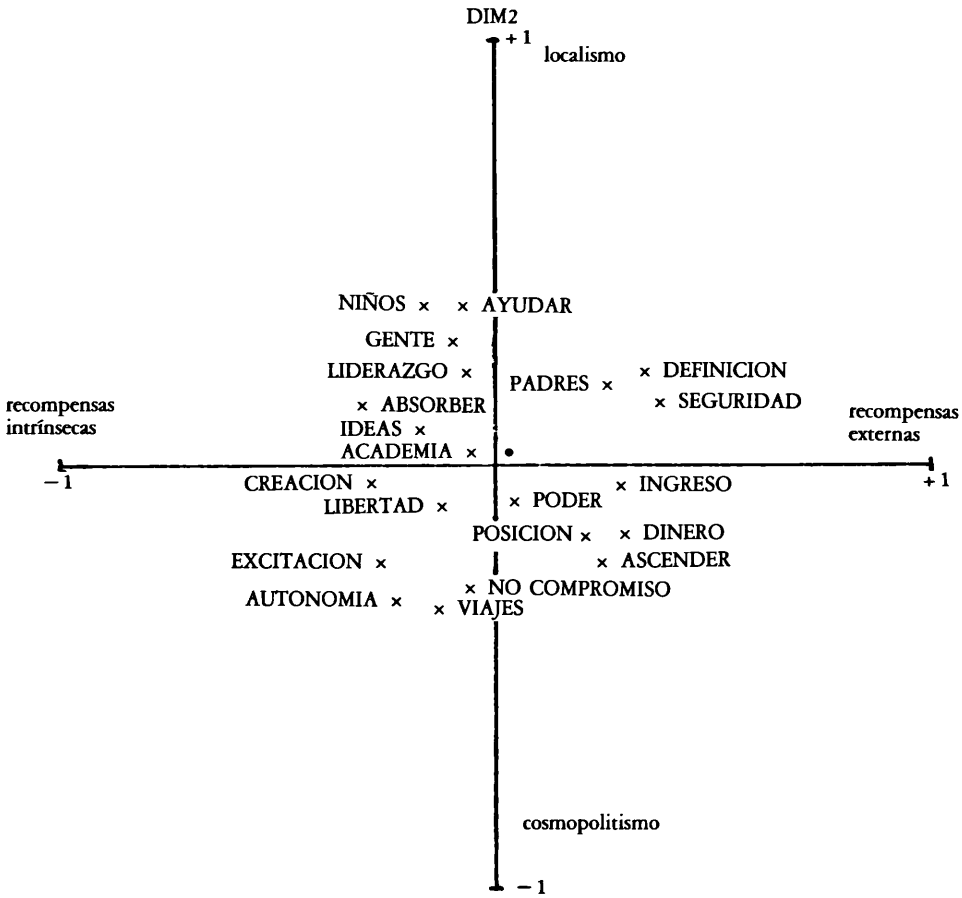


FIGURA 3. Espacio INDSCAL del grupo.

La configuración de los ítems de valores se representa en la solución INDSCAL en dos dimensiones (figura 3). A simple vista se pueden identificar fácilmente grupos de ítems similares, como por ejemplo (a) NIÑOS, GENTE, AYUDA, ABSORBER; (b) POSICION, DINERO, PODER, INGRESO, ASCENDER; (c) EXCITACION, AUTONOMIA, VIAJES, NO COMPROMISO.

En un estudio tradicional el análisis descriptivo y exploratorio podría terminar aquí, pero el modelo INDSCAL nos permite ir más lejos. Específicamente, el investigador puede preguntarse si la configuración global representa adecuadamente las pautas de relación entre los valores ocupacionales para los diferentes subgrupos. Como

ya hemos explicado, el modelo permite representar a los «sujetos» individuales o grupos como una transformación de la configuración conjunta. Esta transformación se lleva a cabo permitiendo que cada «individuo» modifique uniformemente («estire» o «encoja») la escala de cada una de las dimensiones de la configuración global de los estímulos, multiplicándolas por un peso. Los pesos correspondientes a cada uno de los subgrupos, constituidos en «pseudo-sujetos», se muestran en la tabla 5, y los pesos se representan en un espacio «de los sujetos» bidimensional (figura 4). Las diferencias son interesantes. La estructura de valores ocupacionales de los hombres indica que los ítems se organizan predominantemente con respecto a un factor de «recompensas intrínsecas *versus* externas», más que en función del de «localismo-cosmopolitismo». En términos de «distancias privadas» entre ítems, los grupos masculinos dan menos importancia a la segunda dimensión que a la primera. Por contra, los subgrupos femeninos organizan los ítems principalmente a través del factor «localismo versus cosmopolitismo», aunque debe tenerse en cuenta que lo que aquí hemos llamado «localismo» está ligado con los valores ocupacionales unidos al papel sexual; por ejemplo ayudar a los niños pequeños.

En el caso de los grupos de mujeres, se atribuye menos importancia a la primera dimensión que a la segunda. También parece que, dentro de cada sexo, los subgrupos cuyos miembros comenzaron a trabajar en lugar de acceder a la enseñanza superior

TABLA 5. Pesos INDSCAL que expresan las correspondientes transformaciones de la configuración conjunta (véase la tabla precedente) para los seis grupos.

Grupo	Solución 3-D			Solución 2-D	
	I	II	III	I	II
1. Estudiantes de ingeniería masculinos	0.658	0.284	0.418	0.658	0.284
2. Estudiantes de ciencias masculinos	0.588	0.437	0.359	0.588	0.437
3. Empleados masculinos	0.805	0.272	0.220	0.805	0.272
4. Estudiantes femeninas de lengua	0.452	0.683	0.230	0.452	0.683
5. Estudiantes femeninas de profesorado	0.336	0.806	0.156	0.335	0.806
6. Empleadas femeninas	0.739	0.394	0.183	0.739	0.394
Sumas de cuadrados	2.289	1.616	0.463	2.289	1.616
Indicador de la bondad del ajuste	Solución 3-D			Solución 2-D	
Porcentaje de varianza explicada por la solución INDSCAL	72.8%			65.1%	

Nota: Se utilizó la solución secuencial, y por lo tanto los ejes son ortogonales. (De hecho, en ambas soluciones I y II presentan una correlación de 0.049; la correlación entre I y III es 0.015, y entre II y III es 0.084).

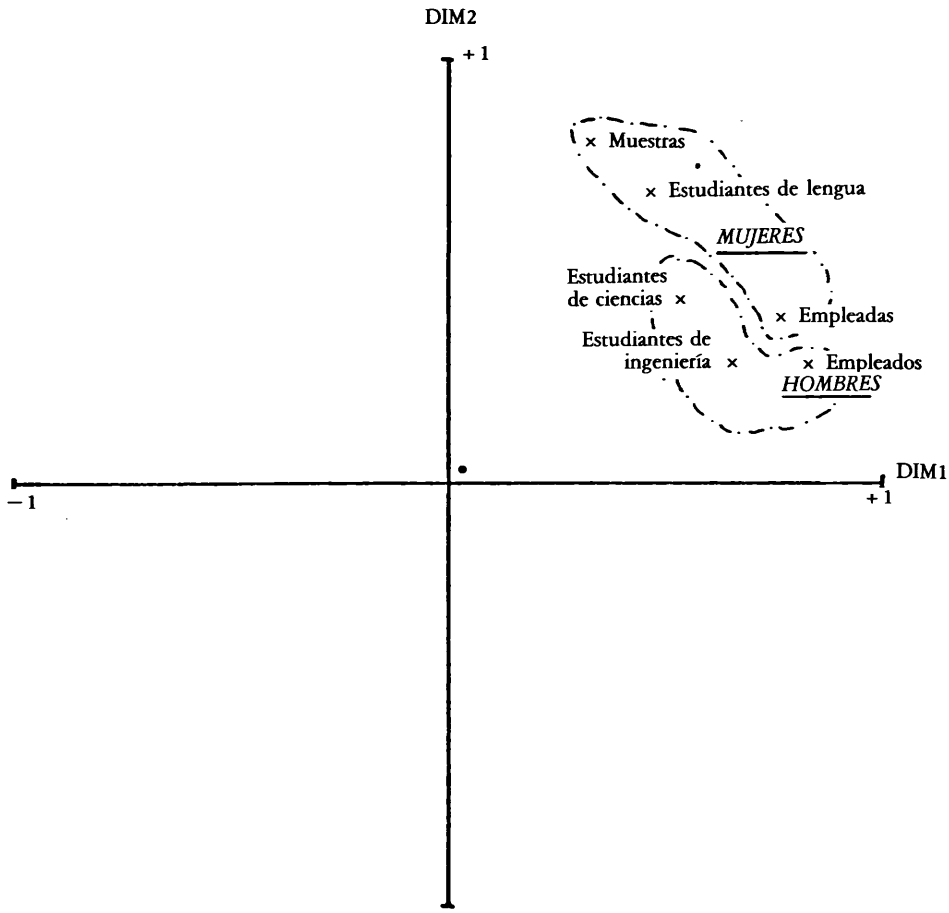


FIGURA 4. Espacio INDSCAL de los sujetos.

tienden a organizar sus valores dando prioridad a la dimensión I (recompensas externas-cosmopolitismo). Por supuesto, se puede usar la figura 4 para comparar los subgrupos en términos de sus pautas relativas de ponderación. Teniendo esto en cuenta es evidente que las mujeres que se pusieron a trabajar tienen unos pesos INDSCAL más parecidos a los de los grupos masculinos que aquellas mujeres que continuaron hacia la educación superior.

Las consecuencias de aplicar pesos «pseudo-sujetos» individuales a la configuración conjunta pueden verse mejor contrastando los «espacios privados» de dos subgrupos «pseudo-sujetos» de características extremas;

(a) Grupo 5: Mujeres que estudian para obtener el título de profesor de escuela primaria.
(b) Grupo 3: Hombres que accedieron a un empleo sin continuar sus estudios.

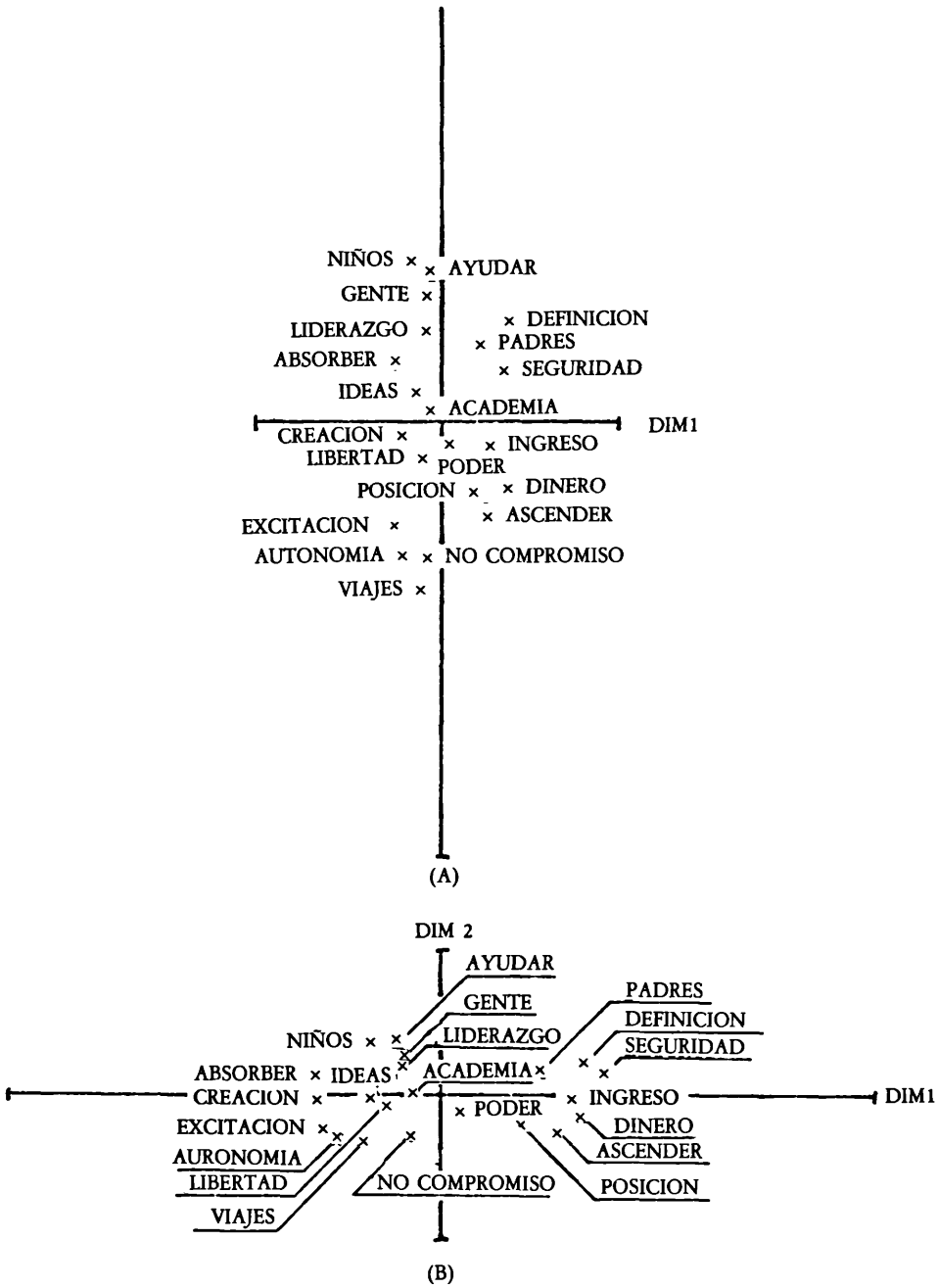


FIGURA 5. INDSCAL: dos espacios «privados».
(A) Grupo 5: mujeres, estudiantes de Magisterio.
(B) Grupo 3: hombres, sin más educación.
PESOS: A — (0.34, 0.91) B — (0.81, 0.27).

Claramente, los «estiramientos» y «encogimientos» uniformes que soporta el espacio global cuando se transforma en los «espacios privados» de estos dos grupos extremos indica plausiblemente el contraste que se da entre los modos en que se relacionan los ítems de los valores ocupacionales. Entre las mujeres que accedieron al curso de profesorado de escuelas primarias, la primera dimensión (horizontal en la figura 5) no discrimina mucho entre los ítems. La segunda (vertical) continúa separando los ítems a lo largo de un continuo que se refiere a la «ayuda a las personas con dedicación continua». El dibujo de la figura 5a contrasta fuertemente con el de la figura 5b. En ésta, la dimensión segunda (la vertical en la figura) se ha encogido, y es la horizontal (la primera) la que sigue discriminando, esta vez sobre la base de una referencia a los sistemas internos o externos de recompensas.

4.5. Conclusiones y desarrollos

Esta aplicación del modelo INDSCAL es útil para resumir y presentar concisamente el tipo de información sobre correlaciones que se genera a menudo en el análisis preliminar de los datos obtenidos en las encuestas sociales. También proporciona una vía de aproximación a un aspecto del problema de la «adecuación de los factores» al que nos referíamos anteriormente: decidir si diferentes muestras de sujetos que se han medido con el mismo conjunto de variables pueden o no ser descritas con el mismo conjunto de factores subyacentes. Parecidas conclusiones se extraen de la descripción y valoración de la interacción concepto-escala en el diferencial semántico (veánse Bynner y Romney, 1972; Jones, 1972 y Coxon, 1972 para continuar la discusión). Sin embargo, el modelo INDSCAL impone una restricción importante: la única transformación válida se deriva del estrechamiento o dilatación diferenciales de las dimensiones fijas de la configuración global. No están permitidas las rotaciones de ejes en el modelo INDSCAL, y la razón es la siguiente. Si todos los sujetos percibieran un espacio euclídeo de los estímulos idéntico, entonces se podría permitir cualquier rotación rígida, porque dejaría invariantes las distancias. Por contraste, el modelo INDSCAL afirma que, mientras que las distancias en cada espacio *privado* son euclídeas, no lo son referidas al espacio del grupo, ya que reciben factores de ponderación idiosincráticos para las «métricas subjetivas» propias de cada individuo (véase la ecuación [5] más arriba).

Cualquier rotación de las coordenadas del grupo $\{x_{ja}\}$ cambia cada una de las métricas subjetivas de los individuos. Esto significa que los valores de los datos predichos por un modelo ajustado cuyas coordenadas del grupo hayan sido rotadas (para facilitar la interpretabilidad o para otra cosa), siempre se ajustarán peor a los datos originales que en el caso de la solución sin rotar. El precio de la rotación es una pérdida de exactitud.

Carroll y Chang (1970) han sugerido que esta propiedad del modelo INDSCAL brinda una ventaja adicional, que es la interpretabilidad directa. Ciertamente, muchos estudios publicados por psicólogos parecen apoyar esta declaración, pero sería difícil argüir que la interpretabilidad sustancial se relaciona directamente con las propiedades formales. Schönemann y Wang (1972, p. 443) establecen una analogía con el método de las componentes principales:

«El hecho de que las componentes principales están únicamente dadas por unas condiciones matemáticamente apropiadas no significa necesariamente que sean también psicológicamente útiles o estén dotadas de sentido.»

La unicidad de la orientación de las dimensiones INDSCAL puede ser un inconveniente sustancial, cuando tal vez una rotación de la configuración del espacio del grupo proporcionaría un marco de referencia más interpretable; un análisis INDSCAL *independiente* para cada conjunto de datos de cada subgrupo revela a menudo notables diferencias en las orientaciones de los ejes. Si tales datos fueran agregados, la orientación de las dimensiones del grupo global sería en gran parte una «componenda» (en el mal sentido) entre las diferentes orientaciones. Un modelo que permitiese una rotación diferencial de los ejes, según sus pesos específicos, comportaría una clara ventaja en tales casos, ya que sería posible representar a los sujetos usando distintas combinaciones de los ejes de referencia para sus juicios. Carroll ha esbozado tal modelo, que es claramente una generalización del INDSCAL, pero para la mayoría de las aplicaciones en el análisis de encuestas los desarrollos como éste no tendrán probablemente un interés mucho más que académico. La importancia de estos modelos es que enseñan bastante bien que si el problema de la agregación en las poblaciones heterogéneas se toma en serio, entonces nuestros modelos de análisis no tendrán más remedio que corresponderse con la complejidad de los datos.

Nota del Editor

Programas de Ordenador

La mayoría de los programas informáticos de Escalas Multidimensionales se encuentran en la serie MDS(X), desarrollada por un equipo de la Universidad de Edimburgo y del University College de Cardiff. En esta serie se halla la siguiente relación de programas:

- CANDECOMP (CANonical DECOMPosition)
- HICLUS (HIerarchical CLUStering)
- INDSCAL-S (INDividual Differences SCALing)
- MDPREF (MultiDimensional PREFerence Scaling)
- MINICPA (Michigan-Israel-Nijmegen Integrated Series: Conditional Proximity Analysis)
- MINIRSA (MINI Rectangular Smallest Space Analysis)
- MINISSA (Michigan-Israel-Nijmegen Interated Smallest Space Analysis)
- MRSCAL (MetRic SCALing)
- MVNDS (Maximun Variance Non-Dimensional Scaling)
- PARAMAP (PARAmetric MAPping)
- PINDIS (Procustrean INdividual DIfferences Scaling)
- PREFMAP (PREFerence MAPping)
- PROFIT (PROperty FITting)
- TRISOSCAL (TRIadic Similarities Ordinal SCALing)
- UNICON (UNIdimensional CONjoint measurent)

La serie está implementada en el centro de procesos de datos de la Universidad Complutense de Madrid.

Para obtener información sobre estos programas se puede escribir a,

The MDS(X) Project
Sociological Research Unit
University College
P.O. Box 78
CARDIFF CF1 1XL
Gran Bretaña

SECCION II

LA CLASIFICACION DE LOS DATOS

5. Introducción

En esta sección incluimos dos técnicas: el Análisis Discriminante y el Análisis de Conglomerados[1]. Ambas técnicas tienen por objetivo la clasificación de los individuos, aun cuando difieran en la forma como se lleva a cabo dicha clasificación y la información que proporcionan. En el análisis de conglomerados hay que agrupar a los individuos (objetos) en una serie de grupos desconocidos por nosotros a priori. Cada grupo está constituido por un conjunto de individuos parecidos entre sí y diferentes al resto. Por el contrario, en el análisis discriminante el investigador define a priori los grupos, y los objetivos del análisis son (1) distinguirlos en base a la información que hay en los datos y (2) clasificar a los individuos en los grupos. Se trata de ver qué variables son las que más discriminan entre los conglomerados, con el fin de predecir la adscripción de los sujetos a los grupos en función de los valores que tomen en esas variables. En un caso, pues, los grupos hay que constituirlos a partir de las variables (análisis de cluster) y en otro los grupos están constituidos y lo que hay que ver es lo específico de cada uno de ellos para poder asignar los individuos a los grupos respectivos (análisis discriminante). Veamos por separado ambas técnicas.

Un ejemplo con dos grupos permitirá ilustrar el objetivo del *análisis discriminante*. Supongamos una muestra de n individuos, medidos en p variables. Y supongamos también que tenemos una variable y, con dos categorías, lo cual permite hablar de dos grupos, según los sujetos se clasifiquen en una u otra categoría. En esta variable damos un valor 1 al individuo que pertenece a un grupo y 0 al que pertenece al otro. Utilizando la regresión múltiple podemos predecir a partir de las p variables independientes el valor de Y para cualquier individuo. Según que la predicción se acerque al 1 o al 0 el individuo se clasificará en uno u otro grupo. Las variables con un mayor coeficiente de regresión serán aquellas que más discriminan, y el conjunto de variables será tanto mejor cuanto mayor sea la equivalencia entre los valores estimados y los reales (entre la clasificación conocida y la estimada por la regresión). El procedimiento serviría para predecir el grupo de pertenencia de aquellos individuos que no estuvieran clasificados en la variable Y (por ejemplo si Y es el voto, podríamos clasificar a los individuos que no contestan a esta pregunta), al tiempo que podríamos ver qué variables son las que más influyen en la clasificación.

En este ejemplo tan simple la regresión ha servido para realizar el análisis discrimi-

[1] En esta introducción vamos a utilizar indistintamente las palabras cluster, conglomerado o grupo.

nante. En situaciones más complejas el procedimiento cambia pero los problemas permanecen. Habrá que partir de unos supuestos del modelo que estemos utilizando. Habrá que seleccionar las variables que discriminan entre los grupos. También habrá que encontrar la(s) función(es) que más discriminan. Y por último habrá que definir un procedimiento de clasificación de los individuos. Respecto del primer punto, podemos enumerar los supuestos de la siguiente manera (véase Klecka, 1975 y 1980):

- es necesario que haya al menos dos grupos;
- para cada grupo se necesitan 2 o más casos;
- se puede utilizar cualquier número de variables discriminantes siempre y cuando su número sea inferior al número de casos menos dos;
- la variable que define los grupos ha de ser nominal, mientras que las variables discriminantes tienen que ser intervales. Cuando las variables sean cualitativas habrá que convertirlas en variables 0-1;
- ninguna variable discriminante puede ser combinación lineal de otras variables discriminantes;
- el número máximo de funciones discriminantes que se pueden calcular podrá ser igual al número de variables discriminantes, siempre y cuando su número no sea mayor que el número de grupos menos uno;
- las matrices de varianza-covarianza de cada grupo han de ser aproximadamente iguales;
- las variables discriminantes han de tener una distribución normal multivariable[2].

Respecto de la selección de las variables discriminantes digamos que en aquellos casos en los que se dude de la capacidad discriminante de las variables se puede hacer una selección entre todas ellas. Por ejemplo, puede haber variables cuyas medias en los diferentes grupos sean iguales o variables que compartan la misma información discriminante aun cuando cada una por separado sea capaz de discriminar. En estos casos se pueden seleccionar las variables siguiendo un procedimieno que, en función de algún criterio, elija la variable que más discrimina en cada paso, añadiendo sucesivamente las variables que forman la combinación más discriminante (véase procedimiento en el texto de Martínez Ramos).

Una vez que tenemos las variables, como indica el autor del capítulo, se trata de «reducir» su número a una, dos o más nuevas variables (factores) que sean combinaciones, lineales o no, de las anteriores. Ellas solas (y generalmente las dos o tres primeras) son capaces de identificar o discriminar a los grupos, previamente constituidos, tan bien como la larga serie de variables introducidas originalmente. Estas nuevas variables (factores) discriminantes reciben el nombre de funciones discriminantes. En términos generales podemos decir que las funciones discriminantes se calculan resolviendo el problema de los vectores propios, ya conocido del análisis factorial. Conceptualmente se trata de calcular los coeficientes para la primera función de manera tal

[2] En la práctica la técnica es muy robusta y no es completamente necesario que se cumplan los dos últimos supuestos. Klecka (1980: 60-62) explica qué ocurre cuando se viola este supuesto.

que las medias de los grupos en esta función sean lo más diferentes posible. Los coeficientes de la segunda función se calculan tratando de maximizar las diferencias entre las medias de los grupos en esa función, con la condición de que los valores de esta función no estén correlacionados con los valores de la primera. Y así sucesivamente con el resto de las funciones. A semejanza con el análisis factorial, una vez definidas las funciones discriminantes hay que interpretar su significado; para ello se procede de manera semejante a como se hace cuando se utiliza aquella técnica.

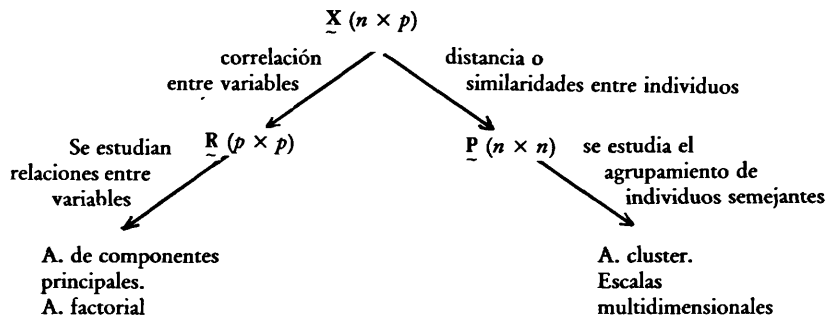
Por último queda la fase de clasificación de los individuos. Se puede hacer a partir de sus valores en las funciones discriminantes o en las variables discriminantes. En el caso de que se haga a partir de las variables discriminantes no se puede hablar propiamente de análisis discriminante, puesto que no se calculan las funciones discriminantes y su peso en la predicción de los grupos —un ejemplo de este tipo es el que se mostraba al comienzo de esta introducción, utilizando la regresión como método de clasificación—. Tal como señala Klecka (1980, pp. 47-48) la clasificación en base a las funciones reduce el trabajo considerablemente. Hay algunas circunstancias en las que los resultados (la clasificación) diferirán según se utilizen variables o funciones discriminantes; por ejemplo cuando las matrices de covarianza de los grupos sean diferentes, debido a que en el cálculo de las funciones discriminantes se hace el supuesto de igualdad mientras que este supuesto no se necesita cuando se utilizan variables discriminantes con el fin de hacer la clasificación. Otra circunstancia que lleva a resultados diferentes se da cuando se ignoran funciones discriminantes debido al hecho de que no son significativas. en este caso los resultados que se consiguen utilizando las funciones discriminantes son preferibles. En su trabajo Martínez Ramos explica el procedimiento de clasificación a partir de las funciones discriminantes.

Desde un punto de vista práctico, el análisis discriminante tiene parecido con el análisis factorial de correspondencias. Como parte del análisis en ambos casos se consigue aislar variables (categorías en el análisis de correspondencias) que diferencian o discriminan a los individuos. Recordemos cómo al hablar de las correspondencias decíamos que uno de los resultados era la formación de estereotipos. Igual ocurre en el análisis discriminante, puesto que las variables o las funciones que más discriminan también se puede decir que definen el estereotipo específico de cada uno de los grupos.

El nombre del *análisis de Cluster* se utiliza para definir una gran variedad de técnicas que tienen por objetivo la búsqueda de grupos en un conjunto de individuos. Junto al nombre de análisis de cluster también se utilizan otros nombres para definir el mismo procedimiento; por ejemplo, en biología se habla de «taxonomía numérica» y en psicología «reconocimiento de patrones». También se conoce la técnica como creación de tipologías o simplemente clasificación. A diferencia con otras técnicas del análisis multivariable los métodos más simples del análisis de conglomerados (métodos de las distancias mínimas, máximas, o distancias entre los centroides, por ejemplo) están desprovistos de todo contenido matemático y se pueden utilizar sin ninguna dificultad haciendo uso de los programas de ordenador.

Si partimos de la matriz de datos $\underset{\sim}{X}$ (n individuos $\times$ p variables), típica del análisis multivariable, las dos preguntas que nos podemos hacer son las siguientes: ¿son las variables similares?, ¿son los individuos similares? A partir de la matriz de correlaciones, obtenida de los datos, $\underset{\sim}{R}(p \times p)$, el análisis factorial da solución a la primera

pregunta. Calculando una matriz de proximidades, $P(n \times n)$, el análisis de conglomerados resuelve la segunda[3].



Básicamente, podemos decir que el análisis de conglomerados consiste en poner límites o barreras a un conjunto de individuos (objetos). Si representamos a los individuos en un espacio euclídeo, donde los ejes son las variables, la posición de los individuos dependerá de los valores que tomen en las variables (fig. 1). El análisis tratará de agrupar a los individuos en función de su similaridad en todas las variables consideradas simultáneamente[4]

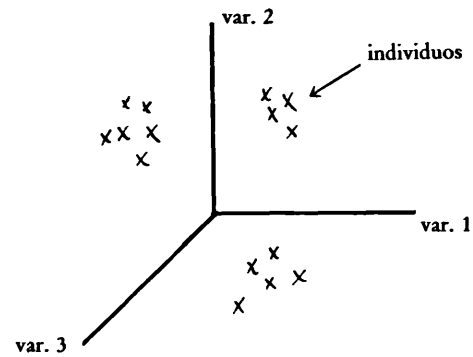


FIGURA 1. Representación de los individuos (objetos) en un espacio euclídeo.

Para proceder a la clasificación de los individuos el análisis sigue dos etapas: medición de la similaridad entre los individuos y aislamiento de los grupos. La solución al

[3] También es posible, aunque menos frecuente, que el análisis factorial se aplique a la matriz P y el análisis de conglomerados a la matriz R. Nosotros no nos vamos a referir en esta introducción a este tipo de análisis.

[4] Bailey (1974: 61) señala con acierto la confusión que se produce al decir que los individuos se asignan a uno u otro grupo, puesto que puede dar a entender que los objetos se mueven en el espacio. Los objetos permanecen estáticos y lo que se mueve son los límites de los clusters con el fin de dar cabida a los individuos que les «pertenecen».

problema de la medición de la similaridad dependerá del tipo de información del que se disponga. En el análisis de conglomerados se admiten variables intervales y variables del tipo ausencia-presencia (binarias). En el caso de tener variables nominales con más de dos categorías es necesario transformarlas en variables 0-1. Martínez Ramos ofrece un repertorio de medidas para situaciones en las que todas las variables son de tipo interval (medidas basadas en las distancias y medidas basadas en la correlación), todas son variables 0-1 (medidas de asociación) y situaciones en las que hay variables de todo tipo (coeficiente de similaridad de Gower).

Cuando las variables son intervales, con el fin de homogeneizar las diferentes escalas de medida hay que proceder a su estandarización, conscientes de que tal procedimiento tiende a diluir las diferencias entre lo grupos. Otros problemas de las variables a los que se refiere Martínez Ramos son la existencia de correlación entre ellas, su número excesivo o la ponderación a la que se les puede someter con el fin de relativizar su importancia. El autor explica las soluciones a estos problemas.

Respecto del problema de los métodos de clasificación, se puede distinguir entre dos grandes bloques: métodos jerárquicos y métodos de optimización (también llamados de reordenación). Todos los métodos del primer tipo tienen 2 cosas en común. Una, que producen una secuencia de particiones de los n individuos en g grupos, con g que va de n a 1 —una partición significa que los grupos son mutuamente excluyentes y que cada individuo está en un solo grupo—. Dos, que los métodos tienen la propiedad de que si a partir de un punto determinado del análisis dos individuos están en el mismo grupo, permanecen juntos para el resto del análisis.

Los métodos jerárquicos se dividen en aglomerativos o ascendentes y asociativos o descendentes. Martínez Ramos explica la diferencia entre ambos métodos e incluye diferentes algoritmos de clasificación para cada tipo. La pregunta de cuál es mejor tiene difícil contestación. Cuando los conglomerados son circulares (figura 2a) todas las técnicas son buenas (detectan bien los grupos). Los problemas surgen cuando los datos tienen las estructuras del tipo 2b o 2c (figura 2).

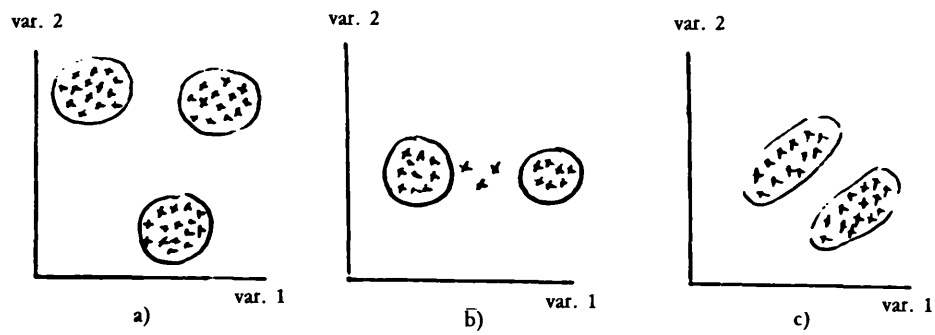


FIGURA 2. Representación de diferentes estructuras de clusters en un espacio bidimensional.

En el caso b, según Everitt (1974 y 1982) el método de las mínimas distancias *(single linkage)* lleva a malos resultados debido a que el «encadenamiento» de indivi-

duos que se establece entre los grupos tiende a diluirlos. En el caso c, métodos como el de la máxima distancia *(complete linkage)* o el de las distancias medias entre los grupos *(average linkage)* tienden a imponer en los datos una estructura esférica, sin detectar su propia estructura. Con el fin de ver la bondad de la estructura descubierta por el análisis se pueden utilizar dos criterios. Uno es que la solución debe de satisfacer la «desigualdad ultramétrica»[5]. Otro es que el coeficiente de correlación cofenético (véase Martínez Ramos) debe de tener un valor superior a .85.

En general, una desventaja de los métodos jerárquicos es que no permiten que los individuos (objetos) que han sido mal clasificados en una fase inicial del análisis puedan ser reclasificados en etapas posteriores.

Los métodos de reordenación implican la optimización de algún criterio de agrupamiento. El objetivo del análisis es encontrar una medida numérica indicativa de la bondad de una buena clasificación, para luego elegir una partición de los individuos en una serie de grupos que optimizen dicha medida. Las medidas utilizadas están basadas en la dispersión o varianza, tratando de lograr que los grupos se constituyan de tal manera que su dispersión intra grupo sea pequeña comparada con la dispersión entre grupos.

En términos prácticos, la diferencia más aparente entre estos métodos y los jerárquicos radica en que en los últimos se tienen las relaciones entre los individuos estructuradas jerárquicamente, mientras que en los primeros el resultado del análisis sólo proporciona los grupos finales con sus individuos. Debido al tiempo de ordenador gastado en los cálculos los métodos de optimización no admiten el análisis de tantos individuos como los jerárquicos —según Everitt (1977), 1.000 individuos y 50 variables superaría ya el tope admisible de los métodos de optimización.

El análisis de conglomerados es una técnica que se puede aplicar junto con las escalas multidimensionales con el fin de definir e interpretar los grupos que se forman como resultado de este método (véase una ilustración en Kruskal y Wish, 1978: 44-46). La matriz de distancias entre los puntos que se obtiene en el análisis multidimensional se puede utilizar como entrada del análisis de conglomerados, tratando de ver los grupos que se constituyen e interpretándolos después con la ayuda de las dimensiones donde se sitúan.

Respecto de la relación entre el análisis de conglomerados y el análisis factorial ya hemos señalado que cuando se utiliza la técnica *Q* el análisis factorial también sirve para reducir el número de individuos. Pero así como en el análisis de conglomerados los grupos satisfacen los requisitos de cualquier tipología, la exhaustividad y la mutua exclusividad, en el análisis factorial un individuo (objeto) puede pertenecer a (saturar positivamente en) más de un factor, puesto que su varianza se divide entre los factores —es decir, los factores no son mutuamente excluyentes.

---

[5] Esta desigualdad dice que si tenemos dos individuos x e *y*, unidos al nivel *d* (donde *d* es el nivel del dendograma) $d(x,y) \leq \max(d(x,z), d(y,z))$

6. Fundamentos del Análisis Discriminante y su aplicación en un estudio electoral

Por *Emilio Martínez Ramos*

6.1. Introducción

Supongamos una muestra de individuos que pertenecen a varias clases o grupos. Por ejemplo, una muestra de individuos distribuidos en función de la intención que tienen de votar a los partidos políticos en unas próximas Elecciones Generales. De cada individuo conocemos una serie de variables que le caracterizan y que definen sus actitudes políticas. La distribución de los individuos en grupos (electorados potenciales) es conocida a priori, antes de tomar cualquier otra información de esos individuos. Pues bien, el objetivo del análisis discriminante consiste en diferenciar los grupos (electorados o potenciales en nuestro caso), previamente definidos, lo más posible teniendo en cuenta exclusivamente la información disponible de cada sujeto y, en el mismo sentido, conocer cuáles son las variables más discriminantes, es decir aquellas que mejor discriminan a la muestra en dichos grupos. Que existen variables que son capaces de discriminar un electorado potencial de otro parece obvio, ya que el tener una intención de voto determinada no es una consecuencia del azar, sino que depende de una serie de actitudes políticas.

En un lenguaje más tecnico, el problema que nos resuelve el análisis discriminante es el de «reducir» el número de variables que discriminan a los grupos a una, dos o varias nuevas variables (factores) que son combinaciones, lineales o no, de las anteriores y que ellas solas (y generalmente las dos o tres primeras) son capaces de identificar o discriminar los grupos, previamente constituidos, tan bien como la larga serie de ellas introducidas directamente en el análisis.

Así, pues, podemos diferenciar los siguientes objetivos del análisis discriminante.

Un primer objetivo, encaminado a explicar el fenómeno que estemos estudiando, consiste en determinar si en función de las variables con las que hemos caracterizado a los grupos, éstos quedan suficientemente discriminados. En la práctica puede suceder que las variables elegidas para caracterizar a los grupos no sean las que mejor los discriminan; y en tal caso, en función de ellas no podemos hablar formalmente de uno u otro grupo. Por ejemplo, si tenemos dos grupos: los votantes de derecha y de izquierda, y para identificarlos hemos incluido variables como «estado civil» o «número de hijos», es evidente que no podemos diferenciarlos. Si sólo dispusiéramos de esta información, no podríamos hablar de un grupo de votantes de derecha y otro de izquierdas, aunque sepamos que existan, y así nos lo hayan declarado los propios indi-

viduos. Ahora bien, si sabemos que existen esos grupos en la realidad, es razonable pensar que existirá un conjunto de variables que los identifique. Precisamente la búsqueda de esas variables discriminantes y la determinación del poder de discriminación de cada una de ellas, es uno de los objetivos del análisis.

Un segundo objetivo, encaminado a predecir un comportamiento, consiste en atribuir o asignar un individuo, del que no conocemos a qué grupo pertenece a priori (individuo anónimo), a uno de ellos, con un cierto grado de riesgo, siempre en función de la información disponible.

Pero el análisis exige unas asunciones previas. Existen tests capaces de comprobar en qué medida se cumplen estos requisitos previos que enumeramos sucintamente a continuación.

Si cada individuo está representado por un vector en el espacio de las variables, suponemos, que esos vectores siguen una ley multinormal y que los centroides de cada grupo se diferencian significativamente.

Igualmente asumimos que, en la población, las matrices de varianza-covarianza de los grupos son iguales o muy parecidas[1].

La muestra deberá ser representativa de cada uno de los grupos que estén constituidos a priori. Sin embargo no es necesario que el tamaño de la muestra de cada grupo sea el mismo.

Las variables deberán ser elegidas de manera que puedan definir y discriminar los grupos. No sabemos con qué importancia participará cada variable en la discriminación, ni qué submuestra de ellas es la que más discrimina. Lo que nos interesa, antes de realizar el análisis, es introducir todas aquellas variables que puedan en mayor o menor grado explicar el fenómeno, es decir, su desagregación en modalidades o grupos. Estas variables deberán ser elegidas de manera que sean lo más independientes posible.

A partir de estas disquisiciones previas vamos a seguir explicando el análisis discriminante sobre el soporte de un caso concreto.

6.2. Selección de las variables discriminantes

Se ha realizado una encuesta en una provincia española a 492 individuos. El estudio es una típica encuesta preelectoral con 26 preguntas sobre actitudes y comportamientos políticos, así como con preguntas sobre la caracterización personal del individuo, y tiene como objetivo estimar el voto de los electores en dicha provincia (véase variables en Apéndice).

Realizada la encuesta, de los 492 individuos, sólo 275 nos declaran su intención de voto y 217 o no lo declaran o no están decididos todavía. La encuesta se realizó una semana antes de las elecciones legislativas de 1982.

---

[1] Algunos programas de análisis discriminante, como por ejemplo el implementado en el paquete BMDP, realizan el test de Box con el fin de contrastar esta hipótesis; en caso de rechazar el supuesto, el programa se detiene no realizando ningún otro cálculo.

La intención de voto declarada en la encuesta es:

	Número de casos por grupo
PSOE	121
AP	77
PCE	4
CDS	6
Otros	13
Abstención	54
TOTAL	275

Estos son los grupos constituidos a priori. Recordemos que faltan 217 individuos que, o son indecisos, o no han querido declarar su voto.

En primer lugar, el análisis nos muestra los valores medios y las desviaciones típicas de cada variable, en cada uno de los grupos constituidos a priori, y en el total de la muestra.

Veamos:

Medias

	P.15	P.50	P.14	P.18	P.53	P.51	P.22	P.54
PSOE	0.00000	0.88430	0.00000	0.02479	0.01653	0.01653	0.78512	0.01653
AP	0.00000	0.03896	0.00000	0.72727	0.00000	0.77922	0.05195	0.07792
PCE	0.00000	0.25000	0.50000	0.00000	0.00000	0.00000	0.25000	0.00000
CDS	0.66667	0.16667	0.00000	0.00000	0.16667	0.00000	0.00000	0.16667
OTROS	0.00000	0.07692	0.00000	0.30769	0.07692	0.53846	0.23077	0.30769
ABSTENCION	0.00000	0.25926	0.00000	0.09259	0.00000	0.16667	0.25926	0.03704
TOTAL	0.01455	0.46182	0.00727	0.24727	0.01455	0.28364	0.42545	0.05455

TABLA 1. Medias y desviaciones típicas de 8 de las 26 variables, para cada uno de los 6 grupos.

A continuación el análisis nos muestra la matriz de varianza-covarianza intra grupos que llamamos *W*. Exponemos a continuación sólo una parte de dicha matriz.

	P.15	P.50	P.14	P.18	P.53	P.51	P.22	P.54
P.15	0.4956629E-02							
P.50	-0.2478315E-02	0.1046099						
P.14	0.0000000E+00	0.1858736E-02	0.3717472E-02					
P.18	0.0000000E+00	-0.1278327E-01	0.0000000E+00	0.9481269E-01				
P.53	-0.2478315E-02	-0.7480241E-02	0.0000000E+00	-0.1328175E-02	0.1384146E-01			
P.51	0.0000000E+00	-0.2594071E-01	0.0000000E+00	0.1980249E-01	-0.2124607E-02	0.9644782E-01		
P.22	0.0000000E+00	0.2500183E-01	0.1858736E-02	-0.1295106E-01	0.4457185E-02	-0.1351617E-01	0.1399015	
P.54	0.1239157E-02	-0.1113472E-01	0.0000000E+00	0.4352500E-02	-0.1886308E-02	-0.2674930E-01	-0.1202720E-02	0.4843085E-01

TABLA 2. Reproducción parcial de la matriz de covarianza intra-grupos.

Esta matriz tiene $(n - g)$ grados de libertad, siendo g el número de grupos y n el número de individuos (275 − 6 = 269).

Tomemos por ejemplo la pregunta 15. En la matriz W, suma de las matrices de covarianza W_i, donde $i = 1, \ldots, 6$, el valor del primer elemento es, según las tablas, de 0.49566×10^{-2}.

Para conocer de dónde sale este valor no tenemos nada más que sumar ese primer elemento de la diagonal en todas las W_i. *El primer elemento de la diagonal en cada* W_i es, naturalmente, la varianza de cada variable.

Posteriormente, el análisis nos proporciona la matriz de covarianza de cada grupo, que llamamos W_A, W_B, W_C... etc.

Y después podemos observar en las tablas la matriz de covarianza total del conjunto de los 275 individuos. A esta matriz la llamamos T (tabla 3). Por ejemplo el elemento de la diagonal de la matriz T correspondiente a la variable P.15 es 0.0143862. Veamos de dónde sale esta cantidad. La desviación típica de la variable P.15 en el conjunto de los 275 individuos es, según las tablas, de 0.11994 (tabla 1), luego su cuadrado es la varianza que corresponde al primer elemento de la diagonal de la matriz T.

Matriz T

	P.15	P.50	P.14	P.18	P.53	P.51	P.22	P.54
P.15	0.1438620E-01							
P.50	-0.6741871E-02	0.2494492						
P.14	-0.1061712E-03	0.2786994E-03	0.7246184E-02					
P.18	-0.3609821E-02	-0.1036629	-0.1804910E-02	0.1868082				
P.53	-0.2123424E-03	-0.6741871E-02	-0.1061712E-03	-0.3609821E-02	0.1438620E-01			
P.51	-0.4140677E-02	-0.1314665	-0.2070338E-02	0.1193895	-0.4140677E-02	0.2039283		
P.22	-0.6211015E-02	0.1495156	0.5441274E-03	-0.9098872E-01	0.4737890E-02	-0.1028666	0.2453351	
P.54	0.2853351E-02	-0.2528202E-01	-0.3981420E-03	0.1201062E-01	-0.7962840E-03	-0.1552754E-01	-0.1234240E-01	0.5175846E-01

TABLA 3. Reproducción parcial de la matriz de varianza-covarianza del conjunto de los 275 individuos.

El programa que hemos utilizado en este caso va paso a paso. El cada paso selecciona la variable más discriminante. Para dicha selección utiliza varios estadísticos que pasamos a estudiar a continuación.

Veamos el primer paso, la selección de la primera variable más discriminante de las 26 que se han utilizado (tabla 4).

La variable más discriminante es la P.15, porque de acuerdo a las tablas es la que ha obtenido un mayor valor en el estadístico F o un menor valor en el estadístico Lambda de Wilks. Observamos que el valor $F = 105.2524$, en la P.15, es mayor que cualquiera de los valores de F en el resto de variables.

El estadístico F responde a la siguiente relación $F = \dfrac{|A|}{|W|}$ en la que $|A|$ es el determinante de la matriz de covarianza inter grupos y $|W|$ es el determinante de la matriz de covarianza intra grupos. F, por supuesto, es un escalar; el numerador nos indica la distancia entre los centroides de los grupos. Evidentemente, cuanto más grande sea esa distancia, mejor, más separados quedan los grupos. El denominador es el determinante de la matriz suma de las matrices de covarianza de los grupos; cuanto menor sea ese valor, mejor, pues más homogéneos serían los grupos en sí mismo. De manera que cuanto mayor sea F, mejor; y el mayor de todos es el correspondiente a la

AT STEP 1, P.15 WAS INCLUDED IN THE ANALYSIS.

		DEGREES OF FREEDOM		SIGNIFICANCE	BETWEEN GROUPS
WILKS' LAMBDA	0.3382534	1	5 269.0		
EQUIVALENT F	105.2524		5 269.0	0.0000	
RAO'S V	526.2618		5	0.0000 (APPROX.)	

---------- VARIABLES IN THE ANALYSIS AFTER STEP 1 ----------

VARIABLE	TOLERANCE	F TO REMOVE	RAO'S V
P15	1.0000000	105.25	

a)

---------- VARIABLES NOT IN THE ANALYSIS AFTER STEP 1 ----------

VARIABLE	TOLERANCE	MINIMUM TOLERANCE	F TO ENTER	RAO'S V
P50	0.9881545	0.9881545	76.480	910.1281
P14	1.0000000	1.0000000	52.809	791.3491
P18	1.0000000	1.0000000	53.449	797.1205
P53	0.9104749	0.9104749	8.7110	643.6270
P51	1.0000000	1.0000000	61.163	836.6047
P22	1.0000000	1.0000000	41.095	737.7559
P54	0.9936035	0.9936035	4.4810	548.9693
P11	0.9879171	0.9879171	7.7956	570.6300
P40	0.9975955	0.9975955	69.346	877.6696
P26	0.9857385	0.9857385	4.8097	557.6419
P03	0.9993795	0.9993795	2.4016	543.4156
P83	1.0000000	1.0000000	0.53201	
P86	1.0000000	1.0000000	1.5287	534.1204
P85	1.0000000	1.0000000	0.97504	
P13.	1.0000000	1.0000000	54.268	801.1477
P90	0.9983970	0.9983970	4.0988	548.4816
P19	1.0000000	1.0000000	0.80980	
P28	1.0000000	1.0000000	38.313	721.1088
P81	1.0000000	1.0000000	0.85892	
P60	0.9987505	0.9987505	1.4857	533.7205
P02	0.9914380	0.9914380	1.5140	541.6813
P33	0.9914149	0.9914149	5.9534	557.5354
P74	0.9955796	0.9955796	1.4427	543.7645
P24	1.0000000	1.0000000	4.7736	550.2989

b)

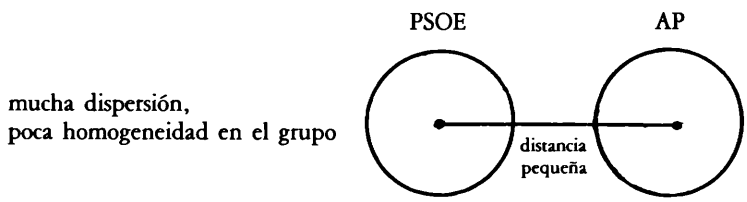
TABLA 4. Selección de las variables que más discriminan. a) variable elegida en el primer paso; b) variables que no se incluyen y sus estadísticos.

variable P.15. De todas las variables, la P.15 es la que ofrece un mayor valor en la relación entre las distancias inter centroides de los grupos y la suma de las varianzas de cada uno de ellos.

Veamos gráficamente estas ideas, en un primer caso en el que el valor de F es grande y en un segundo caso en el que es pequeño.

Primer caso en el que el valor de F es grande (variable muy discriminante).

Segundo caso, en el que el valor de F es pequeño (variable poco discriminante).



Hemos visto en las tablas que el programa utiliza, además de F, otros estadísticos. La Lambda de Wilks va en sentido contrario a F. La relación que existe entre ambos es la siguiente. Llamemos Λ a la Lambda de Wilks.

$$F = \frac{1}{\Lambda} - 1$$

Comprobemos en las tablas esta relación. Los valores de F y Λ en la primera variable que se introduce (P.15) en la función discriminante son:

$$F = 105.2524$$

$$\Lambda = 0.3382534$$

Conociendo Λ calculemos F según la relación anterior con los siguientes grados de libertad.

$$\frac{n - g - p + 1}{g - 1}$$

Siendo n el número de individuos, g el número de grupos y p el número de variables, que en este primer paso es

$$F = \left(\frac{1}{0{,}3382534} - 1\right)\left(\frac{275 - 6 - 1 + 1}{6 - 1}\right) = 105.2524$$

En la cuarta columna del cuadro que hemos expuesto anteriormente el programa nos proporciona la significatividad de F. El mínimo valor por encima del cual deja F de ser significativo es 0.001. Esta condición se establece antes del análisis y puede ser modificable. De manera que contrastamos la hipótesis nula de que no existe diferencia entre las medias correspondientes a esta variable P.15 que acabamos de introducir, en cada uno de los grupos. El valor de F obtenido es como ya sabemos de 105,2524 que se compara con el de las tablas de distribución de F. Los grados de libertad en este caso son, por un lado 5, es decir, 6 grupos menos 1; por otro lado 269, es decir, 275 individuos menos 6 grupos. Si la hipótesis nula de igualdad de medias es cierta, la probabilidad de obtener un valor de F como el conseguido es extremadamente baja, luego se rechaza la hipótesis nula, es decir, hay diferencia significativa entre las medias.

Después de cada paso y antes de introducir la siguiente variable, el programa nos presenta el siguiente cuadro en el que podemos analizar cómo se van diferenciando los grupos (tabla 5).

Para cada par de grupos se calcula el estadístico F. Debajo del valor F aparece el nivel de significatividad. Vemos que hay muchos valores entre grupos que todavía, después de este primer paso, no son significativos.

F STATISTICS AND SIGNIFICANCES BETWEEN PAIRS OF GROUPS AFTER STEP 1
EACH F STATISTIC HAS 1 AND 269.0 DEGREES OF FREEDOM.

GROUP	1	2	3	4	6
2	0.00000E+00 1.0000				
3	0.00000E+00 1.0000	0.00000E+00 1.0000			
4	512.58 0.0000	499.11 0.0000	215.20 0.0000		
6	0.00000E+00 1.0000	0.00000E+00 1.0000	0.00000E+00 1.0000	368.11 0.0000	
7	0.00000E+00 1.0000	0.00000E+00 1.0000	0.00000E+00 1.0000	484.20 0.0000	0.00000E+00 1.0000

TABLA 5. Valor de la F y nivel de significación entre parejas de grupos, después del paso 1.

Después, el programa efectúa el segundo paso en el que se busca una nueva variable, de las 25 que quedan, que forme, conjuntamente con la ya elegida, la P.15, la pareja de variables más discriminante. Para lo cual volvemos a emplear los mismos criterios que en el primer paso.

Según las tablas esta nueva variable es la P.50. Así como el primer paso era muy sencillo, pues al tratarse de una sola variable los cálculos se basaban en la varianza, ahora, con dos variables, la aplicación de los criterios de discriminación empieza a complicarse, puesto que ya hay que trabajar con matrices de varianza-covarianza.

Por ejemplo, el cálculo del criterio Λ (Lambda de Wilks), en este segundo paso tiene la siguiente forma.

$$\Lambda = \frac{\begin{vmatrix} V_{11}V_{12} \\ V_{21}V_{22} \end{vmatrix}_{\text{grupo A}} + \begin{vmatrix} V_{11}V_{12} \\ V_{21}V_{22} \end{vmatrix}_{\text{grupo B}} + \ldots \text{etc.}}{\begin{vmatrix} V_{11}V_{12} \\ V_{21}V_{22} \end{vmatrix}_{\text{en el total}}}$$

Siendo V_{11} la varianza de la variable P.15 en el grupo A, que puede ser el electorado potencial del PSOE; V_{22} la varianza de la variable P.50 en el mismo grupo; V_{12} la covarianza entre las variables P.15 y P.50 en el mismo grupo y así sucesivamente.

Veamos qué ha sucedido en este segundo paso (tabla 6).

Para que una variable pueda introducirse en la función discriminante debe aportar «algo» a la separación entre grupos. Para medir esta aportación ya hemos dicho que utilizamos una serie de criterios, por ejemplo el *F*. Ahora bien para poder entrar, la variable debe superar un valor mínimo de *F* que es 1. Por debajo de este valor de *F* no entra la variable. Por ejemplo, las variables P.83 y P.85 tienen valores en el estadístico *F* menores que 1 y nunca entrarán. Pero además de este criterio podemos utilizar otro que tiene en cuenta lo que en las tablas de ordenador es llamado «nivel de tolerancia». El nivel de tolerancia puede ir de 1 a 0.001. Como vemos existe para cada variable un nivel de tolerancia mínimo, el cual varía en cada etapa y por debajo del cual la variable no entra en la ecuación. Nos conviene que el nivel de tolerancia sea alto, porque la correlación entre las variables ya introducidas y la que se va a introducir, no puede exceder de 1 menos la tolerancia; entonces si la tolerancia fuera baja, la correlación es próxima a 1, y si la variable introducida tiene una alta correlación con las demás de la función discriminante, la matriz de covarianza será singular y no se puede realizar el análisis.

También conviene decir que algunas de las variables ya seleccionadas pueden ser expulsadas de la ecuación, es decir pueden perder su poder de discriminación. Esto ocurre porque se han ido introduciendo variables que conjuntamente están correlacionadas con alguna de las ya introducidas y en tal caso el programa las expulsa, lo que no presupone que no puedan volver a entrar en un paso posterior. Para medir esto se utiliza el «F to remove», el cual debe ser más pequeño que un valor dado antes del inicio del análisis.

En el tercer paso se vuelve a seleccionar una nueva variable y así sucesivamente hasta que el número de etapas coincida con las solicitadas a priori o hasta que la nueva variable no incorpore ningún poder de discriminación entre grupos, o hasta

AT STEP 2, P50 WAS INCLUDED IN THE ANALYSIS.

		DEGREES OF FREEDOM		SIGNIFICANCE	BETWEEN GROUPS
WILKS' LAMBDA	0.1393783	2	5	269.0	
EQUIVALENT F	89.97116		10	536.0	0.0000
RAO'S V	910.1281		10		0.0000 (APPROX.)

---------- VARIABLES IN THE ANALYSIS AFTER STEP 2 ----------

VARIABLE	TOLERANCE	F TO REMOVE	RAO'S V
P15	0.9881545	104.73	
P50	0.9881545	76.480	

a)

---------- VARIABLES NOT IN THE ANALYSIS AFTER STEP 2 ----------

VARIABLE	TOLERANCE	MINIMUM TOLERANCE	F TO ENTER	RAO'S V
P14	0.9910094	0.9792704	53.570	1185.380
P18	0.9833268	0.9716788	28.765	1124.548
P53	0.8573384	0.8573384	11.655	1066.107
P51	0.9325047	0.9214587	25.508	1105.475
P22	0.9567761	0.9454426	9.7423	1027.848
P54	0.9715178	0.9661900	3.0886	925.9125
P11	0.9832503	0.9746537	6.2496	948.5212
P40	0.8825072	0.8741553	19.108	1091.495
P26	0.9716505	0.9708424	4.3297	939.6745
P03	0.9993787	0.9875440	1.4410	927.1879
P83	0.9856399	0.9739645	0.47857	
P86	0.9942224	0.9824453	1.0285	915.5155
P85	0.9733256	0.9617961	0.47391	
P13	0.9878894	0.9761874	28.360	1133.604
P90	0.9199175	0.9104801	1.7493	922.9139
P19	0.9798952	0.9682878	1.6866	925.2934
P28	0.9851899	0.9735198	19.429	1058.613
P81	0.9808914	0.9692722	1.2437	919.4970
P60	0.9973923	0.9868107	1.4375	920.3025
P02	0.9909455	0.9802193	1.5043	926.4417
P33	0.9898427	0.9805201	2.3210	934.1361
P74	0.9805662	0.9732530	2.0360	935.2014
P24	0.9999125	0.9880680	4.4440	933.6431

b)

TABLA 6. Selección de las variables que más discriminan. a) variables elegidas en el 2.° paso; b) variables que no se incluyen y sus estadísticos.

que el valor del estadístico *F* sea menor que la unidad, como es el caso del ejemplo que estamos tratando o hasta que, sencillamente, se hayan agotado las variables.

Y siempre después de cada etapa el programa nos proporciona los valores de *F* para medir la diferencia entre centroides de cada par de grupos (tabla 7).

F STATISTICS AND SIGNIFICANCES BETWEEN PAIRS OF GROUPS AFTER STEP 2
EACH F STATISTIC HAS 2 AND 268.0 DEGREES OF FREEDOM.

GROUP	1	2	3	4	6
2	162.04 0.0000				
3	7.5072 0.0007	0.81610 0.4432			
4	259.41 0.0000	254.33 0.0000	107.92 0.0000		
6	36.875 0.0000	0.77241E-01 0.9257	0.44156 0.6435	186.91 0.0000	
7	70.293 0.0000	7.4232 0.0007	0.15386E-02 0.9985	242.71 0.0000	1.6787 0.1886

TABLA 7. Valor de *F* y nivel de significación entre parejas de grupos después del paso 2.°.

Naturalmente, conforme introducimos nuevas variables, la discriminación de los grupos va a ir mejorando, ya que cada nueva variable introducida, por poco que sea, incorpora su poder de discriminación. Si tal cosa no sucediera el programa se pararía. Vemos en el cuadro anterior que todavía, después del paso 2, el valor de *F* entre algunos pares de grupos, no es significativo. El análisis debe continuar.

En el caso que estamos analizando se registran 22 pasos y al final quedan 3 variables por introducir, precisamente aquellas, como ya hemos dicho, que alcanzan valores por debajo de la unidad en *F*. Veamos las tablas correspondientes al último paso, el 22 (tabla 8).

Y las tablas correspondientes a la situación entre grupos en la que podemos apreciar que todos los valores de *F* ya son significativos por debajo del 0.001 que permite el análisis (tabla 9).

Ya sabemos, pues, qué variables van a participar en el cálculo de la función discriminante. La función discriminante no es otra cosa que un factor, una nueva variable combinación lineal de las anteriores. Puede haber tantas funciones discriminantes co-

AT STEP 22, P24 WAS INCLUDED IN THE ANALYSIS.

		DEGREES OF FREEDOM		SIGNIFICANCE	BETWEEN GROUPS
WILKS' LAMBDA	0.0128866	22	5 269.0		
APPROXIMATE F	15.86650		110 1219.7	0.0000	
RAO'S V	2297.450		110	0.0000 (APPROX.)	

---------- VARIABLES IN THE ANALYSIS AFTER STEP 22 ----------

VARIABLE	TOLERANCE	F TO REMOVE	RAO'S V
P15	0.8238311	104.14	
P50	0.6484833	18.361	
P14	0.8202427	51.430	
P53	0.7738669	10.797	
P51	0.5884843	11.286	
P22	0.7229625	3.0873	
P54	0.6606170	5.2237	
P11	0.2841336	3.3180	
P40	0.6458087	4.9565	
P26	0.2729701	2.3402	
P03	0.6911179	1.2897	
P86	0.8511222	1.7770	
P85	0.8278980	1.3290	
P13	0.2551387	5.9171	
P90	0.8104369	2.4618	
P19	0.6856324	2.3981	
P28	0.2411300	1.8228	
P81	0.8997588	2.2634	
P60	0.8014564	1.8267	
P02	0.6981600	1.4989	
P33	0.9086608	1.5443	
P24	0.6309560	1.0605	

a)

---------- VARIABLES NOT IN THE ANALYSIS AFTER STEP 22 ----------

VARIABLE	TOLERANCE	MINIMUM TOLERANCE	F TO ENTER	RAO'S V
P18	0.1768289	0.1684992	0.75627	
P83	0.5164672	0.2407631	0.80817	
P74	0.8466900	0.2411244	0.54355	

b)

TABLA 8. Selección de las variables que más discriminan. a) variables elegidas al final del proceso; b) variables no seleccionadas y sus estadísticos.

mo variables menos una, pero generalmente las primeras son las que explican una mayor cantidad de varianza, en definitiva las que explican más el fenómeno que estamos estudiando, en nuestro caso la intención de voto.

F STATISTICS AND SIGNIFICANCES BETWEEN PAIRS OF GROUPS AFTER STEP 22
EACH F STATISTIC HAS 22 AND 248.0 DEGREES OF FREEDOM.

GROUP	1	2	3	4	6
GROUP					
2	39.560 0.0000				
3	16.827 0.0000	17.007 0.0000			
4	30.095 0.0000	30.287 0.0000	21.206 0.0000		
6	9.2466 0.0000	3.1721 0.0000	13.808 0.0000	21.223 0.0000	
7	12.215 0.0000	11.335 0.0000	14.776 0.0000	28.157 0.0000	4.2232 0.0000

F LEVEL OR TOLERANCE OR VIN INSUFFICIENT FOR FURTHER COMPUTATION.

TABLA 9. Valor de F y nivel de significación entre parejas de grupos al final del proceso de selección de las variables discriminantes.

En el apartado siguiente explicamos la forma como se calculan las funciones discriminantes, primero analíticamente y después mostrando los resultados obtenidos al analizar nuestros datos.

6.3. Selección de las funciones discriminantes

La función discriminante en notación matricial tiene la siguiente expresión:

$$z = a'(x_{ij} - \overline{x_j})$$

en la que x_{ij} representa las variables de origen (x_{ij} es el valor del individuo i en la variable j), $\overline{x_j}$, las medias y «a» es el vector de pesos. Cada variable queda afectada por

un peso, un valor de «a». Precisamente el objetivo del análisis es estimar los valores óptimos de «a»; conocidos los valores de «a» que corresponden a cada variable es posible estimar un valor z de pertenencia para cada individuo, como nos lo demuestra la fórmula anterior. Veamos cómo estimar los valores óptimos de «a».

Las matrices de covarianza de cada grupo tienen la siguiente expresión

$$W_i = (X_{ij}^i - \bar{X}_j^i)'(X_{ij}^i - \bar{X}_j^i)$$

Y W representa la suma de las matrices de covarianza de todos los grupos.

$$W = \Sigma W_i$$

Por otro lado, T representa la matriz de covarianza total, como si no estuviera la población distribuida en grupos.

$$T = (X_{ij} - \bar{X}_j)'(X_{ij} - \bar{X}_j)$$

Y A toma la expresión

$$A = k\, dd'$$

Siendo k una constante y d la matriz de distancia entre centroides.

A partir de la función discriminante obtenemos la siguiente expresión para un grupo cualquiera

$$\Sigma Z_i^2 = Z_i'Z_i = a'W_i a \text{ (en el grupo } i, \text{ genérico)}$$

La anterior expresión no es otra cosa que la varianza de Z en el grupo i genérico ya que consideramos que los valores de Z están normalizados.

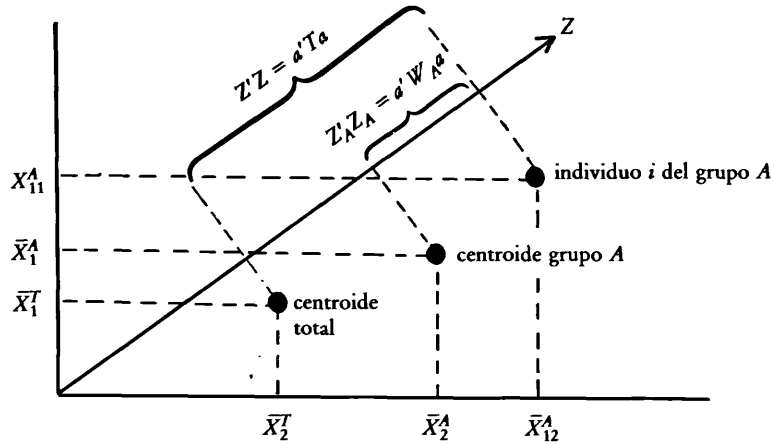
De la misma manera obtenemos la siguiente expresión que representa la varianza de Z en la muestra total.

$$\Sigma Z^2 = Z'Z = a'Ta$$

Y por último la expresión

$$\bar{Z}_i - Z_j = d'a$$

Veamos estas expresiones representadas geométricamente, pero en el caso de dos variables y un solo individuo.



$Z'Z$ es la proyección sobre el eje Z de la matriz de covarianza total T. La matriz de covarianza total si las variables están normalizadas nos indica la distancia desde el individuo a la media del total de la población o, en el caso general, la suma de distancias de toda la nube de individuos al centroide del total de la población $Z'_A Z_A$ es la proyección sobre el eje z de la matriz de covarianza del grupo A. Observemos que los valores de los pesos «a» están dando orientación al eje Z. A distintos valores de «a», encontraremos distintos ejes. Pues bien, de todos los ejes posibles habrá que conseguir uno, el óptimo, en el que la proyección sobre él, es decir $Z'_A Z_A = a' W_A a$, o, lo que es lo mismo, la suma de las distancias de toda la nube A a su centroide sea mínima y $Z'Z = a'Ta$ sea máxima.

El criterio generalmente utilizado para encontrar los valores óptimos de «a» es hacer máxima la siguiente expresión:

$$\frac{a'Ta}{a'Wa} = \frac{\Sigma Z_i^2}{\Sigma\Sigma Z_i^2}$$

Es decir, la varianza total de Z en el conjunto de los grupos, dividida por la suma de varianzas de Z en cada uno de los grupos.

Pero sabemos que:

$$a'Ta = a'Aa + a'Wa$$

luego:

$$\frac{a'Ta}{a'Wa} = 1 + \frac{a'Aa}{a'Wa}$$

El primer eje factorial discriminante será aquel que maximice el cociente anterior. Este eje factorial está, como vemos, en función de los coeficientes a, de manera que estos coeficientes deberán ser tales que maximicen dicho cociente.

Como hemos visto maximizar $\frac{a'ta}{a'Wa}$ es lo mismo que maximizar $\frac{a'Aa}{a'Wa}$. Pues bien, a esta última expresión, la llamamos λ que es un escalar, al cual denominamos valor propio o eigen valor correspondiente a un vector propio.

En realidad el cociente $\frac{a'ta}{a'Wa}$ es la varianza total de z (z eran los valores discriminantes de cada individuo, o lo que es lo mismo la reducción de los valores de las variables en cada individuo en la función discriminante) en el conjunto de los invididuos dividido por la suma de varianzas de Z en cada uno de los grupos. Vemos que cuanto más grande sea esa expresión habrá más diferencia entre el numerador y el denominador y por lo tanto el valor máximo de λ será aquella distribución de los valores de Z en la que, en cada grupo, conseguimos la mayor homogeneidad posible (varianza muy pequeña); estando al mismo tiempo unos grupos muy separados de otros ya que la varianza total o la distancia entre los centroides es muy grande. Precisamente la línea que separa los valores de Z así distribuidos será el primer eje factorial discriminante.

Pero decíamos que maximizar $\frac{a'ta}{a'Wa}$ es lo mismo que maximizar $\frac{a'Aa}{a'Wa}$. Para hacer máxima la expresión $\frac{a'Aa}{a'Wa}$, aplicamos Lagrange. Hay que tener en cuenta que tanto el numerador como el denominador, como el propio λ, son formas cuadráticas respecto «a».

Haciendo operaciones en la anterior expresión llegamos a

$$a'AA - \lambda(a'Wa) = 0$$

Y derivando parcialmente respecto a «a»

$$W^{-1}Aa = \lambda a$$

De lo cual se deduce que los coeficientes discriminantes óptimos que vamos buscando se obtienen de la siguiente ecuación:

$$a = W^{-1}d$$

«a» es el vector propio o eigenvector de la matriz $W^{-1}A$ según se desprende de la ecuación $W^{-1}Aa = \lambda a$; es decir, el primer eje factorial discriminante: en otro sentido, es el vector de pesos que proporciona la discriminación óptima. De manera que los coeficientes de la función discriminante están determinados por los vectores propios de la matriz $W^{-1}A$ que corresponden a los más grandes valores propios o eigenvalores.

Si existieran varios vectores propios, como en nuestro ejemplo, serían ortogonales y harían el papel de nuevas dimensiones (nuevas variables) en un espacio n-1 dimensional. Sus correspondientes valores propios $\lambda_1\lambda_2$ ... etc, serían cada vez más pequeños. Estos valores propios miden el poder discriminante del vector propio correspondiente, es decir el eje factorial. La suma total de los valores propios, como se desprende de lo

dicho anteriormente, indica el porcentaje total de inercia o varianza que queda explicado, es decir que se «conserva» en la operación de reducción de todo el sistema a los ejes factoriales.

Si en la expresión $W^{-1}Aa = \lambda a$, sustituimos A por su valor kdd', tenemos

$$kW^{-1}dd'a = \lambda a$$

y si ahora sustituimos «a» por su valor $W^{-1}d$ llegamos a $\lambda = Kd'W^{-1}d$, que es la distancia entre los centroides de los grupos, multiplicada por una constante, pero en la métrica W^{-1}. Esta es la llamada distancia de Mahalanobis. Si la métrica fuera $W^{-1} = I$ (matriz unidad), la distancia $d'd$ sería la euclideana. Vemos pues que maximizar λ es maximizar la distancia entre los centroides de los grupos en la métrica W^{-1}.

De manera que λ puede interpretarse como el porcentaje de varianza de Z que queda explicada o también como la distancia entre los centroides de los grupos.

Volvamos al ejemplo y observemos el siguiente cuadro (tabla 10).

CANONICAL DISCRIMINANT FUNCTIONS

FUNCTION	EIGENVALUE	PERCENT OF VARIANCE	CUMULATIVE PERCENT	CANONICAL CORRELATION	AFTER FUNCTION	WILKS' LAMBDA	CHI-SQUARED	D.F.	SIGNIFICANCE
					0	0.0128866	1131.4	110	0.0000
1*	3.69296	43.24	43.24	0.8870823	1	0.0604761	729.43	84	0.0000
2*	2.66718	31.23	74.47	0.8528252	2	0.2217768	391.58	60	0.0000
3*	1.47030	17.22	91.68	0.7714867	3	0.5478563	156.45	38	0.0000
4*	0.46041	5.39	97.07	0.5614802	4	0.8000939	57.987	18	0.0000
5*	0.24985	2.93	100.00	0.4471086		1.6430897			
	8.5407								

* MARKS THE 5 CANONICAL DISCRIMINANT FUNCTION(S) TO BE USED IN THE REMAINING ANALYSIS.

TABLA 10. Funciones discriminantes.

Este es el cuadro resumen del programa y por supuesto el más importante de todos. Cada una de las cinco últimas líneas es representativa de una función discriminante. Ya hemos dicho que puede haber tantas funciones discriminantes como variables menos una, pero sólo las primeras son verdaderamente importantes porque son las que están explicando una mayor cantidad de varianza. En el cuadro anterior vemos que la suma de las λ es igual a 8,5407 y que el primer eje factorial discriminante está explicando el 43,24% de la varianza total, lo cual a nuestro juicio es un excelente resultado para este primer factor. Sólo con este factor o esta nueva variable explicamos casi la mitad de la varianza del fenómeno que estamos estudiando y que recordemos que es la intención del voto. El 43,24% se obtiene de la siguiente expresión:

$$\frac{\lambda_i}{\Sigma\lambda_i} \times 100$$

La correlación canónica p, en la cuarta columna, va en el mismo sentido que λ y la relación entre ambas es la siguiente:

$$p = \sqrt{\frac{\lambda}{1 + \lambda}} = \sqrt{\frac{3{,}69296}{4{,}69296}} = 0{,}887082$$

Se llama correlación canónica porque es la correlación entre las nuevas variables y las primeras.

El valor Λ (Lambda de Wilks) correspondiente a la línea 0 es el calculado por el programa con todas las variables a la vez.

En la línea 1 se calcula el valor Λ correspondiente a la primera función discriminante, una vez conocidos los valores «a».

Cuando hay muchos individuos la significatividad estadística de Λ puede medirse por la distribución χ^2, con $(g - 1)p$ grados de libertad siendo g el número de grupos y p el número de variables. Ya que Λ puede transformarse en una χ^2. Barlett dio una aproximación de Λ a la χ^2 en 1938 y es la siguiente:

$$\chi^2 = -m \log_e \Lambda$$

siendo $m = n - 1 - \frac{1}{2}(p + g)$, y siendo n el número de individuos, p el número de variables y g el número de grupos.

Se computa Λ, luego se computa χ^2, se observa en la tabla el valor de χ^2 y se ve si la diferencia es significativa.

El resto de las funciones discriminantes, también se distribuye como una χ^2, así será posible conocer si existe alguna función más que posea poder discriminante. En este caso la aproximación es como sigue.

$$\chi^2 = -\left(n - \frac{p + g}{2} - 1\right) \log_e \Lambda'$$

Siendo p el número de variables, n el número de individuos y g el número de grupos.

Los grados de libertad son $p - k(g - k - 1)$, siendo k el número de funciones obtenidas.

Si el test nos da significatividad estadística, esto nos indica que al menos una función de las que restan por obtener es discriminante.

Pero recordemos que el objetivo era conseguir unos valores de «a», que son los coeficientes de la función discriminante. En el ejemplo estos valores son los de la tabla 11.

Como vemos, cada variable tiene un peso en cada función. Así, ahora será más fácil conseguir los valores z para cada individuo, sin más que multiplicar el valor que éste tiene en cada variable por su peso. Ese valor z es la proyección del sujeto correspondiente sobre la línea que representa la función discriminante.

Hay que tener en cuenta que cada función discriminante representa una nueva variable combinación lineal de las introducidas. En nuestro caso la nueva variable o función primera explica, nada menos que el 43,24% de la varianza. Un paso importante y difícil será dar nombre a esa nueva variable o función que está explicando ella sola

ROTATED STANDARDIZED DISCRIMINANT FUNCTION COEFFICIENTS

	FUNC 1	FUNC 2	FUNC 3	FUNC 4	FUNC 5
P15	1.06439	-0.04516	-0.02478	-0.03627	0.06598
P50	0.27384	-0.84206	0.26118	-0.08505	0.12165
P14	0.03880	0.11651	-0.01664	1.02923	0.06661
P53	0.49171	-0.20340	0.14706	-0.07088	0.36281
P51	0.00444	-0.26000	0.80239	-0.00788	0.26943
P22	-0.12455	-0.33494	-0.00906	0.06403	0.13721
P54	-0.03525	-0.24813	0.36395	-0.09234	0.59237
P11	-0.01148	0.28476	0.23616	0.02184	0.68165
P40	0.08927	0.38094	0.13257	-0.19557	-0.01102
P26	0.10137	-0.27492	-0.23581	0.30232	-0.15245
P03	-0.09548	-0.09990	-0.09216	0.04150	-0.23134
P86	-0.01726	0.03514	0.01016	0.26419	-0.02997
P85	-0.04768	-0.06181	-0.19606	-0.07234	-0.03068
P13	0.03155	0.41359	0.56221	0.20991	-0.49190
P90	-0.09450	0.28318	-0.08101	-0.19564	-0.03838
P19	-0.06710	0.33664	-0.08064	-0.00383	-0.22064
P28	0.03979	-0.43273	-0.00028	-0.10368	0.42677
P81	0.01223	0.23626	0.01641	-0.05059	0.21428
P60	0.09862	-0.12825	0.08350	0.22091	-0.23944
P02	-0.15417	-0.05136	-0.03633	-0.15940	0.27464
P33	0.04763	0.12804	-0.00448	0.18652	-0.02078
P24	-0.00112	0.07567	-0.06934	-0.09326	0.35525

TABLA 11. Coeficientes estandarizados de las funciones discriminantes.

casi la mitad de la varianza del fenómeno estudiado, la intención de voto. Para dar un nombre a la nueva variable será aconsejable reflexionar sobre la siguiente tabla que nos da la correlación entre las funciones discriminantes y las antiguas variables. Puede pues interpretarse el significado de las nuevas variables observando aquellas antiguas variables que tienen muy alta y muy baja correlación. El dar un nombre concrèto a la nueva variable es trabajo de expertos en sociología política.

A continuación exponemos la citada tabla (tabla 12) junto con otra (tabla 13) que nos proporciona el valor de las funciones en la media de cada grupo, pudiendo así ver la importancia de cada función para los diferentes grupos.

Los asteriscos de la tabla indican el grupo de variables que más están dotando de sentido o significado a cada función. Así, la función 1 viene definida por la variable P.15, pudiendo decir que esta función discrimina según que los grupos contesten de una u otra forma a esta pregunta.

La tabla 13 nos proporciona el valor de la función en la media de cada grupo; así

ROTATED CORRELATIONS BETWEEN CANONICAL DISCRIMINANT FUNCTIONS AND DISCRIMINATING VARIABLES
VARIABLES ARE ORDERED BY THE FUNCTION WITH LARGEST CORRELATION AND THE MAGNITUDE OF THAT CORRELATION.

	FUNC 1	FUNC 2	FUNC 3	FUNC 4	FUNC 5
P15	0.85690*	0.08177	-0.07876	-0.03856	-0.05690
P50	-0.03273	-0.68562*	-0.12645	0.03940	-0.11118
P22	-0.08119	-0.50830*	-0.10006	0.01467	0.07822
P40	0.03690	0.41940*	0.40000	-0.15155	0.08069
P33	0.13431	0.14973*	0.07576	0.06689	-0.02545
P03	-0.07787	-0.09620*	-0.03224	0.05917	0.00779
P51	-0.02686	0.12331	0.66607*	0.02055	-0.06638
P13	0.00774	0.08256	0.66446*	0.06244	-0.30410
P18	-0.00947	0.05953	0.64276*	0.04827	-0.23580
P28	-0.01371	0.03312	0.58520*	0.03577	-0.09322
P90	-0.03600	-0.02726	-0.16637*	0.02538	-0.02242
P19	-0.02569	0.11687	-0.12337*	-0.05057	-0.02651
P14	-0.00429	0.02814	-0.03603	0.83946*	0.09633
P86	-0.00512	-0.08056	0.01948	0.13048*	-0.04424
P11	-0.05601	0.15248	-0.11015	0.02212	0.55680*
P54	0.01654	0.01389	0.08040	-0.05332	0.45520*
P26	-0.02715	0.13349	-0.13102	0.06321	0.32902*
P53	0.10500	-0.05056	0.01111	-0.02369	0.25778*
P24	-0.05773	0.15954	-0.16779	0.10621	0.23282*
P81	-0.03996	0.02547	-0.08008	-0.05413	0.18967*
P02	-0.05902	-0.05949	0.01378	-0.03192	0.18967*
P83	-0.01270	0.05537	0.06511	-0.02737	-0.13575*
P60	0.06467	-0.12065	0.09196	0.10873	-0.13415*
P85	-0.02163	0.03950	0.05533	-0.02265	-0.07234*
P74	0.05646	0.04008	0.04638	0.03286	0.06743*

TABLA 12. Correlaciones entre las funciones y las variables discriminantes. Las variables están ordenadas a partir de la función con mayor correlación y la magnitud de esta correlación.

CANONICAL DISCRIMINANT FUNCTIONS EVALUATED AT GROUP MEANS (GROUP CENTROIDS)

GROUP	FUNC 1	FUNC 2	FUNC 3	FUNC 4	FUNC 5
1	-0.18671	-1.69306	-0.99768	-0.11100	-0.15782
2	-0.21439	1.55796	2.08299	-0.08488	-0.18427
3	-0.10051	1.71321	-1.89818	9.39080	0.84210
4	10.68067	0.85354	-1.00557	-0.25706	0.85252
6	-0.10246	1.12579	1.25340	-0.38248	2.23020
7	-0.43057	1.07940	-0.78406	-0.20523	-0.07761

TABLA 13. Valor de las funciones en la media de cada grupo.

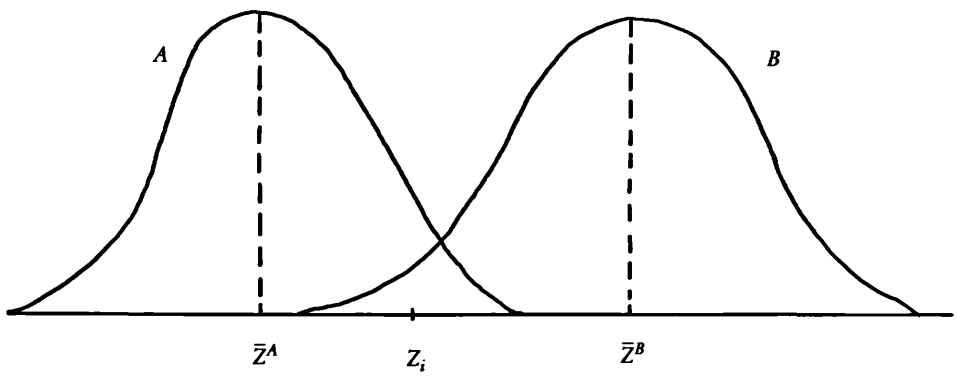
podemos ver los que realmente son muy distintos del resto. Por ejemplo, en la función 1 el grupo 4 se diferencia claramente del resto en su centroide.

6.4. La clasificación de los individuos

Pasemos ahora a estudiar la fase más interesante del análisis, comprobando su poder de predicción o clasificación.

Obtenidos los valores «a» es posible calcular para cada individuo su valor discriminante z correspondiente. Para ello, de acuerdo a la expresión $Z = a'x$, no tenemos nada más que multiplicar en cada individuo el valor de cada variable por su peso correspondiente. Posteriormente normalizamos estos valores z, de manera que podemos utilizar la tabla de la curva normal para estimar probabilidades de asignación.

Ahora bien, nos podemos encontrar con problemas de clasificación en la llamada zona crítica. En dicha zona los valores z_i de cada individuo están muy próximos y no es posible, con certeza, decir que un individuo z_i pertenece a A o a B.



Individuo en
zona crítica

Entonces, para asignar un individuo a un grupo u otro tenemos que diseñar una regla de decisión.

Pero antes veamos qué son las probabilidades a priori y a posteriori.

Observamos a un individuo cualquiera que tiene un valor z_i y nuestro problema es buscar una regla de decisión para clasificarlo en uno de los grupos; o dicho en otras palabras, buscar una regla de decisión que nos separe con el mínimo error posible los individuos pertenecientes a un grupo de los pertenecientes a otro.

Pero, ¿disponemos de alguna información antes de efectuar la clasificación de los individuos, además de, naturalmente, el valor Z?, ¿conocemos en la población la estructura probabilística del fenómeno que estamos estudiando? Puede ser que sí, y

en nuestro caso, en concreto, no estamos totalmente a ciegas. Este conocimiento de la realidad puede proceder de una encuesta anterior, del resultado de unas elecciones anteriores e incluso de los datos de la propia encuesta. Esta información, el estado en el que los individuos se encuentran en la realidad es llamado «probabilidad a priori» y se designará así: $P(A)$, $P(B)$... etc., con la condición, claro está, de que la suma de todas estas probabilidades sea la unidad.

Conviene recalcar que para calcular las probabilidades a priori, no hemos tenido en cuenta todavía los datos de la encuesta o al menos no todos, pero, en nuestro caso es aceptable suponer que a priori la probabilidad de votar PSOE no es la misma que la de votar PCE o CDS. Pues bien, la incorporación de esta probablidad a priori, mejora la predicción final.

Las probabilidades a priori introducidas han sido las siguientes:

Grupo	Probabilidades a priori
PSOE	0,44000
AP	0,28000
PCE	0,01450
CDS	0,02182
Otros	0,04727
Abstención	0,19636

Es decir, están en función del peso del propio grupo en el total de los 275 individuos que han declarado su voto. Puede que esta probabilidad no sea del todo correcta, pero en cualquier caso es mejor que no considerar ninguna y pensar que a priori todos los grupos tienen la misma probabilidad. Si tuviéramos un conocimiento más exacto de la realidad, lo reflejaríamos en estas probabilidades a priori.

Otra probabilidad a considerar es la llamada «a posteriori» y es la que resulta del análisis, al normalizar los valores de Z de cada individuo. La designaremos así $P(A/Z_i)$ y puede leerse como la probabilidad que tiene el individuo i de pertenecer al grupo A teniendo en cuenta las variables introducidas en el análisis. Se trata de una probabilidad condicional, ya que la pertenencia a un grupo está condicionada a la información de que se dispone.

Pues bien, según Bayes,

$$P(A/Z_i) = \frac{P(Z/A) \cdot P(A)}{P(Z/A) \cdot P(A) + P(Z/B) \cdot P(B)}$$

(siendo $P(A)$, $P(B)$... etc., las probabilidades a priori),
que es la regla de decisión utilizada para clasificar a un individuo que muestra un valor Z determinado. Y el individuo será asignado a aquel grupo en el que la probabilidad $P(W/Z_i)$ sea mayor. Si $P(A/Z_i) > P(B/Z_i)$ será asignado a A. Si hay varios grupos será asignado a aquel que dé mayor probabilidad.

Vemos que la tabla 15 nos proporciona para cada individuo el valor $P(W/Z)$ (en la

tabla $P(G/X)$) mayor alcanzado que corresponde con el del grupo al que se le atribuye y el siguiente mayor.

CASE SUBFILE	SEQNUM	MIS VAL	SEL	ACTUAL GROUP	HIGHEST PROBABILITY GROUP	P(X/G)	P(G/X)	2ND HIGHEST GROUP	P(G/X)	DISCRIMINANT SCORES				
NONAME	1			UNGRPD	7	0.2987	0.9366	2	0.0631	0.1535	2.9360	-1.0557	-0.3286	-0.8273
NONAME	2			UNGRPD	1	0.7574	0.9668	7	0.0332	0.1456	-1.3166	-0.7615	-0.3861	-0.8365
NONAME	3.			1	1	0.8339	0.9963	7	0.0037	-0.0001	-2.4758	-0.8452	0.2460	-0.9407
NONAME	4			2 ***	7	0.6959	0.9155	2	0.0826	-0.6117	2.3407	-1.4125	-0.9036	-0.6848
NONAME	5			1	1	0.6379	0.9989	7	0.0011	-0.2402	-3.1840	-1.0843	0.4829	-0.6458
NONAME	6			2	2	0.6575	0.9808	6	0.0188	0.2094	1.8430	3.3135	0.5181	-0.7712
NONAME	7			1	1	0.9753	0.9836	7	0.0164	-0.2556	-1.7969	-0.9897	-0.3401	-0.6718
NONAME	8			UNGRPD	6	0.0000	1.0000	1	0.0000	3.9704	0.1713	1.7977	-0.5212	1.7169
NONAME	9			1	1	0.9318	0.9962	7	0.0038	-0.1302	-2.7248	-0.8562	0.1308	-0.1362
NONAME	10			7	7	0.4313	0.7245	1	0.2661	-1.1941	0.8697	-1.9449	-0.3421	-0.3256
NONAME	11			UNGRPD	7	0.2305	0.9790	4	0.0169	-0.8632	-0.5053	0.2618	-0.2774	1.9665
NONAME	12			UNGRPD	2	0.9349	0.9775	6	0.0211	-0.0513	1.9321	3.1517	-0.1710	-0.2630
NONAME	13			1	1	0.9035	0.9825	7	0.0175	0.0281	-1.6542	-0.8736	-0.1460	-0.8931
NONAME	14			7	7	0.4262	0.9611	2	0.0369	-1.0383	1.6629	-0.9201	-0.1892	1.6386
NONAME	15			7 ***	1	0.9321	0.9669	7	0.0331	-0.6120	-1.3451	-1.2853	-0.4869	-0.0243
NONAME	16			7	7	0.6667	0.7713	2	0.2073	-0.3044	1.2360	-1.1011	0.2138	-1.6714
NONAME	17			2	2	0.2253	0.9907	7	0.0080	-0.7260	0.5751	2.5126	0.1851	-0.7197
NONAME	18			2	2	0.9070	0.8977	7	0.1008	-0.5396	1.7699	1.0941	-0.4086	-0.0356
NONAME	19			1	1	0.9611	0.9949	7	0.0051	0.0269	-2.3704	-0.7402	0.0756	-0.4208
NONAME	20			1 ***	7	0.2197	0.7820	1	0.2177	-1.0117	-0.7414	-0.4518	-0.4628	1.6798
NONAME	21			7	7	0.7333	0.9314	2	0.0356	-0.9531	1.5265	-1.7494	-0.6227	-0.6355
NONAME	22			UNGRPD	7	0.9309	0.9380	2	0.0516	-0.7148	1.8017	-1.6647	-0.2948	-0.6916
NONAME	23			UNGRPD	1	0.0190	0.7836	7	0.1688	0.7254	-0.4619	-1.0677	-0.2717	-1.7773
NONAME	24			1	1	0.9252	0.9919	7	0.0081	-0.0886	-2.0434	-1.0293	0.0503	-0.9647
NONAME	25			6	6	0.2985	1.0000	2	0.0000	-0.8013	1.6416	1.1416	-1.7857	5.7715
NONAME	26			2	2	0.7803	0.9690	6	0.0293	0.0616	1.8703	3.2483	-0.3341	-0.2644
NONAME	27			2	2	0.8223	0.9830	6	0.0165	0.1013	1.8070	3.2881	0.4064	-0.5787
NONAME	28			1	1	0.2373	0.9299	7	0.0677	0.1987	-0.6316	-1.4520	-0.4110	1.2230
NONAME	29			UNGRPD	7	0.3031	0.6443	2	0.3553	-0.9119	0.3603	1.1216	-0.6202	0.8054
NONAME	30			1	1	0.9421	0.9925	7	0.0075	-0.3583	-2.2543	-1.1462	0.1662	0.1407
NONAME	31			UNGRPD	1	0.2671	0.9763	7	0.0237	0.4110	-0.9784	-0.8812	0.2759	1.4051
NONAME	32			UNGRPD	7	0.6020	0.9616	2	0.0251	-0.3816	1.6431	-1.6760	0.5947	0.9327
NONAME	33			UNGRPD	7	0.8872	0.9366	2	0.0610	-0.3347	2.1483	-1.5218	0.1300	-0.8585
NONAME	34			1	1	0.3496	0.9914	7	0.0086	0.5798	-0.9944	-0.9154	0.5993	0.6782
NONAME	35			UNGRPD	2	0.0827	0.7316	7	0.2252	0.0995	1.9572	1.2328	0.7794	1.6962
NONAME	36			UNGRPD	7	0.8394	0.9665	2	0.0331	-0.4661	2.6604	-1.4442	-0.3421	-0.9176
NONAME	37			UNGRPD	1	0.1189	0.9710	7	0.0289	0.4448	-1.1456	-0.9975	0.7588	1.2917
NONAME	38			UNGRPD	2	0.7910	0.9832	6	0.0126	-0.5347	1.2491	2.9057	-0.4848	0.0922
NONAME	39			7 ***	1	0.5247	0.8361	7	0.1623	-0.0607	-0.5219	-0.9644	-0.6415	-0.8549
NONAME	40			UNGRPD	7	0.3472	0.9207	3	0.0448	-0.1836	1.3855	-1.5083	1.0384	0.4518
NONAME	41			1	1	0.0559	0.9108	7	0.0863	-1.4689	0.3101	-2.0365	-0.0896	-0.8164
NONAME	42			UNGRPD	1	0.8756	0.9880	7	0.0120	-0.5224	-2.3262	-1.0893	-0.5328	0.1177
NONAME	43			UNGRPD	2	0.8569	0.8743	7	0.1161	-0.5617	1.5568	1.1086	-0.4121	0.5764
NONAME	44			2	2	0.1211	0.9940	7	0.0039	-0.5240	0.4157	3.0175	0.6915	-0.2547
NONAME	45			UNGRPD	7	0.9279	0.9310	2	0.0628	-0.5527	1.8557	-1.6265	-0.1272	-0.9804
NONAME	46			UNGRPD	2	0.5793	0.8538	7	0.1461	-0.7310	1.5010	0.9988	-0.7340	-0.4570
NONAME	47			UNGRPD	7	0.8604	0.9438	2	0.0551	-0.4639	2.3135	-1.3473	-0.0513	-1.1915
NONAME	48			1	1	0.2585	0.9650	7	0.0302	0.5885	-0.8015	-0.6246	0.2843	-1.0307

TABLA 15. Cuadro de clasificación del análisis parcial.

Y así podemos ver cómo el partido al que el individuo ha declarado su intención de votar en la encuesta puede no corresponder con el partido al cual le ha asignado el análisis. Cuando no coinciden el partido o grupo al que le atribuye el análisis con el declarado por el sujeto, el programa anota 3 asteriscos.

Pues bien, hasta ahora sólo hemos tenido en cuenta a los 275 sujetos que han declarado su voto y por lo tanto formaban parte de alguno de los 6 grupos constituidos a priori. Pero nada hemos dicho todavía de los 217 que no declaran intención de voto y por lo tanto no es posible atribuirles a grupo alguno. Son los que en la tabla 15 aparecen con el nombre de «ungrpd» (no agrupados).

Pues bien, el análisis pasa por cada uno de ellos las funciones discriminantes y calcula una probabilidad de pertenencia, atribuyéndoles a aquel grupo en el que consiga un mayor valor. He aquí, pues, el otro objetivo del análisis, su capacidad de predicción.

De esta manera podemos construir la llamada «matriz de confusión» que relaciona los valores reales de pertenencia a grupos con los valores predichos por el análisis (tabla 16).

CLASSIFICATION RESULTS -

ACTUAL GROUP	NO. OF CASES	PREDICTED GROUP MEMBERSHIP 1	2	3	4	6	7
GROUP 1	121	110	1	0	0	2	8
		90.9%	0.8%	0.0%	0.0%	1.7%	6.6%
GROUP 2	77	2	68	0	0	1	6
		2.6%	88.3%	0.0%	0.0%	1.3%	7.8%
GROUP 3	4	0	0	4	0	0	0
		0.0%	0.0%	100.0%	0.0%	0.0%	0.0%
GROUP 4	6	1	0	0	5	0	0
		16.7%	0.0%	0.0%	83.3%	0.0%	0.0%
GROUP 6	13	0	6	0	0	7	0
		0.0%	46.2%	0.0%	0.0%	53.8%	0.0%
GROUP 7	54	13	10	1	0	1	29
		24.1%	18.5%	1.9%	0.0%	1.9%	53.7%
UNGROUPED CASES	217	71	44	3	3	9	87
		32.7%	20.3%	1.4%	1.4%	4.1%	40.1%

PERCENT OF "GROUPED" CASES CORRECTLY CLASSIFIED: 81.09%

TABLA 16. Matriz de confusión. Relaciona pertenencias reales con las predichas por el análisis.

Vemos que en el 81,09% de los casos, el valor real de pertenencia al grupo declarado por el sujeto coincide con el predicho por el análisis. Este es un excelente resultado. Los casos de no coincidencia pueden deberse a múltiples causas, entre otras a que el sujeto no haya sido sincero en su declaración de intención de voto.

En cualquier caso lo más importante es que los 217 sujetos que no han declarado intención de voto, el análisis los atribuye a grupos concretos.

	Individuos anónimos	
	Números absolutos	%
PSOE	71	33
AP	44	20
PCE	3	1
CDS	3	1
Otros	9	5
Abstenciones	87	40
Total	217	100

Ahora podemos agregar al voto declarado en la encuesta, el voto predicho por el análisis de estos 217 individuos.

	Voto declarado (275) (%)	Voto declarado + Voto predicho (496) (%)
PSOE	44	39
AP	28	25
PCE	1	1
CDS	2	2
Otros	5	5
Abstenciones	20	28
Total	100	100

Apéndice

Variables utilizadas en el análisis

P.15. Pregunta sobre la importancia que le da el entrevistado a un determinado problema de tipo económico. La respuesta se recoge en una escala de 0 a 10 puntos.
P.50. ¿Votó PSOE en las últimas elecciones? Si-No.
P.14. Pregunta sobre la importancia que se da a un problema nacional de tipo autonómico. Respuestas de 0 a 10.

P.53. ¿Votó al CDS? Si-No.
P.51. ¿Votó a AP? Si-No.
P.22. Pregunta acerca del partido que mejor puede resolver un problema nacional de tipo económico. La respuesta es: PSOE-otro.
P.54. ¿Votó a UCD? Si-No.
P.11. Pregunta por la importancia que se da a un problema nacional de tipo social. La respuesta de 0 a 10.
P.40. ¿Autoubicación política del entrevistado en una escala de 0 a 7?
P.26. Pregunta sobre cuál es el partido que mejor puede resolver un problema nacional de tipo autonómico. La respuesta es: PSOE-otro.
P.03. ¿Tamaño del municipio de residencia?
P.86. Pregunta si el entrevistado tiene una determinada ocupación. La respuesta es: Si-No.
P.85. Pregunta similar a la anterior, pero preguntando por otra ocupación.
P.13. Pregunta por la importancia que da el entrevistado a un determinado problema de tipo nacional. La respuesta: Si-No.
P.90. ¿Actitud hacia la religión? Las respuestas en una escala de 1 a 5 puntos.
P.19. Pregunta sobre el acuerdo del entrevistado con una determinada frase en relación a la autonomía de la región. La respuesta es: Acuerdo-Desacuerdo.
P.28. Igual que la anterior pero con una frase sobre las instituciones regionales.
P.81. Pregunta si el entrevistado tiene una determinada ocupación. La respuesta es Si-No.
P.60. ¿El sexo del entrevistado es hombre? ¿Sí-No.
P.02. ¿Edad?
P.33. ¿Nivel de estudios?
P.24. ¿Ingresos?

7. Aspectos teóricos del análisis de cluster y aplicación a la caracterización del electorado potencial de un partido

por *Emilio Martínez Ramos*

7.1. Introducción

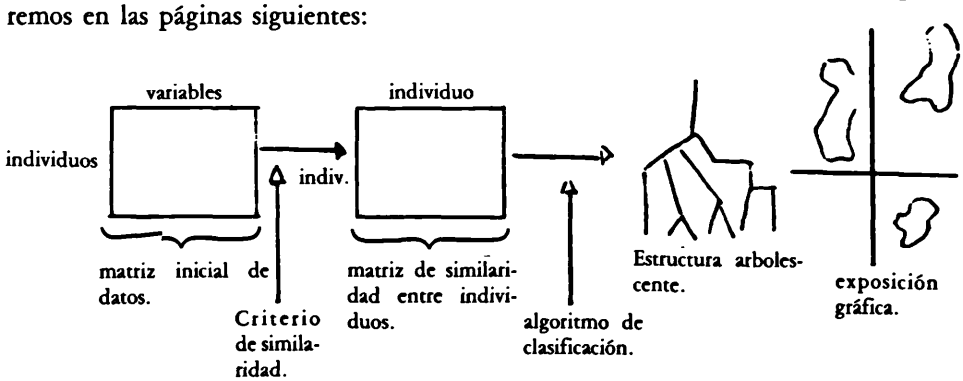
El objetivo del análisis de cluster es el siguiente: dado un conjunto de individuos (M) y teniendo de cada uno de ellos una información (N), el análisis será capaz de clasificarlos en grupos de manera que los individuos pertenecientes a un grupo (y siempre con respecto a la información de que se dispone) serán tan similares como sea posible. Así pues, un cluster es para nosotros un conjunto de individuos similares. De todas maneras, cualquier definición que se dé del cluster a estas alturas, sin profundizar más en el análisis, será ambigua, porque la propia definición de un cluster está en función de los algoritmos de clasificación que se hayan empleado para agrupar a los sujetos; los cuales pueden ser muy simples, con un fondo estadístico muy sencillo (como por ejemplo muchos de los utilizados en biología) o muy complicados.

Esta técnica de clasificación tiene su origen en la biología, ciencia en la que el problema de la clasificación de las especies adquiere gran relieve. El propio Robert R. Sokal (1977) decía que la clasificación es uno de los procesos fundamentales en la ciencia y que los fenómenos deben ser ordenados para que podamos entenderlos.

En este autor, Robert R. Sokal, y en P. H. A. Sneath arranca una de las líneas más fecundas del análisis cluster. Los libros *The principles of numerical taxonomy* (1963) y *Numerical taxonomy* (1972) son aportaciones decisivas para el progreso de esta técnica de clasificación automática.

Aunque esta técnica naciera en la biología, ha sido después aplicada con éxito en muchos campos, incluido el campo de la sociología empírica, pero especialmente en medicina, psiquiatría, arqueología, antropología.

En el siguiente cuadro se reflejan los pasos fundamentales del análisis que explicaremos en las páginas siguientes:



Como en otras técnicas del análisis de datos, partimos de una matriz variables/individuos, N × M, siendo N el número de variables que hemos tenido en cuenta y M el número de individuos[1]. Esta matriz no es otra cosa que el cuestionario cumplimentado por la persona entrevistada. En esta matriz el valor de una celdilla representa la contestación de un individuo a una variable concreta. Estas variables pueden ser de cualquier tipo, ordinales, nominales o intervales, e incluso será posible realizar el análisis con distintos tipos de variables a la vez.

En este tipo de análisis deberemos poner especial cuidado en la selección de las variables de partida que van a caracterizar a cada individuo y en función de ellas se van a agrupar. Hay que tener en cuenta que en este análisis, a diferencia de otros, no hay variable dependiente. Los grupos, aquí, se configuran por sí mismos. De ahí también el carácter exploratorio de este análisis —como decía Kruskal (1977: 17) «aplicar el cluster analysis para vagos propósitos». Junto a una buena selección de las variables (para lo cual tal vez sea necesario realizar previamente otro tipo de análisis de datos) habrá también que poner especial cuidado en el criterio de similaridad que se utilice. Como veremos, la literatura sobre el tema y los programas de ordenador disponibles nos muestran una amplia gama de criterios. Igualmente deberemos seleccionar uno de los procedimientos de agregación o desagregación de los sujetos, mejor llamados algoritmos de clasificación, que nos presentan los programas al uso.

Aunque el objetivo prioritario del análisis es la clasificación de los sujetos, cabe dotarle de una cierta capacidad de predicción. Si un sujeto pertenece a un grupo determinado en el que existe un alto grado de homogeneidad en relación a un conjunto de variables, por ejemplo relativas a la actitud política, será posible atribuir a dicho sujeto, con un cierto riesgo, el valor de una característica que le define pero que desconocemos porque no la ha declarado, por ejemplo, la intención de voto.

De acuerdo con el esquema anterior, vamos a reflexionar primero sobre la elección de las variables y sobre los criterios de similaridad, y después sobre el proceso de agregación o desagregación de los sujetos o algoritmos de clasificación.

7.2. Selección de las variables y criterios de distancia y similaridad

A la hora de analizar los datos nos encontramos con los siguientes problemas que nos pueden plantear las variables seleccionadas. Fundamentalmente hay tres problemas:

—que las variables estén en diferentes unidades
—que las variables estén correlacionadas
—que haya un número de variables muy grande.

En el caso de que las variables estén expresadas en diferentes unidades, lo que es habitual en las investigaciones sociológicas, convendrá «normalizarlas» (llevarlas a una

[1] Puede leerse en el libro *Eléments d'analyse de données* de E. Diday y otros, diferentes tipos de tablas en que pueden venir los datos. Por la brevedad de este artículo nosotros nos referimos a la tabla de datos más habitual, que relaciona los individuos con las variables a través de un número. Esta tabla se llama cuantitativa.

métrica común), con objeto de evitar las posibles incidencias en los grupos formados a posteriori.

Normalizar una variable como sabemos es evitar la incidencia de su unidad de medida. Por ejemplo, si una variable es el ingreso familiar y otra variable es el número de hijos, observamos que ambas variables tienen distinta unidad de medida y que la variable «ingresos» tiene un espectro mucho más amplio de posibles valores; de manera que, en esta variable, los cambios de un individuo a otro pueden ser muy grandes, e influir decisivamente en el índice de similaridad, y por lo tanto en la formación final de los grupos.

Así pues, nos interesa eliminar ese efecto y llevar ambas variables a la misma unidad, para lo cual es necesario «normalizar». La fórmula más habitual para normalizar una variable es:

$$\frac{X - \bar{X}}{\sigma}$$

donde X es el valor de la variable, $\bar{X}$ su media y σ su desviación típica, de esta forma, siempre, la variable normalizada tendrá la media 0 y la varianza 1.

Para resolver el segundo y tercer problemas deberemos usar el método de los componentes principales que, como sabemos, reduce el número de variables a aquellas más significativas, las que explican una mayor cantidad de varianza y que no están correlacionadas entre sí (véase Batista, en este mismo libro)[2].

El método de los componentes principales es una simple rotación de los ejes de la que resulta un nuevo conjunto de coordenadas para cada punto. Aunque hayamos efectuado una rotación de los ejes, las distancias entre los puntos no han sido modificadas si la matriz de partida es no singular; y si es singular, la distancia antes es mayor que la distancia después de la rotación, debido a que incluye información redundante. Una matriz singular quiere decir que una variable está exactamente determinada por una combinación lineal de otras variables.

Cuantas más componentes utilicemos la distancia entre dos puntos después de aplicar la técnica de los componentes principales se aproxima cada vez más a la distancia entre los puntos sobre el espacio de las variables originales. Cuando el número de componentes sea el total posible (tantos como variables), entonces las distancias derivadas de ambos espacios son las mismas, claro está, siempre teniendo en cuenta que la matriz sea no singular.

En este caso, cuando hemos utilizado todos los componentes principales, las distancias son euclideanas y están todas ellas dentro de una elipse multidimensional. Si utilizamos, como es costumbre, sólo algunos de los componentes principales, dos o tres, es lo mismo, también nos encontraremos con distancias (distintas a las del espacio de las variables originales) dentro de una elipse formada por menos dimensiones.

---

[2] El efecto de esta solución dependerá de la técnica de clasificación utilizada. En el caso de algunas técnicas basadas en la optimización (algoritmos de asignación), el uso de las componentes principales sólo servirá para reducir el número de variables (véase Everitt, 1974: 48-49). En este caso, para resolver los problemas de estandarización y de correlación entre las variables es necesario utilizar la distancia de Mahalanobis (véase *infra*).

Un último problema que vamos a mencionar es el hecho de que no todas las variables tienen la misma importancia desde el punto de vista del problema sociológico que nos planteamos.

El hecho de que una igual distancia geométrica entre los puntos, puede no indicar una igual distancia desde un punto de vista sociológico es importante tenerlo en cuenta, y nos introduce en el tema de la ponderación de las variables.

Precisamente una forma para ponderar la importancia de las variables consiste en estandarizar los componentes principales. Esta estandarización convertirá la elipse multidimensional en una esfera multidimensional.

Para resolver dos de los problemas que acabamos de mostrar, uno el de la correlación entre variables y otro el de que las variables vengan en distintas unidades, se suele utilizar la distancia de Mahalanobis.

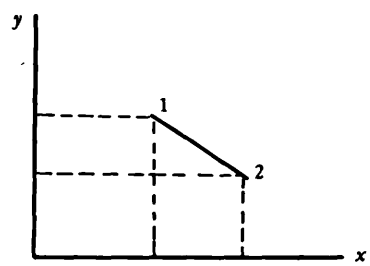
A partir de la matriz de datos de entrada podemos medir lo similares que son dos individuos genéricos en función de los valores que en cada uno de ellos tomen las variables introducidas. Para medir esto es necesario elegir un criterio de similaridad. Supongamos que tenemos dos individuos que han contestado a un cuestionario de 60 preguntas, pues bien, deberemos cuantificar lo similares o no que son esos dos individuos en función de sus contestaciones a esas 60 preguntas. No podemos decir que son muy o poco similares, debemos llegar a una cantidad que mida exactamente su similaridad.

En este sentido existe una legión de índices de similaridad y de distancia entre individuos, y el investigador debe ser consciente de cuál está usando en cada momento.

La mayor parte de ellos pueden ser clasificados en 1) criterios basados en la distancia (considerando a los individuos como vectores en un espacio n-dimensional (apartados 7.2.1 y 7.2.2) 2) criterios basados en coeficientes de correlación (7.2.3 a 7.2.5); 3) coeficientes basados en tablas de datos de posesión o no posesión de una serie de atribuciones (7.2.6). Veamos algunos de ellos.

7.2.1. La distancia euclideana

Esta medida de similaridad es la de más fácil comprensión. Si tenemos dos sujetos y dos variables que les definen puede comprenderse la noción de distancia entre ellos, sin más que considerar a los individuos como puntos de un espacio de dos dimensiones (tantos como variables) y a la variable como la proyección de esos puntos sobre los ejes de coordenadas.



Si los individuos están situados en el punto 1 y 2 del plano y los valores de las variables son las proyecciones de cada punto sobre los ejes (que representan las variables) entonces la distancia entre ellos es:

$$d_{12} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Si en vez de dos dimensiones (variables), tenemos n dimensiones, la distancia entre el punto 1 y 2 será imposible representarla pero fácil expresarla algebraicamente ya que lo único que tendremos que hacer es generalizar la fórmula anterior a n casos.

$$d_{12} = \sqrt{\sum_{k=1}^{n} (x_{2k} - x_{1k})^2}$$

en la que k representa el número de dimensiones o variables.

Si tenemos m individuos llegamos a la siguiente expresión

$$d_{ij}^2 = \sum_{i,j=1}^{m} \sum_{k=1}^{n} (x_{ik} - x_{jk})^2$$

en notación matricial:

$$d_{ij}^2 = (x_i - x_j)'(x_i - x_j) = d'd$$

Naturalmente, cuanto menor sea la distancia entre dos puntos (individuos) más similaridad hay entre ellos.

7.2.2. *La distancia de Mahalanobis (1936)*

Si decíamos que la distancia euclideana en notación matricial era:

$$d_{ij}^2 = (x_i - x_j)'(x_i - x_j) = d'd$$

La distancia de Mahalanobis, también en notación matricial es:

$$d_{ij}^2 = (X_i - X_j)' \; W^{-1}(X_i - X_j)$$

Siendo W la matriz de covarianza.

Como sabemos, la distancia en un espacio vectorial euclideano está configurada por dimensiones que serán vectores unitarios. La distancia de Mahalanobis no es del espacio euclideano, sino que nos situamos en otra métrica.

Un espacio con distinta métrica a la euclideana significa que se han ponderado las variables, se las ha llevado a una misma unidad de medida.

La ventaja de la distancia de Mahalanobis sobre la distancia euclidea es que permite que las variables estén correlacionadas. En el caso de que las correlaciones sean cero, la distancia de Mahalanobis es igual a la euclidea medida con variables estandarizadas.

7.2.3. *El coeficiente de correlación producto momento de Pearson*

Uno de los coeficientes más empleados para medir la similaridad entre dos individuos es el coeficiente de correlación de Pearson. Generalmente se emplea para datos de tipo cuantitativo y preferente en el sistema *Nearest neighbour* que explicaremos más adelante.

7.2.4. *Coeficiente de correlación de rangos de Kendall*

Este coeficiente se usa cuando en la tabla de datos de entrada los individuos ordenan una serie de características.

Kendal propone un coeficiente τ (tau) para comparar dos órdenes. Consideramos dos individuos $i\,j$ que ordenan simultáneamente una serie de características, o una serie de individuos que son ordenados por dos jueces.

Se establece el número de concordancias y el número de discordancias. Para ello se calculan todas las parejas posibles. Por ejemplo si se está ordenando a una serie de individuos por dos jueces, tomamos una pareja $i\,j$; hay concordancia si el orden de $i\,j$ es igual en los dos jueces, es decir si i está delante de j en el 1.er juez y también lo está en el 2.° juez. Hay discordancia si sucede lo contrario.

La fórmula es:

$$\tau = \frac{2}{n(n-1)} (a - b)$$

Si hay concordancia total τ vale 1 y si hay discordancia total τ vale -1. a es el número de concordancias y b el número de discordancias. $n(n-1)$ es el número total de parejas. La fórmula propiamente es

$$\tau = \frac{a - b}{\frac{n(n-1)}{2}}$$

7.2.5. *Coeficiente de correlación de rangos de Spearman*

Este coeficiente se usa cuando tenemos variables ordinales. Por ejemplo una serie de n sujetos son clasificados por dos jueces. A una clasificación la llamamos x, a la otra y.

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} (x_i - y_i)^2}{n(n^2 - 1)}$$

o también

$$r_s = \frac{3}{n-1}\left[\frac{4\sum_{i=1}^{n} X_i Y_i}{n(n+1)}\right] - (n+1)$$

$x_i - y_i$ es la diferencia entre la posición del individuo i según x y según y.

7.2.6. *Los coeficientes de asociación (para variables dicotómicas)*

Si las variables que definen a la población son dicotómicas, es decir de presencia/ausencia; posesión/no posesión, es posible emplear un coeficiente de similaridad entre sujetos llamado *matching type* y que es muy sencillo. Consiste en establecer un cociente entre el número de coincidencias en las variables entre cada dos individuos y el total de variables. Existe una amplia gama de coeficientes *matching type,* expondremos algunos de ellos.

Si tenemos dos individuos definidos por cinco variables, del tipo de las anunciadas anteriormente, de posesión/no posesión, y codificamos la posesión con un 1 y la no posesión con un 0.

	Variables				
Individuos	*a*	*b*	*c*	*d*	*e*
A	1	1	0	0	1
B	0	1	0	1	0

Un coeficiente de similaridad sería:

$$C_1 = \frac{m}{M}$$

Siendo m el número de variables que son comunes a los dos individuos (concordancia) y M el número total de variables. En el ejemplo el coeficiente C sería 2/5.

Otro posible coeficiente sería: (Tanimoto)

$$C_2 = \frac{h}{H}$$

Siendo h, en este caso, el número de variables codificadas con un 1 para los dos individuos, y H el número de variables codificadas con un 1.

$$V = \sum_{j=a}^{e} X_{Aj}(1 - X_{Bj}) = X_{Aa}(1 - X_{Ba}) + X_{Ab}(1 - X_{Bb}) + X_{Ac}(1 - X_{Bc}) +$$

$$+ X_{Ad}(1 - X_{Bd}) + X_{Ae}(1 - X_{Be}) = 1(1-0) + 1(1-1) + 0(1-0) + 0(1-1) + 1(1-0) = 2$$

Llamemos a *U:*

$$U = R + S + T + V = 1 + 1 + 1 + 2 = 5$$

Teniendo en cuenta esta nomenclatura veamos algunos índices de similaridad. Todos estos índices de similaridad varían de 0 a 1.

RUSEEL-RAD $$C_3 = \frac{R}{U} = \frac{1}{5}$$

KENDALL $$C_4 = 1 - \frac{V + T}{U} = 1 - \frac{2 + 1}{5} = 1 - \frac{3}{5} = \frac{5}{5} - \frac{3}{5} = \frac{2}{5}$$

JOCCARD $$C_5 = \frac{R}{R + T + V} = \frac{1}{1 + 1 + 2} = \frac{1}{4}$$

ROGER y TANIMOTO $$C_6 = \frac{U - (T + V)}{U + (T + V)} = \frac{5 - (1 + 2)}{5 + (1 + 2)} = \frac{2}{8} = \frac{1}{4}$$

YULE $$C_7 = \frac{RS - TV}{RS + TV} = \frac{1.1 - 1.2}{1.1 + 1.2} = -\frac{1}{3}$$

OCHIAI $$C_8 = \frac{R}{(R + T)(R + V)} = \frac{1}{(1 + 1) + (1 + 2)} = \frac{1}{6}$$

BRAVAIS-PEARSON $$C_9 = \frac{RS - QTV}{(R+T)(R+V)(S+T)(S+V)} = \frac{1.1 - 1.2}{(1+1)(1+2)(1+1)(1+2)} = \frac{1}{18}$$

SOKAL y SNEATH $$C_{10} = \frac{R}{R + 2(T + V)} = \frac{1}{1 + 2(1 + 2)} = \frac{1}{7}$$

KULCZINSKY (1) $$C_{11} = \frac{R}{T + V} = \frac{1}{1 + 2} = \frac{1}{3}$$

Otros coeficientes de asociación para tablas 0, 1 ó matrices lógicas pueden consultarse en los libros y artículos de R. P. Sokal y P. H. A. Sneath (1963, 1973), o en el libro de J. P. Benzecri (1976). Recogemos a continuación algunos de estos coeficientes del último autor citado, para lo cual habrá que definir la siguiente nomenclatura.

Sobre el ejemplo anterior:

Individuos	Variables *a*	*b*	*c*	*d*	*e*
A	1	1	0	0	1
B	0	1	0	1	0

Llamando a *R*, *S*, *T* y *V*,

$$R = \sum_{j=a}^{e} X_{Aj}X_{Bj} = X_{Aa}X_{Ba} + X_{Ab}X_{Bb} + X_{Ac}X_{Bc} + X_{Ad}X_{Bd} +$$

$$+ X_{Ae}X_{Be} = 1.0 + 1.1 + 0.0 + 0.1 + 1.0 = 1$$

$$S = \sum_{j=a}^{e}(1 - X_{Aj})(1 - X_{Bj}) = (1 - X_{Aa})(1 - X_{Ba}) + (1 - X_{Ab})(1 - X_{Bb}) +$$
$$+ (1 - X_{Ac})(1 - X_{Bc}) + (1 - X_{Ad})(1 - X_{Bd}) + (1 - X_{Ae}(1 - X_{Be}) =$$
$$= (1 - 1)(1 - 0) + (1 - 1)(1 - 1) + (1 - 0)(1 - 0) + (1 - 0) +$$
$$(1 - 1) + (1 - 1)(1 - 0) = 1$$

$$T = \sum_{j=a}^{e}(1 - X_{Aj})X_{Bj} = (1 - X_{Aa})X_{Ba} + (1 - X_{Ab})X_{Bb} + (1 - X_{Ac})X_{Bc} +$$

$$+ (1 - X_{Ad})X_{Bd} + (1 - X_{Ae})X_{Be} = (1 - 1)0 + (1 - 1)1 + (1 - 0)0 +$$
$$(1 - 0)1 + (1 - 1)0 = 1$$

KULCZINSKY (2)

$$C_{12} = \frac{R}{2}\frac{1}{R + T} + \frac{1}{R + V} = \frac{1}{2}\frac{1}{1 + 1}$$

$$+ \frac{1}{1 + 2} = \frac{1}{2}\frac{1}{2} + \frac{1}{3} = \frac{5}{12}$$

DICE/SORENSEN

$$C_{13} = \frac{2R}{2R + T + V} = \frac{2.1}{2.1 + 1 + 2} = \frac{2}{5}$$

De todos estos coeficientes, de acuerdo con Everitt (1974) quizá los más utilizados sean el C_5 y el C_{13}.

7.2.7. *Coeficiente de asociación para variables binarias, cualitativas y cuantitativas*

Gower (1971) define un coeficiente que se puede utilizar con cualquier tipo de datos.

$$S_{ij} = \sum_{k=1}^{p} S_{ijk} / \sum_{k=1}^{p} W_{ijk}$$

El peso W_{ijk} es igual a 1 o a 0 dependiendo de que la comparación considerada sea válida para la variable k, y, excepto en el caso de variables dicotómicas, este peso sólo puede ser cero cuando se desconozca el valor de la variable k para uno o ambos de los individuos. En el caso de variables dicotómicas W_{ijk} es igual a cero cuando la variable k esté «ausente» en ambos individuos.

Los valores de S_{ijk} se calculan de la siguiente manera:

- Datos binarios: en este caso el coeficiente de Gower es igual al coeficiente de Joccard (C_5 en el apartado anterior).
- Datos nominales: en este caso S_{ijk} es igual a 1 si los dos individuos son iguales en la variable k, y $S_{ijk} = 0$ cuando difieren.
- Datos intervales: en este caso $S_{ijk} = 1 - |X_{ik} - X_{jk}| / R_k$, donde X_{ik} es el valor del individuo i en la variable K, y R_k es el recorrido de la variable.

Tomando el ejemplo de Everitt (1974: 55 y 56) vamos a mostrar el cálculo del coeficiente de Gover.

	Altura (pulgadas)	Peso (libras)	Color de ojos	Color de pelo	Fuma/ no fuma
Individuo 1	66	120	Azul	Rubio	Fuma
Individuo 2	72	130	Verde	Negro	Fuma
Individuo 3	70	150	Azul	Rubio	No fuma

Veamos el coeficiente de similaridad para los individuos 1 y 2.

$$S_{12} = \frac{(1 - |66 - 72|/6) + (1 - |120 - 130|/30) + 0 + 0 + 1}{5} = 0.334$$

De igual forma calcularíamos $S_{13} = .446$ y $S_{23} = .220$.

7.2.8. *Transformación de distancias a similaridades: coeficiente de similaridad de Cattell*

Digamos que la diferencia más aparente entre las distancias y las similaridades es que mientras que las primeras pueden tomar cualquier valor positivo, las segundas sólo aceptan valores entre 0 y 1. Sin embargo, es posible transformar cualquier distancia en una similaridad, por ejemplo mediante una transformación como la de Cattell.

El coeficiente de Cattell es una transformación monótona de la distancia euclideana, cuando las variables están normalizadas y tienen varianza la unidad.

Toma la siguiente expresión:

$$c = \frac{E_i - \sum_{i=1}^{n} d_{jk}^2}{E_i + \sum_{i=1}^{n} d_{jk}^2}$$

En la que i es el número de dimensiones o variables; n es el número de pares; d^2_{jk} es la distancia euclideana al cuadrado con variables normalizadas entre los individuos j y k; E_i es dos veces el valor χ^2 para n grados de libertad. Este coeficiente varía de $+1$ a -1.

7.3. Algoritmos de clasificación

De acuerdo con los criterios de similaridad y distancia dados en las páginas anteriores, ya tenemos formada una matriz de similaridad entre los sujetos. Ahora debemos estudiar la manera cómo se pueden formar los grupos de individuos. En cada casilla de la matriz de similaridad o de distancia tenemos un número que es reflejo de la medida de similaridad entre cada par de sujetos. Esta matriz, por ser simétrica, sólo utiliza la mitad de las posiciones y, además, la diagonal no tiene sentido ya que nos daría la similaridad de cada sujeto consigo mismo.

	Individuos 1, 2, 3, ... N
Individuos	1 2 3 . . . . N

Para la constitución de los conglomerados o clusters caben diferentes procedimientos. En este apartado vamos a hacer mención a técnicas jerárquicas, técnicas basadas en la partición y a otras técnicas no clasificables en los grupos precedentes, como el análisis factorial de tipo Q. Dentro de las técnicas jerárquicas, las más simples y comunes, cabe distinguir entre métodos aglomerativos o ascendentes y métodos disociativos o descendentes. En los primeros, mediante la utilización de algún criterio se van agrupando los individuos en cada paso hasta llegar a un conglomerado que engloba a la totalidad. En los segundos, partiendo del conjunto de los individuos como un conglomerado y siguiendo también algún criterio, se procede dividiéndolos en grupos más pequeños y homogéneos hasta llegar en última instancia a cada uno de los sujetos como conglomerado más simple y de máxima homogeneidad. Entre los métodos aglomerativos o ascendentes incluimos en este trabajo el método de las distancias mínimas, el método de las distancias máximas, el método de las distancias entre centroides y el de las distancias ponderadas. Respecto de los métodos divisivos o descendentes incluimos el de William y Lambert.

Las técnicas de partición difieren de las técnicas jerárquicas en el hecho de que admiten reasignación de los individuos (objetos) con el fin de corregir en una etapa posterior una mala partición inicial. La mayoría de estas técnicas se pueden describir co-

mo intentos de asignar los individuos a conglomerados tratando de optimizar algún criterio predefinido. Como ilustración de esta técnica incluimos el método *K*-means (MacQueen, 1966), que será el que se utilice en el ejemplo que utilizamos en este trabajo.

El último de los métodos que vamos a explicar es la técnica *Q* del análisis factorial, que consiste en agrupar los individuos en lugar de las variables.

7.3.1. *Método de las distancias mínimas*

En inglés llamado *Single Linkage* o *Nearest Neighbour*. El proceso comienza con todos los individuos, considerados cada uno como un cluster separado. En primer lugar se calcula la distancia entre cada par de individuos. El proceso continúa uniendo un individuo a un cluster o un cluster a un cluster, de acuerdo al criterio de la mínima distancia entre los dos individuos más próximos, perteneciendo cada uno a cluster separados. En cada etapa del proceso el número de cluster formados disminuye.

Veamos un ejemplo.

Vamos a suponer que tenemos 6 individuos y que las distancias entre ellos vienen representadas en esta matriz.

	1	2	3	4	5	6
1	—	9	9	8	2	9
2		—	4	1	6	4
3			—	7	6	8
4				—	5	5
5					—	3
6						—

Vemos que los dos individuos más próximos son el 2 y el 4, que quedan agrupados. Para calcular la distancia entre el cluster formado por el individuo 2 y 4 y el resto de los individuos se emplea, como hemos dicho antes, el criterio de «distancia mínima». Así la distancia entre el cluster (2,4) y el individuo 1 es 8, porque 8 es la mínima distancia entre el cluster y el individuo 1; en concreto entre el individuo 4 (que forma parte del cluster) y el 1 —del individuo 2 al 1 la distancia sería 9—. Las distancias entre el cluster (2,4) y los individuos 3, 5 y 6 serían 4, 5 y 4, respectivamente:

$$d_{(24)3} = \min\{d_{23}, d_{43}\} = d_{23} = 4$$

$$d_{(24)5} = \min\{d_{25}, d_{45}\} = d_{45} = 5$$

$$d_{(24)6} = \min\{d_{26}, d_{46}\} = d_{26} = 4$$

Entonces se vuelve a elaborar la matriz, pero ahora habiendo unido los individuos 2 y 4.

	1	(2,4)	3	5	6
1	—	8	9	2	9
(2,4)		—	4	5	4
3			—	6	8
5				—	5
6					—

Los dos más próximos, ahora, son el 1 y el 5. Volvemos a calcular las distancias (mínimas) entre el nuevo cluster (1,5) y todos los demás, sean clusters o individuos. Por ejemplo, la distancia entre el cluster (1,5) y el cluster (2,4) es 5, la más próxima entre ambas (entre el individuo 5 y el 4):

$$d_{(15)(24)} = \min\{d_{12}, d_{14}, d_{52}, d_{54}\} = d_{54} = 5$$

Y volvemos a elaborar la matriz, después de calcular el resto de las distancias.

	(1,5)	(2,4)	3	6
(1,5)	—	5	6	5
(2,4)		—	4	4
3			—	8
6				—

Ahora la menor distancia es la que hay entre el grupo (2,4) y el individuo 3. Por lo tanto 2,4 y 3 forman un nuevo conglomerado y procedemos a calcular sus distancias respectivas a los otros grupos o individuos.

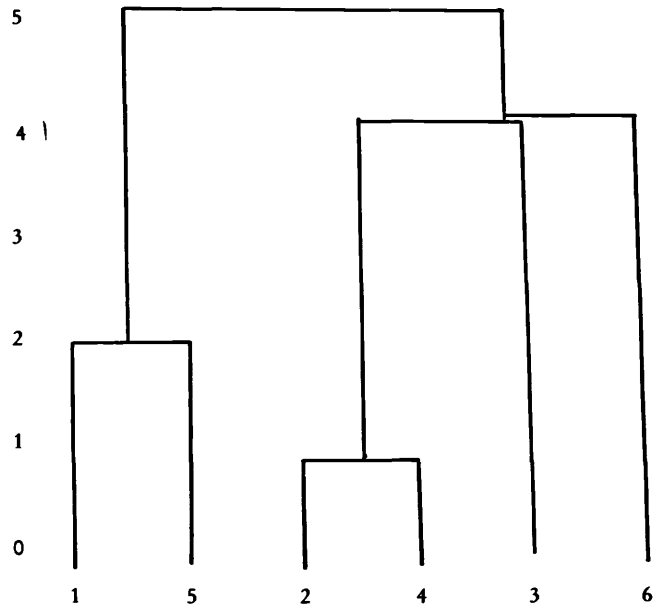
$$d_{(243)(15)} = \min\{d_{21}, d_{25}, d_{41}, d_{45}, d_{31}, d_{35}\} = d_{21} = 5$$

$$d_{(243)6} = \min\{d_{26}, d_{46}, d_{36}\} = d_{26} = 4$$

Con estos datos volvemos a elaborar la matriz

	(1,5)	(2,4,3)	6
(1,5)	—	5	5
(2,4,3)		—	4
6			—

Los dos más próximos son el grupo (2,4,3) y el individuo 6. Por lo tanto unimos ambos clusters. Finalmente se unen los dos grupos (2,4,3,6) y (1,5) para formar un solo grupo.



En los métodos jerárquicos no hay una respuesta fácil a la hora de decidir con cuántos grupos quedarnos tras el análisis. La solución va desde confiar la decisión al conocimiento sustantivo que tenga el investigador del objeto de estudio hasta observar el dendograma y quedarse con el número de grupos que queda tras grandes cambios en las fusiones de los grupos. Sin embargo, en muchos casos no será necesario decidir el número de grupos subyacentes a los datos en la medida en que sólo se busque la estructura jerárquica de clasificación de los objetos. En nuestro ejemplo lo mismo podría hablarse de dos conglomerados, el (1,5) y el (2,4,3,6), que de tres, el (1,5), el (2,4) y el (3,6).

7.3.2. *Método de las distancias máximas*

Se diferencia del anterior en que recoge la distancia máxima entre grupos. Veamos el arranque del método sobre la misma matriz de datos del ejemplo anterior.

Como en el caso anterior, los dos individuos más próximos son el 2 y el 4. Calculemos las distancias del grupo 2,4 con las demás:

2,4 con 1 — 2 con 1 = 9; 4 con 1 = 8 } máxima = 9

2,4 con 3: máxima = 7
2,4 con 5: máxima = 6
2,4 con 6: máxima = 5

La nueva matriz será:

	1	2,4	3	5	6
1	—	9	9	2	9
2,4	—	—	7	6	5
3	—	—	—	6	8
5	—	—	—	—	3

Los más próximos ahora son el 5 y el 6, luego, hallamos el cálculo de las distancias entre el 5 y 6 y los restantes clusters o individuos; y así sucesivamente.

7.3.3. *Método de las distancias entre centroides*

También llamado en inglés *Average Linkage*. En este método la distancia que se computa es la que existe entre los centroides de los grupos. Y se unirán aquellos grupos que tengan sus centroides más próximos.

En el método del centroide utilizamos las siguentes fórmulas para calcular la distancia.

$$D_{i+j,k} = \frac{n_i d_{ki} + n_j d_{kj} - \frac{n_i n_j d_{ij}}{n_i + n_j}}{n_i + n_j}$$

que representa la distancia entre el grupo formado por los individuos o grupos $i + j$, y el individuo o grupo k, siendo n_i el número de individuos del grupo i y d_{ki} la similaridad entre k e i.

O también:

$$D_{i+j,l+m} = \frac{n_i n_l d_{il} + n_j n_l d_{jl} + n_i n_m d_{im}}{(n_i + n_j)(n_l + n_m)} - \frac{n_i n_j d_{ij}}{(n_i + n_j)^2} - \frac{n_l n_m d_{lm}}{(n_l + n_m)^2}$$

En el ejemplo anterior sería:

Individuos

Individuos	1	2	3	4	5	6
1	—	9	9	8	2	9
2	—	—	4	1	6	4
3	—	—	—	7	6	8
4	—	—	—	—	5	5
5	—	—	—	—	—	3

Como siempre, elegimos el par más similar, el 2 y el 4, y se calculan las distancias entre este par y el resto.

	1	2 + 4	3	5	6
1	—	8,75	9	2	9
2 + 4	—	—	5,75	5,75	4,75
3	—	—	—	6	8
5	—	—	—	—	3

$$D_{(2+4),1} = \frac{1 \times 9 + 1 \times 8 - \frac{1 \times 1 \times 1}{2}}{2} = \frac{9}{2} + \frac{8}{2} - \frac{1}{4} =$$

$$\frac{18 + 16 + 1}{4} = \frac{35}{4} = 8{,}75$$

$$D_{(2+4),3} = \frac{1 \times 4 + 1 \times 7 - \frac{1 \times 1 \times 1}{2}}{2} = \frac{8}{4} + \frac{14}{4} - \frac{1}{4} = 5{,}75$$

$$D_{(2+4),1} = \frac{1 \times 6 + 1 \times 5 - \frac{1 \times 1 \times 1}{2}}{2} = \frac{12}{4} + \frac{10}{4} - \frac{1}{4} = 5{,}75$$

$$D_{(2+4),6} = \frac{1 \times 4 + 1 \times 5 - \frac{1 \times 1 \times 1}{2}}{2} = \frac{8}{4} + \frac{10}{4} - \frac{1}{4} = 4{,}75$$

Ahora los más próximos son el 1 y el 5, que pasan a formar un grupo independiente. Así pues, tenemos que calcular ahora las distancias entre:

1,5 y 2,4
1,5 y 3
1,4 y 6

Aplicando la fórmula,

$$D_{(1+5)(2+4)} = \frac{1 \times 1 \times d_{12} + 1 \times 1 \times d_{52} + 1 \times 1 \times d_{14}}{(1 + 1)(1 + 1)} - \frac{1 \times 1 \times d_{15}}{(1 + 1)^2} -$$

$$- \frac{1 \times 1 \times d_{24}}{(1 + 1)^2} = \frac{1 \times 1 \times 9 + 1 \times 1 \times 6 + 1 \times 1 \times 8}{4} -$$

$$- \frac{1 \times 1 \times 2}{4} - \frac{1 \times 1 \times 1}{4} = \frac{23 - 2 - 1}{4} = 5$$

$$D_{1+5,3} = \frac{1 \times d_{13} + 1 \times d_{53} - \frac{1 \times 1 \times d_{15}}{1 + 1}}{1 + 1} =$$

$$= \frac{1 \times 9 + 1 \times 6 - \frac{1 \times 1 \times 2}{2}}{2} = 7$$

$$D_{1+5,6} = \frac{1 \times d_{16} + 1 \times d_{56} - \frac{1 \times 1 \times d_{15}}{1 + 1}}{1 + 1} =$$

$$= \frac{1 \times 9 + 1 \times 3 - \frac{1 \times 1 \times 2}{2}}{2} = 5,5$$

Así que la nueva matriz sería:

	2 + 4	1 + 5	3	6
2 + 4	—	5	5,75	4,75
1 + 5	—	—	7	5,5
3	—	—	—	8

Los grupos más próximos ahora son el 2 + 4 y el 6. Luego tenemos que calcular la distancia del nuevo grupo, 2 + 4 + 6, con todos los demás

2 + 4 + 6 con 1 + 5
2 + 4 + 6 con 3

$$D_{(\substack{2\\i}+\substack{4\\j}+\substack{6\\k})(\substack{1\\i}+\substack{5\\j})} = \frac{2 \times d_{2+4/1+53} + 1 \times d_{1+5/6} - \frac{2 \times 1 \times d_{2+4/6}}{2 + 1}}{2 + 1} =$$

$$= \frac{2 \times 5 + 1 \times 5,5 - \frac{2 \times 1 \times 4,75}{3}}{3} = \frac{30 + 16,5 - 9,5}{9} = 4,1$$

$$D_{(\substack{2\\i}+\substack{4\\j}+\substack{6\\k})(\substack{3\\i})} = \frac{2d_{3/2+4} + 1d_{6/3} - \frac{2 \times 1 \times d_{2+4/6}}{2 + 1}}{2 + 1} =$$

$$= \frac{2 \times 5,75 + 1 \times 8 - \frac{2 \times 1 \times 4,75}{3}}{3} = \frac{34,5 + 24 - 9,5}{9} = 5,4$$

La nueva matriz sería:

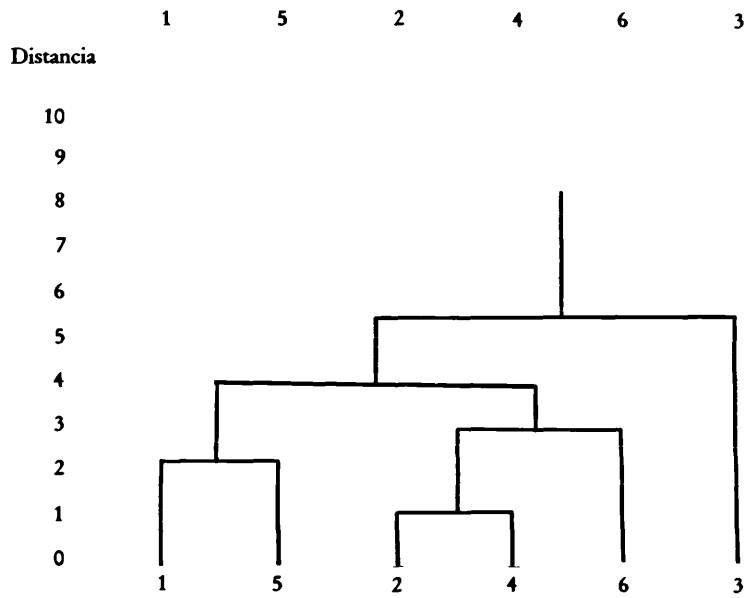
	2 + 4 + 6	1 + 5	3
2 + 4 + 6	—	4,1	5,4
1 + 5	—	—	7

Los grupos más próximos ahora son el 1,5 con el 2,4,6; luego debemos calcular la distancia entre el grupo 2 + 4 + 6 + 1 + 5 y el 3

$$D_{\substack{2+4+6+1+5/3 \\ i \quad j \quad k}} = \frac{3 \times d_{3/2+4+6} + 2 \times d_{3/1+5} - \dfrac{3 \times 2 \times d_{2+4+6/1+5}}{3+2}}{3+2} =$$

$$= \frac{3 \times 5{,}4 + 2 \times 7 - \dfrac{3 \times 2 \times 4{,}1}{5}}{5} = \frac{81 + 70 - 24{,}6}{25} = 5{,}056$$

Podemos crear una estructura arborescente, mostrando cómo en cada paso se van uniendo los individuos.



7.3.4. *Método de las distancias ponderadas (Weighted Pairgroup)*

Se parte de la matriz de coeficiente de similaridad, ya configurada, y se establece el par, de todos los posibles, que tiene la más alta similaridad (coeficiente de correlación producto momento).

A partir de este par, ya establecido, se van agrupando el resto de individuos a dicho par de acuerdo a unas condiciones de aglomeración.

Se establece una nueva matriz de similaridad formada por los mismos elementos de la anterior y sustituyendo las columnas y las filas de los dos individuos agrupados por una sola columna y fila.

El resto de individuos de la matriz puede agruparse con este primer núcleo, para lo cual se observa, en primer lugar, los individuos que tienen una más alta correlación media con los miembros del grupo.

El límite de la formación del grupo se da cuando decrece el nivel de la correlación media de los miembros del grupo hasta un nivel dado.

Veamos un ejemplo:

Supongamos que tenemos la matriz de similaridad siguiente (los datos de similaridad son coeficiente de correlación):

	1	2	3	4	5	6
1		0,54	0,81	0,90	0,65	0,45
2	—	—	0,76	0,90	0,96	0,67
3	—	—	—	0,88	0,43	0,35
4	—	—	—	—	0,95	0,43
5	—	—	—	—	—	0,63
6	—	—	—	—	—	—

Elegimos el coeficiente más alto, que corresponde al par de individuos más similar, en nuestro caso, el par (2,5). Con un coeficiente de 0,96. Así pues, este es el núcleo de partida. Veamos la nueva correlación de todos los individuos con el par elegido.

Caso de 1 con el par (2,5)

1 con 2 tiene	0,54
1 con 5 tiene	0,65
	1,19

Correlación media entre 1 y el par (2,5) es igual a $\frac{1 \cdot 19}{2} = 0,59$

Caso de 3 con el par (2,5)

3 con 2 tiene	0,76
3 con 5 tiene	0,43
	1,19

Correlación media = 0,59

Caso de 4 con el par (2,5)

4 con 2 tiene	0,90
4 con 5 tiene	0,95
	1,85

Correlación media = 0,92

Caso de 6 con el par (2,5)

6 con 2 tiene	0,67
6 con 5 tiene	0,63
	1,30

Correlación media = 0,65

Así pues, ahora podemos formar la nueva matriz, uniendo en una sola columna y una sola fila los 2 individuos más similares:

	1	(2,5)	3	4	6
1	—	0,59	0,81	0,90	0,45
(2,5)	—	—	0,59	0,92	0,65
3	—	—	—	0,88	0,35
4	—	—	—	—	0,43
6	—	—	—	—	—

El individuo que tiene una correlación media más alta con el núcleo formado es el 4. ¿Se incorpora el individuo 4 al grupo (2,5)? Sí, siempre que no baje la correlación media del grupo por debajo de un valor dado que es elegido por el investigador, en nuestro caso puede ser el 0,03; veamos la correlación media del grupo creado.

	Corr.
2 con 5	0,96
2 con 4	0,90
4 con 5	0,95
	2,81

Correlación media del grupo = 0,933

Si la correlación del grupo (2,5) es de 0,96 y ahora es de 0,933, ha bajado un 0,027 que es menor que el 0,030 permitido, luego el individuo 4 queda integrado al grupo.

Se calculan las distancias.

1 con (2,5,4)

	2	-	0,54
1	5	-	0,65
	4	-	0,90
			2,09

$$\text{Distancia media} = 0{,}696 = \frac{2{,}09}{3}$$

3 con (2,5,4)

	2	-	0,76
3	5	-	0,43
	4	-	0,88
			2,07

$$\text{Distancia media} = \frac{2{,}07}{3} = 0{,}696$$

6 con (2,5,4)

	2	-	0,67
6	5	-	0,63
	4	-	0,43
			1,73

$$\text{Distancia media} = \frac{2{,}09}{3} = 0{,}58$$

Volvemos a formar la nueva matriz:

	1	(2,5,4)	3	6
1	—	0,696	0,81	0,45
(2,5,4)	—	—	0,696	0,58
3	—	—	—	0,35

El coeficiente de correlación más alto ahora es el 1 con el 3 (0,81). Calculamos los nuevos índices de similaridad de la pareja (1,3) con 6 y con (2,5,4).

(1,3) con 6

	1	-	0,45
6	3	-	0,35
			0,80

$$\text{Distancia media} = 0{,}40$$

Calculamos ahora:

(2,5,4) con (1,3)

	2	-	0,54	
1	5	-	0,65	Suma: 2,09;
	4	-	0,90	
	2	-	0,76	
3	5	-	0,43	Suma: 2,07
	4	-	0,88	4,16

$$\text{Distancia media: } \frac{4,16}{6} = 0,69$$

Luego:

	(2,5,4)	(1,3)	6
(2,5,4)	—	0,69	0,58
(1,3)	—	—	0,40
6	—	—	—

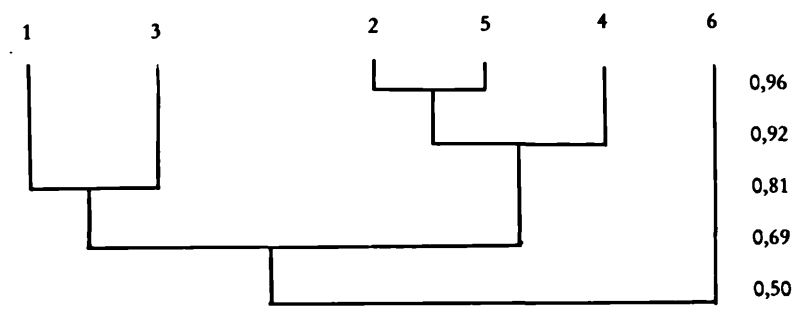
Elijo 0,69, luego, uno (2,5,4) con (1,3).
Ahora tenemos que calcular la relación de (1,2,3,4,5,) con 6

G { 1 - 0,45; 2 - 0,67; 3 - 0,35; 4 - 0,43; 5 - 0,63 } Suma: 2,53

$$\text{distancia media: } \frac{2,53}{5} = 0,506$$

	(1,2,3,4,5,)	6
(1,2,3,4,5)	—	0,506
6	—	—

Ahora podemos establecer una estructura arborescente que nos indique las uniones sucesivas, siendo el coeficiente de correlación los valores en ordenadas. Un valor en abscisas no significa nada.



7.3.5. *Método de William y Lambert*

Este método divide a la población en grupos, en lugar de lo que hacen otros que forman los grupos mediante agrupamiento de individuos. El artículo original donde se explica el método trata de una división de cuadrantes de terreno según tengan o no cinco especies distintas de plantas (A. C. 1975).

Los datos de entrada son, pues, la presencia o ausencia de las especies en los cuadrantes, que toman valores (1,0) respectivamente. A nuestros efectos podríamos hablar de individuos y de características, siendo de interés dividir los individuos.

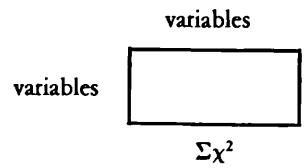
Se establece una primera matriz que correlaciona las variables dos a dos, todas con todas.

La correlación entre una variable y otra será un índice de asociación que definimos como el valor χ^2 que se obtiene de una tabla de contingencia 2 × 2. Por ejemplo:

	Variable A	Variable B
N.º de individuos que tienen la característica	SI	SI
N.º de individuos que no tienen la característica	NO	NO
Total de individuos		

A partir de una tabla de este tipo para todas las combinaciones posibles de variables dos a dos podemos obtener el valor de χ^2 que es el que hemos llamado índice de asociación.

Así, podemos construir la matriz de asociaciones de todas las variables consigo mismas.



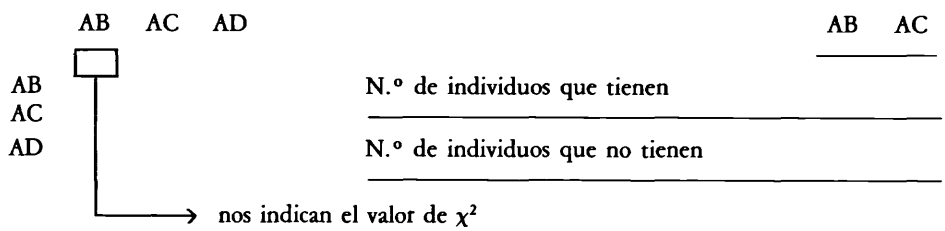
el valor de las casillas de esta matriz es el valor de χ^2 obtenido mediante la anterior tabla de equivalencia.

Esta matriz es simétrica, y además no existen valores en la diagonal que nos indicarían las asociaciones de las variables consigo mismas. Los valores no significativos al computar la χ^2 son tratados como ceros en la matriz.

Obtenemos la suma de los χ^2 en todas las columnas y establecemos que la primera segmentación se efectuará por aquella variable cuya χ^2 sea más alta.

Para efectuar la segunda segmentación se establece la siguiente segunda matriz, donde aparecen las variables asociadas dos a dos.

Población según tenga o no cada característica	(Distintos universos) Número de individuos	Asociaciones binarias									
		AB	AC	AD	AE	BC	BD	BE	CD	CE	DE
Tiene A											
No tiene											
Tiene B											
No tiene											
Tiene C											
No tiene											
Tiene D											
No tiene											
Tiene E											
No tiene											



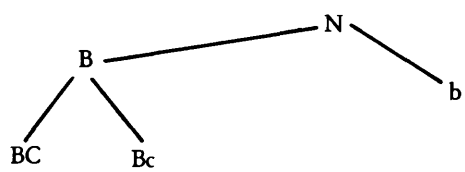
En cada subpoblación se calculan los valores de χ^2 para la asociación de variables que indican las columnas.

Lógicamente no se calcularán estos valores de χ^2 en las poblaciones donde estén presentes algunas de las variables; los valores de χ^2 no significativos son tomados como 0.

Habíamos dicho que la primera división se efectuaba por la χ^2 más alta en la primera matriz, que, por ejemplo, puede corresponder a la variable B.

Pues bien, observamos los valores de las asociaciones entre variables en esta fila donde puede dividirse la población según tenga o no la variable B y observamos el valor de χ^2 más alto en la fila que, por ejemplo, es la asociación AC. ¿Por cuál de los dos segmentamos a la población? Por aquel que en la primera matriz tenga un valor χ^2 más alto que, por ejemplo, en nuestro caso es la C.

Así:



Bc será el número de individuos entre los que tienen B que además tengan C.

Ejemplo de aplicación del método de Willian y Lambert, tomado de su artículo.

	Variables				
	A	B	C	D	E
A	—	51,31	45,66	x	x
B	51,31	—	93,76	(23,63)	(68,64)
C	45,66	93,76	—	(4,84)	(14,08)
D	x	(12,62)	(4,84)	—	6,92
E	x	(68,64)	(14,08)	6,92	—
$\Sigma\chi^2$	96,97	226,33	158,34	24,38	89,64

x - significa que los valores de χ^2 no son significativos y serán tratados como 0.
0 significa que la asociación es indeterminada en esa población.
() significa que el tipo de asociación es negativo.
— significa que la asociación es necesariamente indeterminada en función de las características.

La primera segmentación se efectuará por B ($\Sigma\chi^2$ más alto).

Hay 615 individuos.

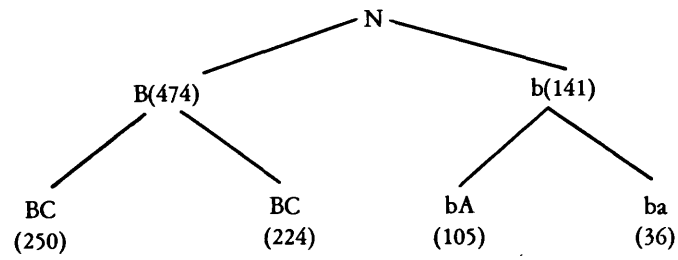
Se establecen los valores de χ^2 en los grupos de individuos que señala el cuadro siguiente, sin calcular las variables en las que se han dividido los individuos.

COMBINACIONES DE DOS EN DOS

Individuos	N.º de casos	AB	AC	AD	AE	BC	BD	BE	CD	CE	DE
Con variable A	556	—	—	—	—	73,29	(19,34)	(65,41)	(7,18)	(12,50)	9,25
Sin variable A	59	—	—	—	—	0	0	x	0	0	0
Con variable B	474	—	24,8	x	0	—	—	—	x	0	0
Sin variable B	141	—	x	4,34	x	—	—	—	x	x	x
Con variable C	259	0	—	0	0	—	x	0	—	—	0
Sin variable C	356	16,16	—	x	x	—	(25,63)	(35,06)	—	—	3,85
Con variable D	29	0	0	—	0	5,63	—	x	—	x	—
Sin variable D	586	58,12	47,91	—	(5,20)	83,08	—	(59,22)	—	(11,55)	—
Con variable E	21	0	0	—	—	0	0	—	0	—	—
Sin variable E	594	48,50	43,99	x	—	77,88	(7,69)	—	x	—	—

La primera segmentación es en B, luego observamos la fila B. El valor más alto de χ^2 en la fila B es AC. ¿Por cuál segmentamos, por *A* o por *C*. Vamos a la primera matriz y vemos que C tiene un χ^2 más alto, luego seleccionamos C.

En la fila b (sin variable B) observamos que AD tiene el valor más alto. A tiene un χ^2 más alto que D, luego segmentamos por A.



7.3.6. *El método K-means (Dixon, 1981)*

Este método arranca ya de una configuración de los grupos y no como en los métodos anteriores que arrancaba de cada uno de los individuos como clusters separados. Esta primera configuración de grupos puede ser al azar, ya que el objetivo del método es ir mejorándola paso a paso. En cada paso se crea una nueva configuración o distribución de casos con el mismo número de grupos, de manera que se va reduciendo la distancia media al cuadrado desde todos los componentes de cada grupo a su centroide. Si en un paso esta distancia media disminuye, el método continúa; y si aumenta se para el proceso.

En este método es necesario elegir previamente el número de clusters que queremos conseguir. Sobre ese número, sin modificarlo, se optimiza la distribución de casos. Para solucionar este inconveniente será posible realizar varios procesos con distinto número de clusters de partida.

Es posible utilizar también cualquier otro método de agrupamiento de los expuestos en estos papeles, y cuando los clusters están ya definidos, reordenar los casos de cada cluster según este método.

Esta técnica es útil cuando en vez de estar interesados por la estructura jerárquica de la clasificación de los individuos (objetos) tan solo tenemos interés en conocer el número de grupos constituidos y sus características.

7.3.7. *Q-technique*

En el artículo «A Statistical Method for Evaluating Systematic Relationships» de Robert R. Sokal y Charles D. Michener (1958), se expone el método Q-technique. Este método consiste en establecer correlaciones entre individuos basadas sobre medidas de las características que tienen en común. Por ejemplo, en psicología puede aplicarse en la interpretación de tests realizados a varios sujetos. En el fondo la Q-technique es un método para determinar dimensiones, y puede decirse que es un análisis factorial. El objetivo de este método es conseguir tipos puros, aquellos marcados por las dimensiones que haya producido el análisis. Esta técnica ha sido criticada, porque el análisis factorial no está especialmente diseñado para clasificar, sino para buscar las dimen-

siones subyacentes de un fenómeno. Según la Q-technique los individuos se clasifican en función de variables ortogonales que son las que proporcionan el marco de referencia del espacio, de ahí la calificación de puros a los individuos que quedan clasificados en los distintos grupos por dichas dimensiones. Este método ha sido criticado por Cattell (1978). La primera objeción es que en dicha técnica no se tiene en cuenta la importancia de cada factor, sustituyéndose por un factor común a todos los individuos que van a ser clasificados. La segunda objeción es que no es razonable hablar de la estructura simple en la factorización que se hace en esta técnica. La tercera objeción es que la factorización se hace sobre unos pocos sujetos, generalizándose después al resto. La siguiente objeción es que si bien los tests en psicología son relativamente permanentes, no ocurre lo mismo con las personas.

7.4. El dendograma

Por cualquiera de los procedimientos explicados anteriormente (a excepción del K-means), conseguimos formar un árbol en el que podemos ver con toda claridad cómo se van agrupando los sujetos.

Este árbol se llama en la terminología inglesa dendograma.

En el dendograma la abcisa no tiene realmente significado; la ordenada representa el nivel o los distintos niveles de similaridad donde se han ido agrupando los sujetos, de acuerdo a la medida que hayamos elegido.

De Sneath y Sokal tomamos la idea de Phenon, que no es otra cosa que un indicador de posición de cada unión en el dendograma. El indicador Phenon es interesante desde dos puntos de vista:

1. Con él es posible comparar dendogramas, de forma que podemos establecer la agrupación utilizando diversos métodos y posteriormente podemos comparar y medir estos resultados.
2. Sirve para establecer una medida de ajuste entre los datos de partida y la estructura del dendograma.

Para calcular el coeficiente *phenon* se dividen los rasgos de similaridad en intervalos iguales, tantos como sean necesarios, procurando que no se agrupen muchas uniones en cada intervalo y estableciendo los mismos intervalos en los distintos dendogramas que queramos comparar. Cada intervalo es codificado con un número sobre una escala que va de 1 en adelante, siendo el valor 1 el intervalo que recoge la última agrupación, es decir, aquella pareja que se forma en último lugar.

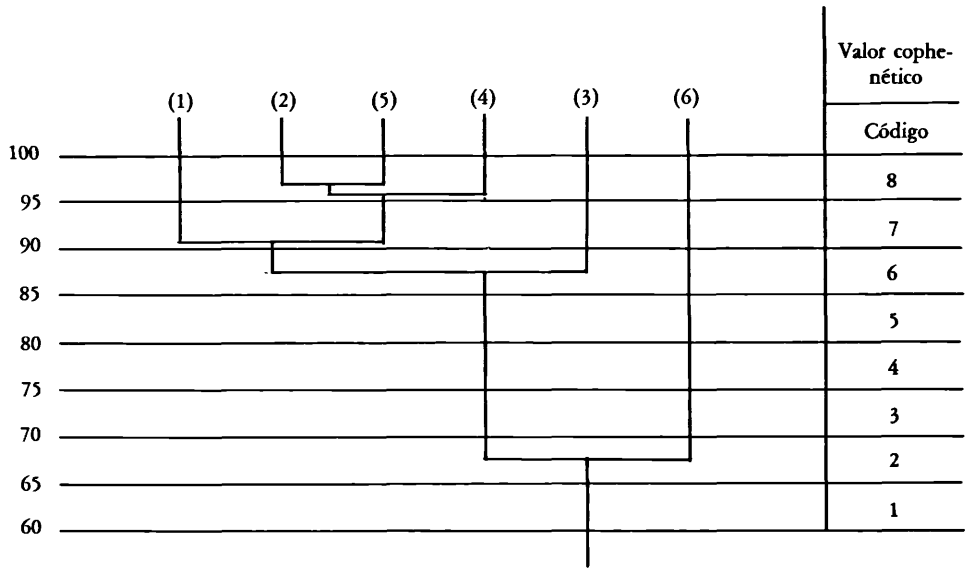
El valor «cophenético» de una agrupación puede, pues, medirse en el código donde se encuentran.

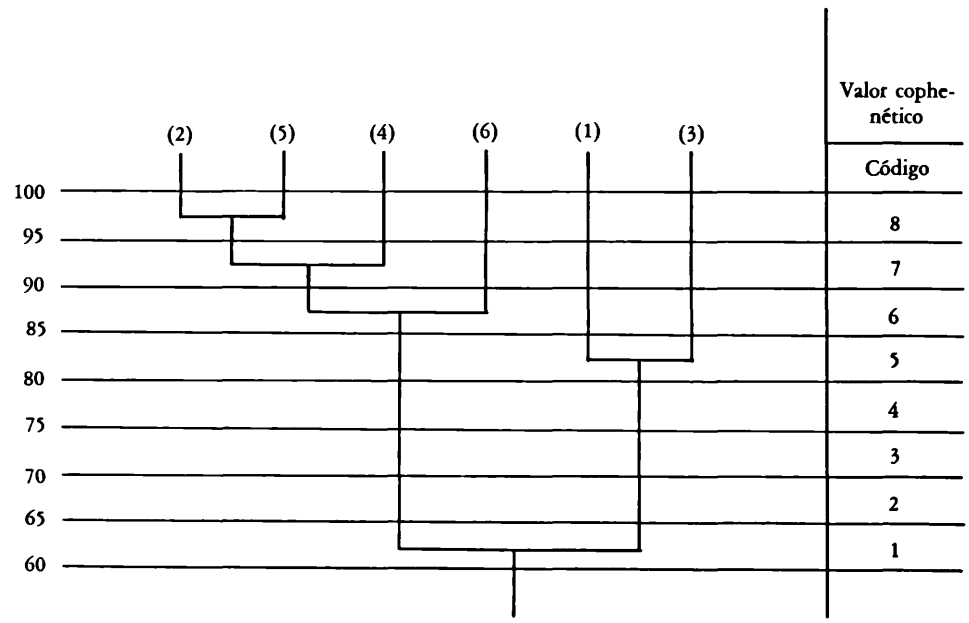
Si, por ejemplo, el individuo 3 y el 5 se agrupan en el *phenon* 5, su «valor cophenético» es 5.

Posteriormente, se llevan estos valores a una matriz, en la que tanto en filas como en columnas están los sujetos que se han agrupado y en los puntos de intersección ponemos el «valor cophenético» correspondiente. De esta manera, si tenemos dos dendogramas y por lo tanto dos matrices de valores cophenéticos es posible establecer

entre ellas un coeficiente que nos indique la correlación que exista. Este coeficiente se llama el «coeficiente de correlación cophenético».

Veamos un ejemplo. Para ello expongamos a escala, y con los valores cophenéticos correspondientes, dos dendogramas.





Del primer dendograma obtenemos la siguiente matriz:

	1	2	3	4	5	6
1		7	6	7	7	2
2			6	8	8	2
3				6	6	2
4					8	2
5						2

Del segundo dendograma obtenemos la siguiente matriz:

	1	2	3	4	5	6
1		1	5	1	1	1
2			1	7	8	6
3				1	1	1
4					7	6
5						6

Valor cophenético	En el primer dendograma hay	En el segundo dendograma hay
8	3 asociaciones	1 asociación
7	3 asociaciones	2 asociaciones
6	4 asociaciones	3 asociaciones
5	0 asociaciones	1 asociación
4	0 asociaciones	0 asociaciones
3	0 asociaciones	0 asociaciones
2	5 asociaciones	0 asociaciones
1	0 asociaciones	8 asociaciones
	15	15

Establecemos ahora la matriz que tenga como filas y columnas los valores cophenéticos del primer y segundo dendogramas.

									Primer dendograma
	1	2	3	4	5	6	7	8	
8							2	1	3
7	3								3
6	3				1				4
5									0
4									0
3									0
2	2								5
1									0
Segundo dendograma	8	0	0	0	1	3	2	1	

Entre estas dos distribuciones puede establecerse por ejemplo el coeficiente de correlación de Pearson.

7.5. Aplicación del análisis de cluster a una encuesta de actitudes políticas

Veamos una aplicación del análisis de cluster. Este análisis es muy útil para resolver muchas situaciones o investigaciones de la sociología y el marketing. Hemos elegido una que nos parece muy adecuada.

Se trata de conocer la composición interna del electorado de un partido político. No ya las variables sociodemográficas o psicográficas que caracterizan a sus electores potenciales, sino el peso de las tendencias políticas dentro del partido. Algunos partidos políticos son votados por una gran masa de individuos, heterogénea en cuanto a sus ideas y creencias políticas. Se trata entonces de conocer el peso y la importancia de esas distintas tendencias para actuar en consecuencia. De gran interés para el partido político será observar y medir la evolución de esas tendencias a lo largo del tiempo. Suponiendo que estamos ante el electorado potencial de un gran partido de la derecha, se trataría de conocer qué importancia tiene la tendencia política de centro, la derecha tradicional y católica, el componente liberal..., etc.

Para lo cual se ha realizado una encuesta sobre actitudes políticas. No se incluyen las preguntas en este artículo debido a la confidencialidad del tema. En cualquier caso las preguntas deberán ser suficientes y relativas a todos aquellos ítems que pueden constituir una actitud política. Cuanto mayor sea el número de variables, mejor será para definir los clusters. En nuestro caso se han introducido 40 variables.

El programa utilizado ha sido el K-means del paquete BMDP, que ha sido explicado anteriormente.

Como hemos dicho al explicar este programa, hay que definir previamente un número de grupos. El análisis lo que hace es reordenar los individuos en función de las variables seleccionadas, de manera que al final del proceso los individuos pertenecientes a un grupo sean lo más similares posible. Naturalmente, los grupos finales, compuestos cada uno de ellos por individuos extremadamente parecidos en cuanto a una

serie de actitudes políticas, nos descubrirán las distintas tendencias dentro del electorado potencial del partido, su importancia, su fuerza y a través de un análisis longitudinal, su evolución.

Para solucionar uno de los problemas de este programa en el que necesariamente hay que definir a priori el número concreto de grupos que se desean formar, es recomendable preparar el programa para que repita el proceso de reordenación con distinto número de grupos.

En nuestro caso, como no sabemos con certeza cuántas tendencias diferenciadas existen en el interior del electorado del partido, decimos al análisis que procese los datos con 6, 5 y 4 grupos de partida.

En primer lugar el programa procesa los datos con 6 clusters. Realiza 40 reordenaciones de sujetos entre los grupos. Como ya hemos dicho, la primera distribución de los individuos en grupos es al azar o, en otro caso, es el proceso final de un análisis cluster de sujetos al estilo tradicional.

Todos los individuos que han sido tratados por el análisis, tienen intención de votar a un partido concreto de la derecha.

Sería posible, aunque no lo hemos hecho en este caso, realizar previamente un análisis discriminante para aislar y sólo trabajar con aquellos electores del partido especialmente puros, cercanos al centroide; de manera que los grupos constituidos al final reflejen con más claridad la realidad, excluyendo en definitiva aquellos electores potenciales del partido no claros, ambiguos, con posibilidad de ser perdidos, muy influenciables a campañas políticas o con fuerte tendencia a la indecisión o a la abstención. El número de datos que ha tratado el análisis es de 1.500 electores potenciales de un determinado partido.

Este programa trabaja con la distancia euclideana. La primera matriz de datos se configura con la distancia euclideana de cada individuo a todos los demás. Un individuo cualquiera pertenecerá al grupo o cluster cuya distancia a su centroide desde el individuo sea más pequeña. La *K* que incluye el nombre del programa se refiere al número de grupos que se desee formar. En nuestro caso hemos realizado el análisis con distintos valores de $k = 6, 5$ y 4.

Después de las 40 reordenaciones, el análisis nos presenta los siguientes resultados y nos proporciona la siguiente información para cada cluster logrado.

Por ejemplo, para el cluster n.° 1 nos da el número de casos que contiene, dato éste que nos va a indicar la importancia de la tendencia que representa el cluster 1 en el electorado potencial del partido.

En este sentido, el programa nos presenta dos gráficos.

1. Muestra la distancia desde el centroide del cluster a cada uno de los individuos.

Sobre la línea de puntos es posible observar la distancia desde cualquier individuo del cluster n.° 1 al centroide de dicho cluster. Naturalmente, cuanto más cerca del centroide están los componentes del grupo, más homogéneo será éste; en definitiva, más precisa y nítida es la tendencia política que en él se encierra.

Obsérvese en el gráfico n.° 1, que el dígito que aparece, el 1, nos indica precisamente el grupo al que corresponde; cada dígito es un individuo. En horizontal vemos claramente las distancias desde cualquier individuo al centro del grupo. En vertical podemos ver el número de individuos en cada intervalo de distancia. Como se ve se

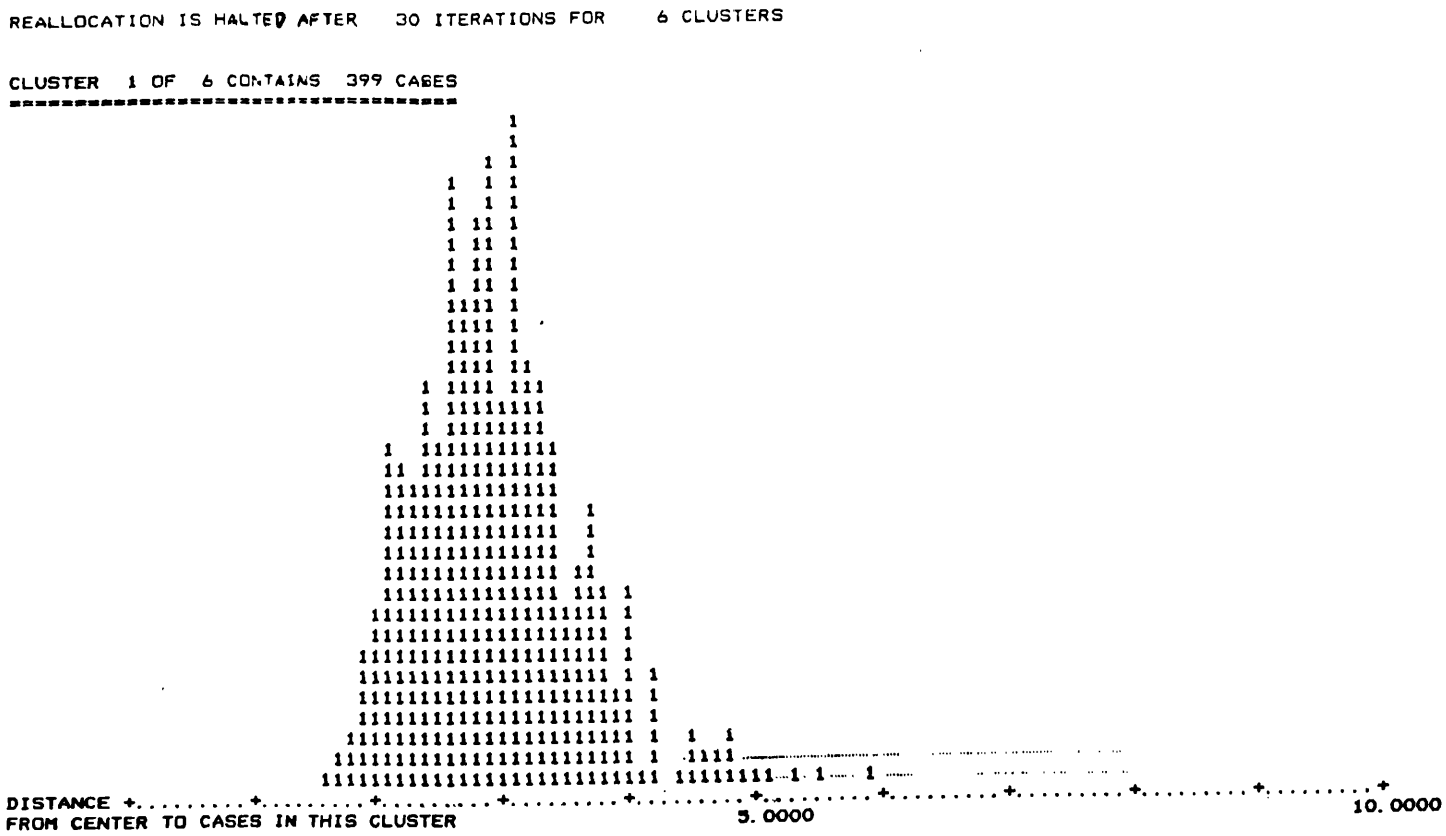


FIGURA 1. Distancia del centroide del cluster 1 a cada uno de los individuos de ese cluster. El n.º de individuos es de 399.

crea una distribución de frecuencias en la que pocos individuos están muy próximos al centroide del grupo, y, también pocos individuos están muy lejanos al centroide del grupo; pero, en cualquier caso, por muy lejanos que estén, están más próximos al centroide de este grupo que a otro cualquiera. La forma de esta distribución de frecuencias nos indica el grado de homogeneidad de esta tendencia política del electorado potencial del partido.

2. En segundo lugar, podemos analizar otro gráfico que nos presenta, de la misma manera que el anterior, la distancia desde el centroide del cluster n.° 1 al resto de los individuos de otros clusters.

Naturalmente, no sólo es necesario que los individuos pertenecientes a un cluster estén próximos a su centroide, sino que también será necesario que el resto de individuos pertenecientes a otros clusters estén lo más lejos posible de dicho centroide.

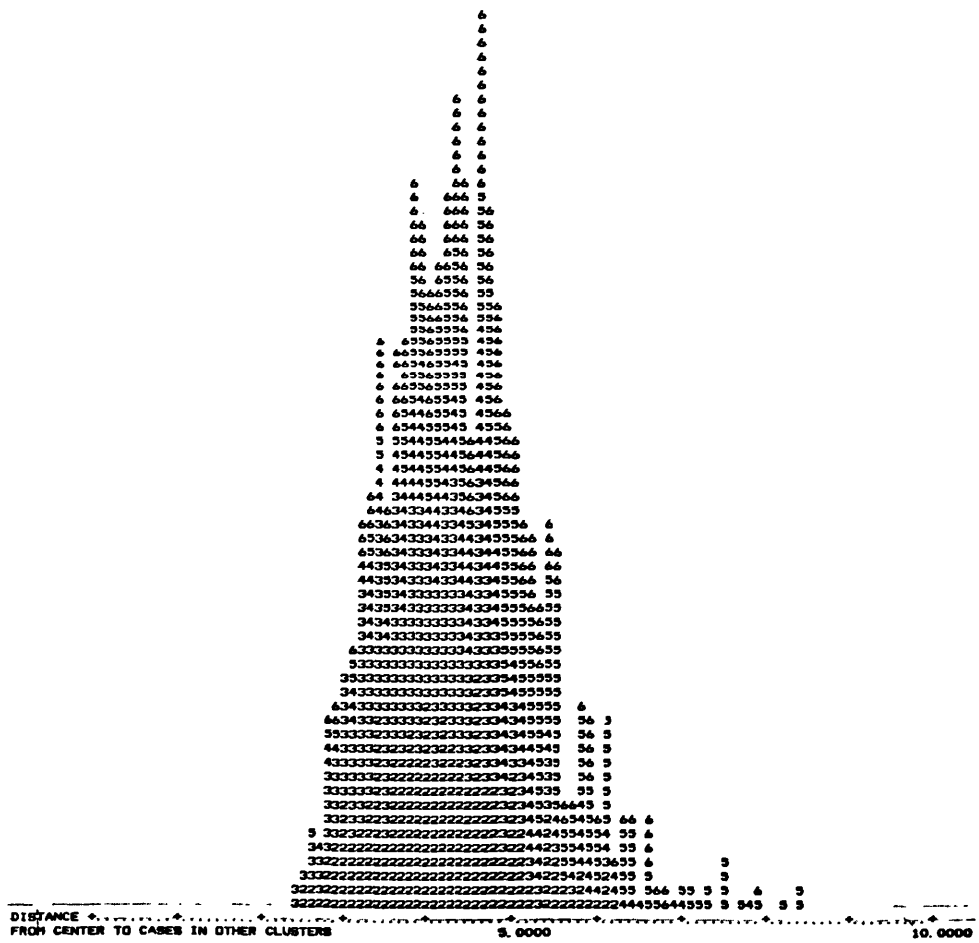


FIGURA 2. Distancias de los individuos de los otros clusters al centroide del cluster 1.

La visión conjunta de las dos distribuciones de frecuencias nos indicará la nitidez y fuerza de la tendencia que se esconde detrás del cluster n.º 1. Vemos que la mayor parte de los componentes del grupo 1 se solapan con los componentes de otros grupos. Este hecho disminuye la fuerza de la corriente sociopolítica que está detrás del cluster n.º 1, ya que vemos que hay individuos que están más próximos al centroide del grupo n.º 1 que muchos de los componentes de este grupo, estando a la vez mucho más próximos, todavía, a los centroides de otros grupos; también podemos ver gráficamente qué grupo es el más distante del n.º 1. En el ejemplo este grupo es el n.º 5.

Más tarde, el programa nos proporciona los mismos datos ya expuestos, pero para cada individuo. En la tabla siguiente sólo expondremos los datos correspondientes a algunos individuos, aunque la salida de datos del programa nos proporcionará los datos de todos los individuos.

Case	Weight	Distance
0016	1,0000	2,7479
0017	1,0000	3,7311
0032	1,0000	2,9086
0037	1,0000	3,9930
0038	1,0000	2,8443
003	1,0000	2,8031
0100	1,0000	2,0911
0108	1,0000	3,9637
0132	1,0000	2,2599
0153	1,0000	2,0805
0154	1,0000	3,7910
0162	1,0000	4,5282
0168	1,0000	2,9001
0170	1,0000	2,7247
0172	1,0000	3,2667
0182	1,0000	2,6688
01 3	1,0000	3,2517
0235	1,0000	2,1285
0247	1,0000	2,6049
0260	1,0000	3,4942

TABLA 1. Distancia de algunos de los individuos al centroide del cluster 1.

En la tabla 1 la primera columna nos indica el número correspondiente a cada individuo. La siguiente columna nos muestra el peso de dicho individuo. Es la misma información del gráfico de la figura n.º 1, pero con la distancia, en concreto, de cada individuo a su centroide.

Otra información de gran importancia es el comportamiento de las variables introducidas en el grupo. Veamos el siguiente cuadro.

Variable	Minimum	Center	Maximum	St. dev.
2 P5	1,0000	2,5670	6,0000	0,7408
3 LIDER 1	1,0000	4,3655	5,0000	0,5912
4 LIDER 2	2,0000	4,1877	5,0000	0,5667
5 LIDER 3	1,0000	1,7445	5,0000	0,6952
6 LIDER 4	3,0000	4,6152	5,0000	0,5443
7 LIDER 5	3,0000	4,4054	5,0000	0,5617
8 UCD	2,0000	4,2269	5,0000	0,5923
9 PSOE	1,0000	1,7132	5,0000	0,6750
10 PCE	2,0000	4,2376	5,0000	0,5877
11 AP	3,0000	4,5423	5,0000	0,5327
12 CDS	2,0000	4,1403	5,0000	0,6342
13 ITEM 1	1,0000	1,2816	2,0000	0,4504
14 ITEM 2	1,0000	1,1709	2,0000	0,3769
15 ITEM 3	1,0000	1,0892	2,0000	0,2855
16 ITEM 4	1,0000	1,1429	2,0000	0,3504
17 ITEM 5	1,0000	1,0102	2,0000	0,1006
18 ITEM 6	1,0000	1,0615	2,0000	0,2406
19 ITEM 7	1,0000	1,0496	2,0000	0,2174
20 ITEM 8	1,0000	1,3377	2,0000	0,4735
21 ITEM 9	1,0000	1,6140	2,0000	0,4875

TABLA 2. Valores mínimo y máximo y desviación típica de algunas de las variables para el grupo 1. Media de esas variables en el cluster 1.

A través de esta información podemos descubrir la tendencia que se oculta en cada uno de los cluster. Será conveniente analizar, a la vez, el comportamiento de cada variable, en cada uno de los grupos. Para descubrir la tendencia sociopolítica que se oculta en el cluster será necesario, en esta etapa, no sólo la presencia de un analista sino también de un experto en sociología política. Ya hemos dicho que los grupos se constituyen «naturalmente», por sí mismos, de manera que, si hemos introducido suficiente información, el grupo constituido nos indica «una agrupación natural» que existe en la realidad. Sólo la aportación de un experto en investigaciones sociopolíticas podrá dar nombre a esa tendencia, o al conjunto homogéneo de actitudes políticas que describe cada uno de los clusters.

La tabla 2 nos proporciona el valor de cada variable en el centroide del grupo 1, es decir, la media; por otro lado nos da los valores máximo y mínimo de la variable en el grupo, además de su desviación típica. Naturalmente un grupo con alta desviación típica y con mucha distancia entre los valores máximo y mínimo, no estaría suficientemente cristalizado; no podríamos hablar, entonces, propiamente, de un grupo que represente una tendencia política nítida. Así pues, nos interesaría encontrar en esta tabla 2, un valor medio muy alto (lo que quiere decir que existe una variable que está dotando de significado al grupo) unos valores mínimos y máximos muy próximos a la media y una desviación típica muy pequeña.

Por ejemplo la variable P.5, que se refiere a la importancia que un problema socioeconómico tiene para la población, es poco representativa para significar al grupo, ya que, además de tener una media baja, tiene un recorrido muy amplio (de 1 a 6) y una desviación típica alta en comparación con las demás. En cambio tanto las va-

riables relativas a la cercanía a los líderes como a partidos son las más importantes para dar sentido a este grupo.

Otra información de interés aparece en la siguiente matriz.

PAGE 5

DISTANCES BETWEEN CLUSTER CENTERS

	1	2	3	4	5
2	2.87514				
3	2.54388	3.91284			
4	2.94460	3.94996	3.22731		
5	3.84182	4.99938	2.95049	2.73653	
6	3.34423	2.50565	2.65542	3.79370	3.61105

TABLA 3. Matriz de distancias entre los centroides de los grupos.

Esta matriz nos indica la distancia que hay entre los centroides de los grupos. Si estos están demasiado próximos, si no hay una clara zona o línea de separación, difícilmente podríamos hablar de tendencias naturales dentro del electorado potencial del partido. De ahí que sea recomendable, aún a costa de perder información, realizar este análisis con individuos próximos a los centroides respectivos de los grupos. Otros análisis realizados previamente podrán detectar estos individuos que hemos llamado puros (por ejemplo, el análisis discriminante).

Si proyectamos todos los individuos en las dimensiones representadas por los centroides de los grupos obtenemos el gráfico 3.

Este gráfico nos muestra geométricamente (y por lo tanto de una manera más clara) la distancia o proximidad entre unos grupos y otros, e, incluso, entre los individuos de un mismo grupo.

En la tabla 4 el análisis nos proporciona, una vez constituidos los grupos, la media de cada variable en cada grupo, así como la desviación típica. Aquí podemos ver el valor de cada variable en cada grupo.

Junto a esta información disponemos de un análisis de la varianza de la misma variable en el conjunto de todos los grupos (Between/Within).

Cuanto mayor sea el valor del estadístico F para las variables, más homogéneo será el grupo y más se distinguirá y diferenciará del resto de los grupos. En definitiva, cuanto mayor sea el valor de F más nítida será la tendencia del electorado potencial que representa cada grupo.

Una vez que ha acabado de dar toda la información correspondiente a $K = 6$, el programa vuelve a arrancar con $K = 5$, y luego con $K = 4$, dándonos para estos casos la misma información que para el primero.

Un análisis posterior nos indicará cuál de las tres agrupaciones es la que mejor refleja las tendencias del electorado potencial del partido.

La conclusión de un estudio de este tipo es evidente. Podemos proporcionar a los dirigentes del partido:

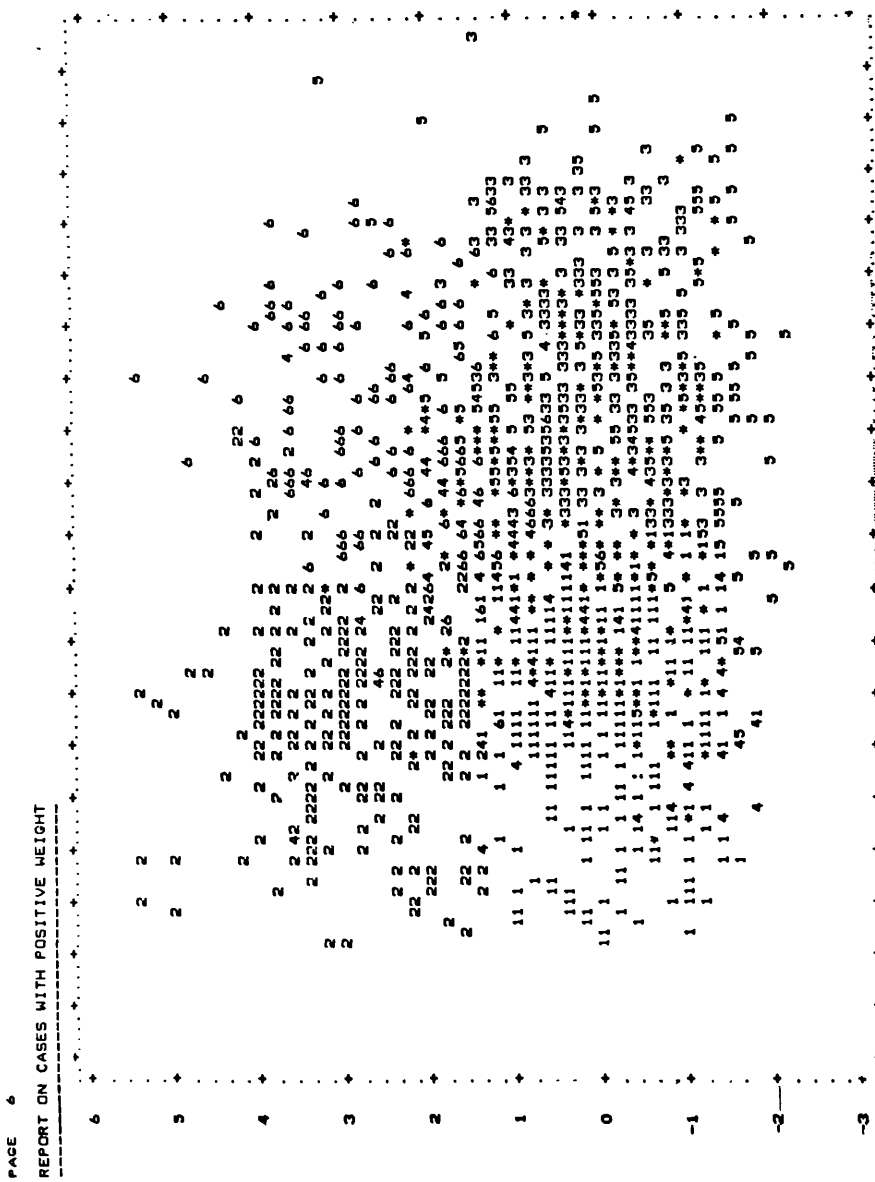


FIGURA 3. Proyección de los individuos en las dimensiones representadas por los centroides.

TABLA 4. a) media de las variables en los grupos, b) desviación típica de las variables en los grupos, c) análisis de la varianza de cada variable en el conjunto de los grupos.

PAGE 7

a)

CLUSTER MEANS

	PS ITEM2 ITEM14	LIDER1 ITEM3 ITEM15	LIDER2 ITEM4 ITEM16	LIDER3 ITEM5 ITEM17	LIDER4 ITEM6	LIDER5 ITEM7	UCD ITEM8	PSOE ITEM9	PCE ITEM10	AP ITEM11	CDS ITEM12	ITEM1 ITEM13
1	2.5662	4.3724	4.1885	1.7359	4.6107	4.4123	4.2305	1.7069	4.2387	4.5399	4.1435	1.2823
	1.1692	1.0844	1.1410	1.0102	1.0616	1.0472	1.3420	1.6158	1.1496	1.8553	2.3886	1.3542
	1.4925	2.6964	2.5561	2.3501								
2	2.1666	4.4545	4.1882	1.5136	4.7716	2.5529	4.2879	1.5397	2.3187	4.7519	4.1028	1.1891
	1.1138	1.0583	1.1883	1.0080	1.0207	1.0488	1.3610	1.7251	1.0625	1.7642	2.3164	1.2160
	1.4901	2.1015	2.2500	1.7283								
3	2.7604	3.8797	2.4865	1.3927	4.5518	4.5266	3.5536	1.4763	4.4246	4.5307	2.6357	1.2607
	1.0934	1.0686	1.1272	1.0102	1.0474	1.0642	1.2622	1.6163	1.1189	1.4599	2.0133	1.3870
	1.3856	2.5170	2.3910	2.1254								
4	2.7348	3.9610	3.8258	1.6956	2.8101	4.1795	3.8344	1.8269	4.0466	2.9866	3.7603	1.3176
	1.1554	1.1428	1.1544	1.0201	1.0616	1.0577	1.1507	1.3284	1.0903	1.4167	1.7329	1.4645
	1.3487	1.9539	1.8039	1.7532								
5	3.8212	2.9549	2.8250	1.8975	3.0672	4.1848	2.7879	1.9286	4.1139	3.0296	2.9220	1.5000
	1.1382	1.1485	1.1267	1.0176	1.0755	1.0620	1.2364	1.3262	1.1475	1.3275	2.0162	1.7168
	1.5877	2.6800	2.6960	2.7149								
6	2.5172	3.7677	2.6218	1.4528	4.5860	2.6392	3.5987	1.5126	2.6000	4.4258	2.6845	1.1360
	1.0526	1.0537	1.0855	1.0128	1.0200	1.0446	1.3699	1.6457	1.1007	1.8636	2.1446	1.2208
	1.4740	2.5223	2.5454	2.1923								
GRAND MEAN	2.7095	3.9671	3.4351	1.6229	4.1921	3.8761	3.7744	1.6606	3.7409	4.1642	3.4620	1.2812
	1.1261	1.0889	1.1394	1.0122	1.0493	1.0539	1.2960	1.5630	1.1158	1.6343	2.1479	1.3904
	1.4688	2.4598	2.4129	2.1764								

c)

PAGE 8

CLUSTER STANDARD DEVIATIONS

	P5 ITEM2 ITEM14	LIDER1 ITEM3 ITEM15	LIDER2 ITEM4 ITEM16	LIDER3 ITEM5 ITEM17	LIDER4 ITEM6	LIDER5 ITEM7	UCD ITEM8	PSOE ITEM9	PCE ITEM10	AP ITEM11	CDS ITEM12	ITEM1 ITEM13
1	0.7415	0.5838	0.5672	0.6945	0.5670	0.5625	0.5940	0.6734	0.5882	0.5370	0.6404	0.4501
	0.3750	0.2780	0.3480	0.1006	0.2405	0.2122	0.4744	0.4864	0.3567	0.8881	0.7835	0.6745
	0.7810	0.6286	0.7333	0.8433								
2	0.5455	0.5649	0.6103	0.5373	0.4891	0.8190	0.6042	0.6191	0.6575	0.4924	0.6555	0.3916
	0.3176	0.2344	0.3909	0.0894	0.1425	0.2154	0.4803	0.4465	0.2421	0.8514	0.8277	0.5073
	0.7833	0.9043	0.8892	0.8609								
3	0.6464	0.8139	0.6357	0.5017	0.5361	0.5379	0.8307	0.5193	0.5469	0.6102	0.7388	0.4390
	0.2910	0.2528	0.3332	0.1003	0.2126	0.2451	0.4398	0.4863	0.3237	0.7356	0.8562	0.6497
	0.6898	0.7636	0.8336	0.9064								
4	0.8604	0.7803	0.8203	0.5688	0.8506	0.8281	0.7408	0.6809	0.9043	0.8717	0.7878	0.4655
	0.3623	0.3499	0.3613	0.1405	0.2405	0.2332	0.3577	0.4696	0.2866	0.6976	0.8470	0.6842
	0.6811	0.9201	0.8786	0.8850								
5	1.1093	0.7399	0.7434	0.6788	0.8524	0.8596	0.6599	0.7035	0.8417	0.7993	0.6829	0.5000
	0.3452	0.3556	0.3326	0.1316	0.2641	0.2411	0.4248	0.4688	0.3546	0.6818	0.8213	0.8143
	0.7926	0.6358	0.6436	0.6229								
6	0.6114	0.7605	0.7011	0.6886	0.6083	0.7566	0.7808	0.7092	0.7238	0.7700	0.6563	0.3428
	0.2233	0.2254	0.2797	0.1125	0.1400	0.2064	0.4828	0.4783	0.3009	0.8533	0.8381	0.5375
	0.7746	0.7704	0.7567	0.8779								

b)

MEAN SQUARES												
BETWEEN	61.7706	72.0553	171.5270	9.6337	164.4701	193.7061	78.6417	7.8870	198.5013	140.8777	127.7398	2.9796
	0.4084	0.3342	0.2232	0.0045	0.1097	0.0159	1.3601	5.1822	0.2806	13.8237	13.6094	7.4903
	1.5477	21.2260	18.4919	31.4798								
WITHIN	0.5699	0.4874	0.4388	0.3850	0.4134	0.5041	0.4831	0.4201	0.4772	0.4342	0.4744	0.1923
	0.1093	0.0802	0.1197	0.0122	0.0468	0.0513	0.2045	0.2263	0.1019	0.6430	0.6819	0.4327
	0.5731	0.5772	0.6236	0.7052								
D.F.-S	5.1320	5.1461	5.1501	5.1527	5.1506	5.1506	5.1471	5.1500	5.1481	5.1477	5.1420	5.1384
	5.1436	5.1388	5.1421	5.1460	5.1411	5.1493	5.1416	5.1239	5.1340	5.1460	5.1520	5.1469
	5.1474	5.1470	5.1461	5.1478								
F-RATIO	108.384	147.837	390.862	25.024	397.884	384.236	162.784	18.774	415.986	324.429	269.251	15.497
	3.736	4.168	1.864	0.368	2.336	0.310	6.652	22.901	2.755	21.497	19.959	17.312
	2.701	36.776	29.651	44.641								
P-VALUE	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.001	0.000	0.084	0.900	0.030	0.932	0.000	0.000	0.012	0.000	0.000	0.000
	0.013	0.000	0.000	0.000								

1. El número de tendencias sociopolíticas que se ocultan detrás de su electorado potencial y, además, la importancia de cada una. Podríamos llegar a decir que este partido que estamos analizando se compone de:

—un 15% de liberales conservadores
—un 20% de derecha civilizada
—un 20% de católicos tradicionales
—un 15% de centro progresista
—un 10% de populistas tradicionales
—un 8% de nostálgicos del franquismo
—un 6% con netas tendencias ultras
—un 6% «captados» por la personalidad del líder del partido

2. Además de esta información, y después de haber estudiado la salida de datos, podremos proporcionar a los dirigentes del partido la nitidez, fuerza y cohesión de cada una de esas tendencias. En definitiva podremos decir «algo» sobre su permanencia en un futuro próximo.

3. De la misma manera podemos dar información sobre el tipo de personas que componen cada grupo. Así, podríamos decir que el 15% de liberales conservadores tienen entre 45 y 65 años, con rentas por encima del millón de pesetas anuales, viven en grandes ciudades..., etc.

4. Incluso podríamos decir, introduciendo la variable intención de voto, en qué grupo se encuentra la mayor probabilidad de ganar o perder votos en unas próximas elecciones.

5. Naturalmente, si se mantiene esta estructuración del electorado potencial, es posible conocer en qué medida ha afectado a cada uno de ellos una concreta del partido, una declaración de un líder, una intervención parlamentaria concreta o, en otro sentido, lo que el electorado potencial (no en su conjunto sino segmentado en tendencias) piensa acerca de un problema nacional como puede ser la permanencia en la OTAN, la Ley de Educación, etc...

Nota del Editor

Programas de Ordenador

El programa de ordenador utilizado en este ejemplo del análisis está incluido en el paquete estadístico BMDP. La mayoría de los paquetes generales de análisis de datos (BMDP, SAS, OSIRIS, etc.) incluyen programas de análisis de conglomerados, con opciones para elegir entre diferentes métodos (jerárquicos, de partición, etc.).

Existe un paquete específico de análisis de conglomerados, CLUSTAN, desarrollado por David Wishart a finales de los 60 y principios de los 70. El paquete incluye diferentes medidas de distancia y de similaridad y también un amplio número de algoritmos de clasificación. Se puede obtener información sobre el paquete en:

The Clustan Project
Computer Centre
University College London
London WC1H OAH
Inglaterra

Los dos libros que ofrecemos a continuación incluyen programas de análisis de conglomerados escritos en Fortran:

Anderberg, M. R., *Cluster Analysis Aplications,* Nueva York, Academic Press, 1973.
Hartigan, J. A., *Clustering Algorithms,* Nueva York, Wiley & Sons, 1975.

SECCION III

AJUSTE DE MODELOS

8. Introducción

Las técnicas que hemos explicado hasta este momento tienen como función la de *describir* los fenómenos sociales, bien sea clasificando o reduciendo la información. Como resultado de las técnicas clasificatorias llegaremos a conclusiones del tipo «en el colectivo de los votantes del PSOE se puede ditinguir una serie de grupos o conglomerados» (análisis de conglomerados), o, en función de la información disponible cabe clasificar a tal individuo, que no contesta sobre su intención de voto, entre los votantes del PCE» (análisis discriminante). Como resultado de las técnicas de reducción podemos ver, por ejemplo, cómo se agrupan las variables en un número menor de factores, sin pérdida de la información original (análisis factorial); cuáles son las dimensiones que subyacen a la percepción de los líderes políticos por parte de un grupo de entrevistados (escalas multidimensionales), o cuáles son las categorías de un grupo de variables que se asocian con la categoría de otra variable (análisis de correspondencias).

Las técnicas que vamos a introducir en esta sección pretenden *explicar* los fenómenos sociales, partiendo de la formalización de una teoría que determina las variables pertinentes para su explicación. Para ello se formularán modelos que postulen las interrelaciones entre las variables; mediante una serie de operaciones estadísticas se puede ver la bondad del modelo que se formula y también estimar la influencia de unas variables explicativas sobre otras explicadas, que a su vez pueden ser también explicativas.

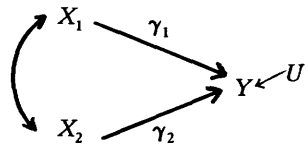
Desde un punto de vista práctico la diferencia fundamental entre los 4 artículos radica en el tipo de variables utilizadas: los dos primeros explican la construcción de modelos causales con variables intervales, mientras que los dos últimos parten de las tablas de contingencia (con variables nominales). El hecho de utilizar diferente nivel de medida hará que los métodos de estimación de los efectos entre las variables sean distintos. Sin embargo los principios fundamentales son semejantes. Veamos por separado cada uno de los artículos.

Alberto Satorra introduce los modelos causales con variables cuantitativas observadas. El punto de partida de su artículo es la explicación de los modelos uniecuacionales. Tales modelos pretenden contestar a la pregunta: «¿qué variables explicativas (independientes) influyen en una variable explicada, respuesta o endógena (dependien-

te)?». El modelo más frecuentemente utilizado se puede expresar en la siguiente ecuación:

$$Y = a + \Sigma\gamma_i X_i + U$$

donde Y es la variable endógena, a es un valor constante que afecta a todos los casos de la población, X_i es la variable explicativa, γ_i es el efecto de X_i sobre Y, y U es el término de error o de perturbación del modelo, que recoge la variación de Y que no puede ser explicada por las variables endógenas, X. Un modelo explicativo de este tipo con dos variables explicativas se puede representar mediante un diagrama como el que ofrecemos a continuación,



en donde una flecha simboliza un efecto «directo» de una variable sobre otra explicado por el modelo, mientras que una flecha doble representa la correlación entre dos variables no explicada por el modelo.

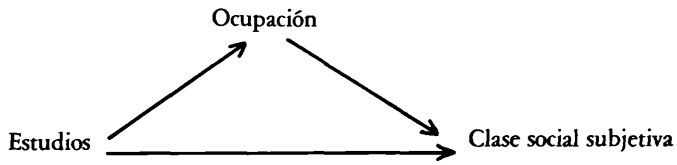
El modelo precedente es la formalización de una teoría causal en donde varias variables explicativas se combinan para dar como resultado un valor de la variable endógena. Este modelo es lineal y aditivo, lo cual significa que el efecto de X_i sobre Y es siempre el mismo cualquiera que sea el valor de X_i o de las restantes variables explicativas (véase *infra*). También se hace el supuesto de que las variables están medidas sin error, pues de lo contrario se producirían sesgos importantes en los resultados —en su artículo, Saris explica la solución a adoptar cuando se sospecha que las variables tienen error.

Respecto del término de error o de perturbación, U, se hace el supuesto de que se trata de una variable aleatoria en la que 1) el valor de U para un individuo (observación) no está correlacionado con el valor de U para otro individuo (observación) —no existe autocorrelación—; 2) tampoco U está correlacionado con las variables explicativas. Al tiempo hay que hacer el supuesto adicional 3) en el sentido de que la varianza de U permanece constante cualesquiera que sean los valores que tome la variable X_i —hipótesis de homoscedasticidad—. Cuando se cumplen estos requisitos estamos ante el modelo de regresión lineal múltiple, siendo posible estimar los efectos de las variables explicativas mediante el método de los mínimos cuadrados ordinarios (MCO).

A partir del modelo causal explicado se pueden considerar otros modelos en los que 1) las variables explicativas en una ecuación sean variables explicadas (endógenas) en otra —modelos multiecuacionales—, y 2) en donde se rompan algunos de los supuestos hechos previamente.

El tipo de modelos de los que se habla en 1) son útiles cuando se trata de estudiar temas tales como los mecanismos a través de los que una variable explicativa influye

en otra variable respuesta. Por ejemplo, se puede observar que el nivel de «estudios» está relacionado con la «clase social subjetiva» de las personas; un modelo causal multiecuacional trataría de ver qué variables pueden intervenir en esta relación. Cabe pensar que los «estudios» también están relacionados con la «ocupación» y que esta variable se relaciona a su vez con la «clase social subjetiva» (véase representación gráfica en la figura siguiente).



Un modelo uniecuacional que incluyera las 3 variables permitiría conocer cuál es el efecto de cada una de las 2 variables explicativas sobre la «clase social subjetiva»; un modelo multiecuacional consideraría la existencia de 2 variables endógenas (dependientes), «ocupación» y «clase social subjetiva», y estudiaría los efectos directos de «estudios» sobre estas dos variables y su efecto indirecto sobre «clase social subjetiva» vía «ocupación». La relación original entre «estudios» y «clase social subjetiva» quedaría así descompuesta en estos dos nuevos efectos. Con el fin de dar solución al modelo precedente se hace necesario resolver tres problemas, relacionados entre sí: la especificación, la identificación y la estimación de los parámetros del modelo.

La *especificación* del modelo hace referencia a la necesidad de indicar, en base al conocimiento teórico sobre el tema, la pauta de relaciones que establecen todas las variables entre ellas y con los términos de error, y de estos últimos entre sí. Un tipo particular de modelo con el que nos vamos a encontrar a la hora de hacer la especificación es el *recursivo;* en este tipo de modelo se asume que si una variable endógena influye en otra variable endógena, esta segunda no tiene un efecto, directo o indirecto, sobre la primera. También se asume que los términos de error están incorrelacionados entre sí e incorrelacionados con las variables explicativas. Caso de que alguna de estas condiciones no se cumpla estamos ante un modelo *no-recursivo.*

El problema de la *identificación* tiene que ver con la cantidad de información necesaria para estimar los parámetros del modelo. En función de la especificación que hayamos hecho hay que construir un sistema de ecuaciones que incluya como incógnitas aquellos efectos que hay que estimar. Para proceder a la resolución del sistema contamos con la información procedente de las relaciones (varianzas y covarianzas, o correlaciones) entre las variables. Si la información no es suficiente, los parámetros del modelo no se pueden identificar. Cuando se trata de modelos recursivos sin error de medida en las variables siempre es posible identificar los parámetros. Cuando el modelo sea *no-recursivo* hay que ver si el sistema de ecuaciones cumple las condiciones necesarias para proceder a una estimación única de los parámetros.

Especificado el modelo y visto que es posible la identificación única de los parámetros hay que proceder a su *estimación,* dado que los parámetros son desconocidos por el investigador. En los modelos recursivos el problema se reduce a aplicar la regresión múltiple (que utiliza la estimación mínimo cuadrática) a cada una de las

ecuaciones del modelo. Cuando los modelos son *no-recursivos* los MCO no son un buen método de estimación. En su lugar se pueden utilizar dos tipos de métodos: los llamados de *ecuación singular* (o *información limitada*), que estiman las diferentes ecuaciones del modelo ecuación por ecuación; y los métodos de estimación de *sistema completo (información completa)*, que estiman conjuntamente todos los parámetros. El más común de los primeros métodos es el denominado de los mínimos cuadrados en dos etapas (MC2E). Entre los segundos Satorra incluye las estimaciones de mínimos cuadrados generalizados o ULS *(unweighted least squares)* y de máxima verosimilitud (MV). Todos ellos son explicados por el autor, con sus ventajas e inconvenientes, ilustrando su aplicación con dos ejemplos donde se comparan los resultados obtenidos con los diferentes métodos de estimación.

Cuando el modelo está sobreidentificado (véase texto), utilizando MV Satorra explica cómo es posible contrastar estadísticamente la validez global del modelo.

El artículo de Willem Saris es la continuación del trabajo de Alberto Satorra. Da soluciones al problema que se plantea en las ciencias sociales cuando se trata de explicar algún fenómeno y la información que obtenemos contiene errores de medida. En el trabajo en cuestión se discuten tres enfoques diferentes para resolver el problema del error aleatorio de medida en los modelos de educaciones lineales estructurales; se presentan los modelos apropiados y se explican su estimación y verificación.

El punto de partida es la necesidad de estudiar las relaciones entre las variables haciendo uso de los métodos explicados por Alberto Satorra. Cuando las variables están medidas con error, la estimación puede llevar a conclusiones erróneas sobre el impacto causal de las variables explicativas sobre las explicadas, y casi siempre implica una sobrevaloración del efecto perturbación —es decir, de la influencia de las causas desconocidas sobre la variable dependiente. (P. V. Marsden, 1981: 203).

Por lo demás, tales errores son frecuentes en la medición que se hace en las ciencias sociales y es conocido, por ejemplo, el efecto que puede tener el uso de diferentes entrevistadores o la distinta formulación de una pregunta sobre las contestaciones de los entrevistados. Cuando tal cosa ocurre, la matriz de varianzas-covarianzas de las variables observadas no es útil par estimar los efectos causales, y previamente a la estimación es necesario estimar la matriz de varianzas-covarianzas de las variables verdaderas (no observadas). A tal fin se pueden seguir diferentes alternativas, aun cuando todas ellas tienen en común el hecho de que obtienen varias medidas de cada variable para utilizarlas después en modelos que operacionalizan las teorías existentes sobre la medida y el comportamiento de los errores. Estas mediciones múltiples, que se hacen tanto de las variables explicativas como de las explicadas, son las variables observadas, mientras que los valores verdaderos aparecen como variables no observadas.

Junto a variables observadas y no observadas hay que considerar los errores en las variables (distintos de los errores en la ecuación o término de perturbación), que pueden considerarse aleatorios o correlacionados, debido a la existencia de una fuente común de sesgo en las mediciones (por ejemplo, la tendencia de algunos entrevistados a decir «sí» a todas las preguntas, al margen de cuáles sean éstas).

Una vez que hemos conseguido obtener la matriz de varianzas-covarianzas de las variables verdaderas estamos en condiciones de estudiar las influencias o efectos de unas variables sobre otras, utilizando los métodos de estimación de los modelos causales. LISREL es un sistema general, implementado en un programa de ordenador, que

permite la resolución de ambos problemas. Siguiendo el trabajo de los psicólogos en el campo del error de medida incorpora la distinción entre variables teóricas y variables observadas, permitiendo estudiar la relación entre ambas (modelo de medida); a partir del trabajo de los económetras en el campo de la causalidad recíproca y en el desarrollo de métodos eficientes de estimación permite la estimación de los parámetros desconocidos (modelo causal o de ecuaciones estructurales). En este sentido, la suma de los trabajos de Satorra y de Saris ofrece una visión completa del sistema, al abordar los autores las cuestiones inherentes a cada una de las dos problemáticas.

En su artículo, Saris plantea tres enfoques diferentes para estimar las varianzas-covarianzas de las variables teóricas (a partir de las relaciones entre las variables observadas y las teóricas). El primero de ellos consiste en el uso de *indicadores múltiples* para cada variable teórica; estos indicadores pueden ser paralelos tau-equivalentes o congenéricos —siendo los dos primeros un caso particular del tercero—. Otro enfoque consiste en tomar información de la misma variable en diferentes momentos del tiempo *(replicación)*. Por último el autor plantea la posibilidad de utilizar la replicación con *indicadores múltiples* para cada observación. En el trabajo en cuestión se discuten las ventajas y desventajas de cada uno de los enfoques y se ofrece un ejemplo que ilustra el tercero de ellos, dado que los dos primeros ya han sido discutidos por otros autores.

Las técnicas que vamos a explicar a continuación difieren básicamente de las precedentes en que el tipo de información que utilizan es de tipo cualitativo (variables nominales, fundamentalmente). Este hecho condiciona el procedimiento a seguir para determinar la relación entre las variables y el efecto de unas sobre otras. Sin embargo, tanto los modelos lineales logarítmicos como los sistemas de la D son un intento de seguir el trabajo de los económetras en el desarrollo de modelos causales, sin olvidar que la mayoría de las variables que se utilizan en las ciencias sociales son cualitativas. En ambos casos, las técnicas que comentamos ahora tienen limitaciones respecto de los modelos causales con variables cuantitativas:

1. En primer lugar, ninguna de las técnicas permite introducir en la estimación de los efectos el error de medida.
2. Tampoco es posible resolver modelos con efectos causales recíprocos (modelos no-recursivos).

A estas dos limitaciones, en el caso de los modelos lineales logarítmicos hay que añadir la imposibilidad de descomponer el tamaño de los efectos de las variables, midiendo los efectos causales (directos e indirectos) y los espúreos. Por lo demás, con las singularidades que vamos a exponer a continuación ambas técnicas tratan de resolver los mismos problemas que la correlación o la regresión (simples o múltiples) y las ecuaciones estructurales cuando se tienen variables cuantitativas.

Resuelto el problema de la especificación del modelo, y sin que haya problema de identificación en este caso, tanto los modelos lineales logarítmicos como los sistemas de la D proceden a la estimación de los parámetros. Puesto que ambs técnicas están basadas en medidas de asociación diferentes, también será diferente el procedimiento de estimación. En los modelos lineales logarítmicos la medida de la intensidad de la relación son las *razones* (comparación de las frecuencias de las categorías entre sí) y en los sistemas de la D la diferencia de proporciones (comparación de las frecuencis de las

categorías con los marginales). A partir de estas medidas, lo mismo que en los modelos precedentes se trataba de ver cómo se descomponía la correlación entre dos variables buscando las fuentes de la covariación, es decir, explicando los mecanismos a través de los que se generaba la correlación, ahora se buscarán las fuentes que determinan una razón (o logaritmo de la razón) o una diferencia de proporciones, estimando la influencia de esas fuentes.

Cuando se trata de ver el efecto de una variable explicativa sobre otra variable respuesta, en los modelos con variables intervales se estimaba un único número. Esto era posible dada la naturaleza continua de las variables intervales (los individuos son «más» o «menos» en la dimensión que se estudia, por ejemplo la edad). Sin embargo, cuando se tienen variables nominales, tal continuo no existe, por el contrario aparecen estados o categorías (el católico, el protestante, etc., son estados diferentes, ni más ni menos, en la dimensión religiosa de los individuos), y ello obliga a que no haya que calcular un solo efecto de «toda» la variable explicativa sobre «toda» la variable explicada, sino que haya diferentes efectos para las diferentes categorías o estados de ambas variables.

A diferencia con los modelos con variables intervales un aspecto que tiene especial interés en los modelos lineales logarítmicos y en los sistemas de la D es la interacción entre las variables. En el caso de las variables intervales un ejemplo de interacción sería el siguiente: el precio de la cesta de la compra (variable dependiente) viene determinado por el precio de las patatas (efecto de la variable «patatas») por su cantidad más el precio de las lechugas (efecto de la variable «lechuga») por su cantidad (ambas variables independientes). Esto sería así en el supuesto de que no hubiera una oferta que hiciera que comprando una determinada cantidad de lechugas y de pan los precios se rebajaran. Es decir, el precio total no siempre es la suma de los productos de la cesta, sino que determinadas combinaciones de valores (kilogramos) de las variables pan y lechuga interactúan modificando el precio total. En el caso de las variables nominales la interacción significa que la relación entre dos variables está condicionada por las categorías de una tercera variable. Por ejemplo, entre los individuos sin estudios (categoría 1 de la tercera variable, estudios) existe una relación muy fuerte entre la edad y la práctica deportiva; sin embargo, entre los individuos con estudios superiores (categoría 2 de la tercera variable) la relación casi se desvanece: tanto jóvenes como menos jóvenes practican deporte, es decir, la edad es independiente de la práctica deportiva. Pues bien, uno de los posibles efectos a considerar cuando se trata de modelos causales con variables cualitativas es esta posible influencia de la interacción, que hay que añadir como uno más de los factores que explican la relación entre dos variables.

En la medida en que utilizamos los mismos datos con ambas técnicas es posible comparar los métodos en función de los resultados obtenidos[1]. Mirando los dos gráficos que muestran la estimación de los efectos en ambos casos (gráficos 9 y 4 en los artículos sobre los modelos lineales logarítmicos y los sistemas de la D, respectivamente) vemos que: 1) el signo de los efectos es el mismo; 2) sus tamaños relativos son también similares, excepto en el caso del efecto de Permiso sobre Ingresos que es ma-

[1] Se puede ver una discusión más técnica sobre las diferencias que existen entre los enfoques para el análisis de tablas de contingencias en Kritzer (1979).

yor cuando se utilizan proporciones; y 3) en ningún caso hay interacciones de 3.º y 4.º orden[2].

Tal como señala James Davis, parece que la elección entre uno u otro método (modelos lineales logarítmicos o sistemas de la D) es un problema de «gusto»; algunas personas se sentirán más cómodas tratando con razones, mientras que otras encontrarán las proporciones como una medida más natural. La elección, pues, queda en la mano del lector.

[2] Su contraste se omite en el caso de los sistemas de la D, aunque se puede ver su cálculo en la obra citada en el capítulo respectivo, Sánchez Carrión (1983).

9. Introducción a los modelos de causalidad *

por *A. SATORRA* ** (autor principal) *y L. H. STRONKHORST*

9.1. Estudios descriptivos y explicativos

Las investigaciones empíricas en el ámbito de las ciencias sociales pueden clasificarse, a grandes ragos, en dos tipos: en descriptivas y explicativas. El objetivo principal de una investigación de tipo descriptivo es la presentación de cifras o datos relativos a la situación o estado de un determinado fenómeno social o humano; por ejemplo, a la pregunta «¿ha aumentado la delincuencia en los últimos años?» corresponde un estudio de tipo descriptivo que presente las cifras actuales de delincuencia y las compare con las cifras de delincuencia habidas en años anteriores. En cambio, un estudio de tipo explicativo no se limita simplemente a la presentación de datos, sino que pretende «explicar» unos fenómenos o acontecimientos a partir de la ocurrencia o no de otros fenómenos o sucesos. Por ejemplo, a la pregunta «¿influye la cantidad de violencia que se proyecta en la televisión sobre la cantidad de violencia presente en la calle?», corresponde un estudio de tipo explicativo que investigue la posible interrelación o relación de «causalidad» entre las dos formas de violencia mencionadas. En general, un estudio de tipo descriptivo comporta la recolección de datos, tabulación y representación gráfica de los mismos, etc.; mientras que un estudio de tipo explicativo implica la formulación de hipótesis estadísticas o «modelos» de interrelación entre variables, la contrastación estadística de las mismas, así como la estimación de la magnitud de los «efectos» hipotéticos entre variables.

En este capítulo presentaremos la metodología estadística adecuada para el tratamiento de los modelos de causalidad (también denominados modelos de ecuaciones estructurales, sistemas de ecuaciones simultáneas, modelos del «path analysis», etc.). Dicha metodología, propia de los estudios de tipo explicativo, es adecuada para el tratamiento estadístico de datos multivariantes generados en un contexto «no-experimental»[1]. Dado el carácter introductorio del presente capítulo, supondremos

* Parte del presente trabajo ha sido posibles gracias a una beca concedida al primer autor por la Dutch Organization for Advancement of Pure Research (ZWO).

** Quiero agradecer las sugerencias obtenidas en el departamento de Estadística y Econometría de la Universidad de Barcelona, que han permitido mejorar una versión preliminar de este artículo.

[1] Para una clarificación de los conceptos relativos a la investigación empírica con datos «no experimentales» véase Blalock (1961) y Saris y Stronkhorst (1983).

que todas las variables relevantes son manifiestas (es decir, observables)[2] y medidas en escala intervalo[3].

9.2. Modelos uniècuacionales

La situación más simple de aplicación de un modelo causal es aquella en la que se pretende estudiar los efectos que sobre una variable Y, que llamaremos variable dependiente o *endógena,* tiene la variación de una variable X, que llamaremos variable *explicativa* o independiente. En dicha situación puede proponerse una ecuación como la siguiente:

$$Y = f(X) \qquad [1]$$

en donde $f(X)$ representa una función específica de X. Dicha ecuación [1] implica que, para un individuo cualquiera de la población, el valor de la variable X determina con exactitud el valor de la variable Y. Sin embargo, a pesar del carácter altamente explicativo que tendría un modelo «exacto» o determinístico como el propuesto en [1], es impensable que dicho modelo se ajuste a una realidad observacional o empírica. Efectivamente, si x_i e y_i son respectivamente los valores de X e Y en el individuo i-ésimo de la muestra (i es un subíndice que caracteriza el individuo en la muestra y que supondremos varía de 1 a N) entonces, lo más probable es que el valor o diferencia $y_i - f(x_i)$ no sea exactamente igual a cero. Esta discrepancia entre Y y $f(X)$ puede ser debida a diferentes razones, como, por ejemplo, a que:

a) La especificación funcional $f(X)$ considerada es incorrecta.
b) Existen otras «causas» de variación de Y distintas de $f(X)$.
c) Los valores x_i e y_i no son los valores «verdaderos» de las variables X e Y (es decir, existe «error de medida»).

En cualquier caso, como solución a la escasa plausibilidad de un modelo determinístico como el propuesto en [1], puede proponerse el modelo alternativo siguiente:

$$Y = f(X) + U \qquad [2]$$

en donde la nueva variable U, que corresponde a la diferencia $(Y - f(X))$, se denomina *término de perturbación* del modelo. Dicha variable U no se corresponde con ninguna característica sustantiva (teórica) de los individuos (es simplemente un rudimento estadístico que recoge la discrepancia entre Y y $f(X)$ y es en general una variable no observable. Sin embargo, a fin de que [2] implique una estructura específica de variación de Y respecto de X (de momento [2] es solamente una identidad), será necesa-

---

[2] La problemática de las variables no observables y del error de medida será tratada en el próximo capítulo.

[3] Una referencia importante para la clarificación de los conceptos relativos a las diferentes escalas de medida es la de Stevens (1946).

rio incorporar a dicha ecuación algún supuesto relativo al comportamiento de dicho término de perturbación U. Concretamente, tendremos que introducir algún supuesto operativo el cual implique que el término de perturbación U recoge solamente aquellos factores o «causas» de variación de Y que no tienen ninguna relación o característica común con la variable X (de lo contrario $f(X)$ no recogería toda la variación de Y debida a X). Ello se consigue suponiendo que si u_i es el valor de U en el individuo i-ésimo de la muestra (recordemos que i varía de 1 a N), entonces dicho valor es el resultado de un fenómeno aleatorio; es decir, se supone que u_i es la realización de una variable aleatoria U_i asociada al individuo i-ésimo de la muestra. Con respecto a la distribución de probabilidad de las variables aleatorias U_1, U_2, ..., U_i, ..., U_N, asociadas a los diferentes individuos de la muestra, supondremos lo siguiente:

(i) La distribución de probabilidad de U_i ($i = 1, 2, \ldots, N$) es independiente de los valores que toma la variable X.

(ii) Las variables aleatorias U_1, U_2, ..., U_N son idénticamente distribuidas e independientes entre sí.

(iii) La esperanza de U_i ($i = 1, 2, \ldots, N$) es cero.

(iv) La distribución de probabilidad de U_i ($i = 1, 2, \ldots, N$) es normal.

La ecuación [2] junto con los supuestos (i) a (iv) constituye una especificación completa de lo que se denomina un *modelo estocástico* (obviamente, el calificativo de estocástico proviene del carácter estocástico o aleatorio que hemos supuesto tiene el término de perturbación U)[4]. La naturaleza estocástica del modelo posibilitará al investigador que dispone de observaciones de las variables X e Y en N individuos de la población (o sea, al investigador que dispone de «información muestral») aplicar técnicas inferenciales de estadística a fin de obtener información sobre las características «poblacionales» del modelo. Por ejemplo, el investigador podrá estimar los coeficientes involucrados en la expresión $f(X)$ que a priori no vienen determinados por la teoría, o contrastar hipótesis estadísticas relativas a dichos coeficientes.

La expresión de [2] más común, y considerada en este capítulo, es la siguiente:

$$Y = \gamma X + U \qquad [3]$$

en donde se supone que las variables Y y X vienen expresadas en desviaciones respecto a sus medias[5] y que γ es un coeficiente de valor desconocido «a priori». Como argumento a favor de la expresión lineal [3] podemos citar, por ejemplo, que una función lineal es una «buena» aproximación de cualquier función cuando el rango de variación del argumento es pequeño; o bien, que [3] es una expresión operativa en la que el

[4] Nótese que la suposición de que u_i es la realización de una variable aleatoria U_i implica que el valor y_i de la variable endógena Y es también la realización de una variable aleatoria. Es decir, la variable Y es también de naturaleza estocástica.

[5] Es decir, se supone que la media de las observaciones de cada variable es igual a cero; o sea,

$$\sum_1^N x_i = \sum_1^N y_i = 0.$$

coeficiente γ tiene una significación empírica concreta: γ mide el cambio esperado en Y provocado por un incremento unitario en X. Dicho coeficiente γ, así como la varianza de U, σ^2, son denominados parámetros (estructurales) del modelo; siendo, en general, la determinación de la magnitud de los mismos uno de los objetivos centrales del investigador. Hay que señalar que la magnitud de σ^2 informa sobre la cantidad de variación de Y «no explicada» por la variable X.

Con respecto a los supuestos (i) a (iv), necesarios para la validez de las técnicas propuestas en este capítulo, cabe comentar una serie de extremos. En primer lugar, la suposición (ii) implica, en particular, que la varianza de U_i es constante respecto a i; es decir, es idéntica para todos los individuos de la muestra. Dicha implicación, que se conoce como la hipótesis de *homoscedasticidad,* puede resultar insostenible en algunas aplicaciones (como, por ejemplo, cuando los individuos son agregados de «tamaños» muy desiguales). Cuando la varianza de U_i varía atendiendo al valor de la variable X, se dice entonces que el término de perturbación es heteroscedástico y, en este caso, previo a la utilización de las técnicas propuestas en el presente capítulo, es necesario aplicar alguna transformación específica a las variables o ponderar las observaciones. No insistiremos sobre este extremo y simplemente remitimos al lector interesado en el problema de la heteroscedasticidad a que consulte cualquier manual de econometría (como, por ejemplo, Johnston, 1963; Goldberger, 1970).

Respecto a la suposición (ii), de independencia estocástica de las diferentes realizaciones de U, cabe señalar que, si bien es una suposición «natural» en un contexto de datos de «corte-transversal» (por ejemplo, con datos de encuesta), cuando los datos son de «serie-temporal» es probable que la serie $U_1, U_2, \ldots, U_N$ presente «autocorrelación» en cierto grado. Los económetras han desarrollado ampliamente técnicas específicas para el tratamiento del problema de la «autocorrelación» (véase, por ejemplo, los manuales de econometría anteriormente citados); sin embargo, puesto que el presente capítulo va dedicado principalmente al tratamiento de datos de «corte-transversal» (datos que, por otra parte, son abundantes en las ciencias sociales), la problemática de la autocorrelación no será objeto de atención en este capítulo.

Una generalización inmediata del modelo [3] se obtiene cuando en lugar de una sola variable X consideramos una serie $X_1, X_2, \ldots, X_K$ de variables explicativas. En este caso, la ecuación del modelo será:

$$Y = \gamma_1 X_1 + \gamma X_2 + \ldots + \gamma_K X_K + U \qquad [4]$$

en donde se supone que U verifica también los supuestos (i) a (iv), teniendo en cuenta que el supuesto (i) se convierte ahora en:

(i) La distribución de probabilidad de $U_i (i = 1, 2, \ldots, N)$ no depende de los valores que toman las variables $X_1, X_2, \ldots, X_K$.

El coeficiente $\gamma_j (j = 1, 2, \ldots, K)$, que aparece en la ecuación [4], se denomina coeficiente de *regresión parcial* correspondiente a X_j y expresa la variación provocada en Y por un incremento unitario de X_j en el supuesto que las restantes variables explicativas de la ecuación permanecen fijas (o «controladas»).

Esperamos que el lector haya reconocido en [4] al modelo de regresión lineal múl-

tiple con K variables explicativas[6]. Además el lector debe también reconocer en los supuestos (i) a (iv) las hipótesis básicas que garantizan que el método de los mínimos cuadrados ordinarios (de aquí en adelante, MCO) conduce a un tratamiento inferencial «óptimo». Cabe señalar también que los supuestos (i) a (iv) garantizan que el método de los mínimos cuadrados ordinarios sea equivalente al método de la máxima verosimilitud propuesto, más adelante, en el apartado de este capítulo correspondiente a la estimación.

En el contexto del *path analysis* una forma habitual de representar la especificación de un modelo es mediante los denominados diagramas *path*. En el caso de un modelo uniecuacional con dos variables explicativas, dicha representación sería la que muestra la figura 1.

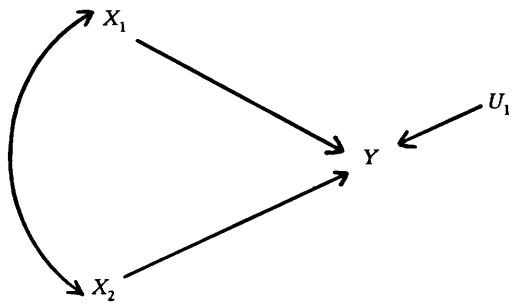


FIGURA 1. 1) Diagrama «path» de un modelo uniecuacional con dos variables explicativas.

En el diagrama que presenta la figura 1 una flecha simboliza un «efecto directo» de una variable (extremo inicial) sobre otra variable (extremo final) explicado por el modelo, mientras que una doble flecha representa una correlación entre dos variables no explicada por el modelo.

9.3. Modelos multiecuacionales

Un investigador que intente operativizar o contrastar una teoría se verá obligado, muchas veces, a considerar modelos más complejos que el simple modelo uniecuacional presentado en el apartado anterior. Puede suceder, por ejemplo, que

[6] Un supuesto adicional de los autores respecto al actual lector es que, éste, al empezar el capítulo, estará ya familiarizado con el modelo de regresión a un nivel de curso introductorio en estadística. En caso de que dicho supuesto no sea cierto, remitimos al lector a que consulte, por ejemplo, Thomas (1980).

alguna de las variables que actúan como explicativas en una ecuación sea al mismo tiempo una variable «explicada» o dependiente (endógena) en otra ecuación del modelo. Supongamos, a efectos ilustrativos, un estudio hipotético en el ámbito educacional en el que se especifica una ecuación de regresión que «explica» las aspiraciones educacionales (AE) (variable dependiente) de los estudiantes de BUP por medio de una serie de variables explicativas entre las que se incluye el rendimiento académico (RA) del estudiante. En dicho estudio, cabe pensar que el rendimiento académico RA será a su vez variable «explicada» o dependiente en otra ecuación de regresión del modelo (sugerida también por la «teoría») en la que entre las variables explicativas figurará la variable de aspiraciones educacionales AE. El hecho de que entre las variables explicativas de una ecuación de regresión figuren variables que simultáneamente son variables endógenas o «explicadas» en otra ecuación del modelo implicará, en general, la utilización de técnicas estadísticas más complejas que las meras técnicas de regresión.

Efectivamente, la naturaleza endógena de una variable explicativa implicará, en general, la existencia de correlación entre dicha variable explicativa y el término de perturbación de la correspondiente ecuación y, por tanto, que los estimadores MCO sean estimadores inconsistentes[7]. Siguiendo con el mencionado estudio en el ámbito educacional, la existencia de correlación entre el término de perturbación y las variables explicativas queda manifiesta considerando el modelo de dos ecuaciones representado en la figura 2 (en donde I representa el resultado de un test de inteligencia y NEP es una medida del nivel educacional de los padres).

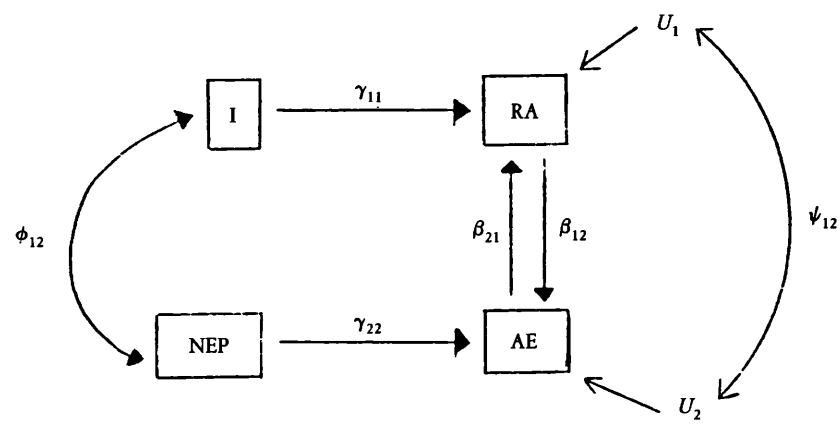


FIGURA 2. Diagrama *path* correspondiente al modelo educacional.

[7] Recordamos al lector que, al estimar una ecuación de regresión, un supuesto básico para asegurar la consistencia de los MCO es que las variables explicativas estocásticas estén incorrelacionadas con el término de perturbación de la ecuación.

En dicho modelo, la variable AE (ídem, RA), que es una de las variables explicativa de la primera (segunda) ecuación, estará correlacionada con el término de perturbación U_1 (U_2) como consecuencia de los efectos recíprocos entre RA y AE (nótese que U_1 (U_2) interviene, a través de RA, en la «generación» de los valores de AE (RA)).

La correlación entre variables explicativas y términos de perturbación aparecerá siempre que una variable tenga indirectamente (a través de otras variables) un «efecto» sobre sí misma o, simplemente, cuando los términos de perturbación de las diferentes ecuaciones del modelo estén correlacionados. Dicha problemática conduce a tener en cuenta la simultaneidad de las diferentes ecuaciones propuestas en el modelo (i), por tanto, a introducir el modelo típicamente econométrico de ecuaciones estructurales lineales o, denominado también, *sistema de ecuaciones simultáneas* (interdependientes)

En un sistema de ecuaciones simultáneas puede distinguirse una serie de G variables $Y_1, \ldots, Y_2, \ldots, Y_G$, denominadas endógenas, la variación de las cuales viene explicada por las ecuaciones del modelo y otra serie de K variables $X_1, \ldots, X_K, \ldots, X_K$, denominadas variables *exógenas,* cuya variación no viene explicada por el modelo. Corresponde a cada una de las variables endógenas una ecuación de regresión lineal en la que como variables explicativas pueden figurar tanto variables exógenas como otras variables endógenas del modelo; es decir, un sistema de ecuaciones simultáneas viene caracterizado por un sistema de ecuaciones del siguiente tipo:

$$\left.\begin{aligned} Y_1 &= 0 + \beta_{12}Y_2 + \ldots + \quad \ldots + \beta_{1G}Y_G + \gamma_{11}X_1 + \ldots + \gamma_{1k}X_K + U_1 \\ Y_2 &= \beta_{21}Y_1 + 0 + \beta_{23}Y_3 + \ldots \quad + \beta_{2G}Y_G + \gamma_{2K}X_1 + \ldots + \gamma_{2k}X_K + U_2 \\ Y_G &= \beta_{G1}Y + \ldots \quad \beta_{GG-1}Y_{G-1} + 0 + \gamma_{G1}X_1 + \ldots + \gamma_{Gk}X_K + U_G \cdot \end{aligned}\right\} [5]$$

en el que algunos de los coeficientes (betas y gammas) a priori serán igual a cero o sometidos a restricciones específicas. Con respecto a los términos de perturbación U_g ($g = 1, 2, \ldots, G$) de las distintas ecuaciones supondremos que son también de naturaleza estocástica como en [4]. Si U_{gi} representa la perturbación (variable aleatoria) asociada a la ecuación g-ésima del modelo e individuo i-éstimo de la muestra, supondremos que:

(i) La distribución de probabilidad de U_{gi} ($g = 1, 2, \ldots, G$ e $i = 1, 2, \ldots, N$) es independiente de los valores que toman las variables exógenas $X_1, X_2, \ldots, X_k$ *del modelo*[8].

(*ii*) Para toda ecuación 9 del modelo ($g = 1, 2, \ldots, G$), $U_{g1}, U_{g2}, \ldots, U_{gN}$ son variables aleatorias estocásticamente independientes entre sí.

---

[8] Precisamente la «exogeneidad» de una variable queda caracterizada por el hecho de que la distribución de probabilidad de los términos de perturbación es independiente de los valores que toma dicha variable.

(iii) Para cada individuo i de la muestra ($i = 1, 2, \ldots, N$), la distribución de probabilidad conjunta de U_{1i}, U_{2i}, ..., U_{Gi} es normal multivariante con vector de medias igual a cero. La matriz de varianzas y covarianzas de dicha distribución vendrá denotada por ψ (psi).

Utilizando notación matricial podemos escribir [5] de la siguiente forma:

$$Y = BY + \Gamma X + U \qquad [6]$$

en donde, obviamente, Y, X y U son los vectores de variables siguientes:

$$Y' = (Y_1, Y_2, \ldots, Y_G)$$

$$X' = (X_1, X_2, \ldots, X_K)$$

$$U' = (U_1, U_2, \ldots, U_G)$$

Y B y Γ las matrices:

$$B = \begin{bmatrix} o & \beta_{12} & \ldots & \beta_{1G} \\ \beta_{21} & o & \ldots & \beta_{2G} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ \beta_{G1} & \beta_{G2} & \beta_{GG1} & o \end{bmatrix}$$

$$\Gamma = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \ldots & \gamma_{1K} \\ . & & & . \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ \gamma_{G1} & \gamma_{G2} & \ldots & \gamma_{GK} \end{bmatrix}$$

Si I representa la matriz identidad de dimensiones $G \times G$, supondremos (sin que ello comporte pérdida de generalidad) que la matriz $(I - B)$ es una matriz no singular. La expresión [5] ó [6] se denomina la *forma estructural* del modelo, siendo los elementos de las matrices B, Γ y Ψ con valor desconocido a priori los denominados *parámetros (estructurales)* del modelo. El conjunto de dichos parámetros vendrá denotado por θ y los valores «verdaderos» (poblacionales) de los mismos por θ^o. En general, será un objetivo fundamental del investigador la obtención de información relativa a

θ^o. Dicha información se «extraerá» de la «muestra» utilizando técnicas estadísticas apropiadas.

Dado que la matriz $(I\text{-}B)$ se ha supuesto no singular, una expresión del modelo equivalente a [6] es la siguiente:

$$Y = \Pi x + V \qquad [7]$$

en donde:

$$\Pi = (I - B)^{-1}\Gamma \qquad [8]$$

y

$$V = (I - B)^{-1}U \qquad [9]$$

Los supuestos estocásticos (i) a (iii) relativos a U y la igualdad [9] implicarán que el vector aleatorio V tenga distribución normal multivariante con vector de medias igual a cero y matriz de varianzas y covarianzas Ω dada por:

$$\Omega = (I - B)^{-1}\Psi(I - B)'^{-1} \qquad [10]$$

La expresión [7] se denomina *forma reducida* del modelo y especifica la distribución del vector de variables endógenas Y condicionada al vector de variables exógenas X. A los $(G \times K + G(G + 1)/2)$ coeficientes de las matrices Π y Ω, que caracterizan dicha distribución condicional de Y respecto X, los denominaremos *parámetros estadísticos* del modelo. El conjunto de dichos parámetros vendrá denotado por Π y los valores «verdaderos» de los mismos por δ^o.

Una clase importante de modelos multiecuacionales son los denominados modelos recursivos, que vienen caracterizados por las dos condiciones siguientes:

a) Existe una ordenación de las ecuaciones del modelo en la que la matriz B correspondiente es triangular.

b) La matriz Ψ de varianzas y covarianzas de los términos de perturbación es diagonal. Es decir, los términos de perturbación de las diferentes ecuaciones del modelo están incorrelacionados entre sí.

El aspecto más relevante del carácter recursivo de un modelo es que la problemática propia de la interdependencia entre ecuaciones desaparece; de tal manera que los MCO aplicados independientemente en cada ecuación conducen a estimaciones consistentes y «óptimas» de los parámetros del modelo. Efectivamente, se demuestra fácilmente que los supuestos de recursividad [(a) y (b)] imposibilitan la correlación entre variables endógenas explicativas y los términos de perturbación correspondientes.

Los modelos recursivos han tenido un papel importante en el desarrollo de la metodología de las ecuaciones estructurales. Algunos autores han argumentado afirmando el carácter intrínsecamente recursivo de todo modelo de causalidad (véase Wold, 1964). Cabe decir también que las técnicas del *path analysis* (Wright, 1934) y, en par-

ticular, las técnicas de la correlación parcial (Simon, 1957 y Blalock, 1964), que contribuyeron de forma importante a introducir los modelos de causalidad en el ámbito de la Sociología, son principalmente técnicas de modelos recursivos.

9.4. Identificación

Un problema típico de los modelos de ecuaciones estructurales no recursivos, que es previo a todo proceso de estimación y contraste, es el denominado *problema de la identificación*. Dicho problema plantea la cuestión fundamental de si mediante la «información muestral» es posible obtener información sobre los valores «verdaderos» de los parámetros estructurales.

La forma reducida [7] del modelo, presentada en el apartado anterior, implica que la distribución del vector de variables endógenas Y condicionada al vector de variables exógenas X es normal multivariante con vector de medias igual a ΠX y matriz de varianzas y covarianzas Ω; de manera que los parámetros estadísticos Π (es decir, el conjunto de elementos de las matrices Π y Ω) caracterizan dicha distribución condicional. Pues bien, el modelo estadístico que tenemos planteado es de tal naturaleza que la «información muestral» posibilita la estimación consistente de los parámetros estadísticos del modelo; o sea, aumentando suficientemente el tamaño de la muestra podemos aproximar los elementos de Π y Ω con la exactitud deseada. El problema de la identificación consiste en averiguar si la determinación de Π y Ω implica también la determinación de θ; solamente si ello es así, la «información muestral» posibilitará también la estimación consistente de los parámetros estructurales del modelo.

Fijada la información «a priori» que hace referencia a los valores específicos de algunos elementos de las matrices B, Γ y Ψ, se establece a través de [8] y [10] una correspondencia entre θ (parámetros estructrurales, o sea, valores concretos de los elementos «libres» de las matrices, B, Γ y Ψ) y δ (parámetros estadísticos, o elementos de las matrices Π y Ω); es decir [8] y [10] determinan una función específica

$$\delta = \delta(\theta)$$

Un modelo diremos que está *identificado* si no existen dos conjuntos θ' y θ'' de parámetros estructurales a los que les corresponde el mismo conjunto δ de parámetros estadísticos; o sea, el modelo estará identificado si:

$$\delta(\theta) = \delta(\theta') \Rightarrow \theta' = \theta''$$

Diremos que el modelo está *localmente identificado* en θ^o si

$$\delta(\theta) = \delta(\theta') \Rightarrow \theta' = \theta^o$$

para cualquier θ' en un entorno de θ^o.

En el mismo sentido, diremos que un parámetro estructural del modelo está identificado si no hay dos valores distintos de dicho parámetro compatibles con un mismo conjunto δ de parámetros estadísticos. Un grupo de parámetros estructurales diremos

que está identificado si cada uno de los parámetros del grupo está identificado. En particular, una ecuación del modelo diremos que está identificada si cada parámetro de la ecuación está identificado. Obviamente, la estimación de un parámetro del modelo tendrá sentido solamente en el caso de que dicho parámetro esté identificado (al menos localmente).

En general, el problema de la identificación de un modelo se resolverá considerando las ecuaciones [8] y [10] e investigando si es posible la obtención de los elementos libres de las matrices B, Γ y Ψ (o sea, los elementos de θ) en función de los elementos de las matrices Π y Ω (elementos de Π). Si los elementos de Π determinan a los elementos de θ, entonces el modelo está identificado. Dado que el número de elementos de Π es igual a $(G \times K + G(G + 1)/2)$, que es el número de ecuaciones implícitas en [8] y [10], si el número de parámetros estructurales funcionalmente independientes (elementos de θ) es r, una condición necesaria para que el modelo esté identificado es la siguiente:

$$d = \{(G \times K + G(G + 1)/2) - r\} \geq 0 \qquad [11]$$

Si el modelo está identificado y d es igual (mayor) que cero, entonces diremos que el modelo está *exactamente (sobre) identificado.*

Cuando el modelo no impone «a priori» restricciones sobre la matriz Ψ de varianzas y covarianzas de los términos de perturbación y las restricciones sobre B y Γ son de «exclusión» (elementos de B y Γ iguales a cero), las condiciones típicas que determinan el estado de identificación de cada ecuación del modelo son las siguientes:

Condición de orden: una condición *necesaria* para que una ecuación concreta del modelo esté identificada es que el número de variables excluidas de la ecuación sea igual o mayor que el número de ecuaciones del modelo menos una.

Condición de rango: Denotando por A la matriz particionada siguiente:

$$A = [I - B \vdots \Gamma] \; . \qquad [12]$$

una condición *necesaria y suficiente* para que una ecuación concreta del modelo esté identificada es que la matriz formada mediante las columnas de A que tienen valor cero en la fila de la ecuación considerada (o sea, aquellas columnas que corresponden a variables excluidas de la ecuación considerada) sea de rango $G - 1$ (es decir, que pueda extraerse de dicha matriz un determinante de orden $G - 1$ distinto de cero).

Si una ecuación del modelo está identificada y verifica además la condición de orden con igualdad (desigualdad) estricta, entonces diremos que dicha ecuación está exactamente(sobre) identificada.

Como ilustración de dichas condiciones de rango y orden, consideremos el modelo especificado en la figura 2 anterior. En dicho modelo, es fácil verificar (véase apartado 9.7 de este capítulo) que las dos ecuaciones verifican las condiciones de orden con igualdad estricta. Puesto que además cada ecuación verifica la condición de rango (con tal de que γ_{11} y γ_{22} sean distintos de cero) resulta que ambas ecuaciones del modelo están exactamente identificadas. En cambio, si el modelo especificase adicionalmente un «efecto» directo de NEP sobre RA, entonces el modelo dejaría de estar identificado.

Es importante que el lector tenga en cuenta que las condiciones de orden y rango anunciadas anteriormente son relevantes solamente en el caso de que no haya restricciones sobre Ψ, y que las restricciones sobre B y Γ sean de «exclusión». A veces se imponen «a priori» restricciones sobre Ψ resultando que modelos que no verifican las condiciones de orden y rango enunciadas anteriormente pueden estar identificados. Un ejemplo típico de dicha situación nos lo ofrecen los modelos recursivos, para los que se puede probar que están siempre identificados (que los MCO proporcionen estimaciones consistentes de los parámetros estructurales es ya una prueba de ello), independientemente de que las diferentes ecuaciones del modelo verifiquen o no las condiciones de orden y rango.

Para un tratamiento más detallado del problema de la identificación, el lector interesado puede consultar los manuales de econometría anteriormente citados. Para un tratamiento exhaustivo del problema de la identificación del modelo de ecuaciones estructurales introducido en este capítulo véase Fisher (1966).

9.5. Estimación

Cuando el modelo está identificado, y la especificación del mismo es correcta, existe una gran variedad de métodos de estimación que conduce a estimaciones consistentes de los parámetros estructurales. Puede estimarse cada ecuación del modelo separadamente de las demás mediante uno de los denominados métodos de ecuación singular o *información limitada,* o bien puede utilizarse uno de los métodos de estimación de sistema completo, o *información completa,* que estiman conjuntamente todos los parámetros de las ecuaciones del modelo.

Los métodos de estimación de ecuación singular, no tienen en cuenta la intercorrelación entre los términos de perturbación de las diferentes ecuaciones del modelo y pueden, por tanto, resultar menos eficientes que los métodos de estimación de sistema completo (que si incorporan en el proceso de estimación la información relativa a la intercorrelación entre los términos de perturbación). Sin embargo, los métodos de ecuación singular pueden resultar más «robustos» frente a los errores de especificación que los métodos de sistema completo ya que, al tratar cada ecuación por separado de las demás, los errores de especificación de una ecuación «perturban» solamente las estimaciones de los parámetros de esta ecuación.

El método de estimación de ecuación singular más común es el denominado de los mínimos cuadrados bietápicos (MC2E). La naturaleza de dicho método de estimación es simple: si los MCO aplicados a una ecuación del modelo resultan inconsistentes debido a la correlación del término de perturbación con variables endógenas explicativas de la ecuación, los MC2E sustituyen en la estimación MCO dichas endógenas explicativas por «instrumentos» incorrelacionados con el término de perturbación. El «instrumento» de una variable será la combinación lineal de variables exógenas del modelo que más se correlaciona con ella. La obtención de los «instrumentos», por medio de la regresión MCO de cada variable endógena explicativa sobre todas las variables exógenas del modelo, constituye la primera etapa de los MC2E; mientras que, la estimación MCO de la ecuación del modelo con «instrumentos» en lugar de endógenas explicativas constituye la segunda etapa de los MC2E.

A fin de facilitar la exposición de los método de estimación de sistema completo más habituales, introduciremos el vector Z de variables del modelo dado por:

$$Z' = (Y', X') ,$$

la matriz S de varianzas y covarianzas muestral de Z[9], particionada de la forma siguiente:

$$S = \left[\begin{array}{c|c} S_{yy} & S_{yx} \\ \hline S_{xy} & S_{xx} \end{array}\right]$$

y, finalmente, la matriz Σ, que denominaremos matriz de varianzas y covarianzas poblacional, dada por [10]:

$$\Sigma = \left[\begin{array}{c|c} \Pi S_{xx} \Pi' + \beta^{-1} \Psi \beta'^{-1} & \Pi S_{xx} \\ \hline S_{xx} \Pi' & S_{xx} \end{array}\right]$$

Puesto que la inferencia estadística se efectúa condicionada a los valores observados de las variables exógenas, la matriz S_{xx} puede considerarse fija y conocida, de manera que la especificación del modelo induce una función específica $\Sigma = \Sigma(\theta)$ que hace corresponder a cada θ una matriz Σ de varianzas y covarianzas poblaciones.

Diferentes métodos de estimación de sistema completo, entre ellos el que corresponde a la aplicación del principio de la máxima verosimilitud, se derivan de la minimización de una *función criterio* $F = F(S, \Sigma)$ o medida del ajuste («discrepancia») entre S y Σ. Puesto que Σ es función de θ, fijada la muestra (y, por tanto, fijada la matriz S) se obtiene una función específica $F(\theta) = F(S, \Sigma(\theta))$ de θ. Definiremos la estimación de θ asociada a la función criterio F como aquel valor $\hat{\theta}$ de los parámetros del modelo que satisface:

$$F(\hat{\theta}) \leq F(\theta) \quad , \quad \forall \theta \in \Theta$$

en donde Θ denota el espacio de parámetros (valores posibles de θ) asociado a la especificación del modelo. Obviamente, las cualidades de $\hat{\theta}$ como estimación de θ dependerán de la elección hecha de la función criterio F.

---

[9] Si $z_i (i = 1, 2, \ldots, N)$ representa la observación del vector Z correspondiente al individuo i-ésimo muestral, la matriz S vendrá definida por:

$$S = 1/(N-1) \sum_{i=1}^{N} (z_i - \bar{z})(z_i - \bar{z})'$$

en donde

$$\bar{z} = 1/N \sum_{i=1}^{N} z_i$$

[10] Véase, por ejemplo, fórmula [4] en Joreskog (1973).

Expresiones típicas de F con las siguientes:

(i) $$F_{\text{ULS}}(S, \Sigma) = 1/2Tr(S - \Sigma)^2 ,$$

que conduce a los denominados estimadores de mínimos cuadrados no ponderados (ULS: «Unweigthed Least Squeares»),

(ii) $$F_{\text{GLS}}(S, \Sigma) = 1/2Tr[(S - \Sigma)S^{-1}]^2$$

que corresponde a los denominados estimadores de mínimos cuadrados generalizados (*GLS*: «Generalized Least Squares»), y, finalmente,

(iii) $$F_{\text{ML}}(S, \Sigma) = Tr(\Sigma^{-1}S) - \log|\Sigma^{-1}S| - (G + K)$$

que es la función criterio correspondiente a las estimaciones *máximo verosímiles* (con información completa) (ML: «Maximum Likelihood»). Los estimadores ULS y GLS vienen tratados, por ejemplo, en Browne (1977) mientras que la función criterio de ML viene propuesta en Joreskog (1973). Puede demostrarse que tanto los estimadores ML como los ULS y GLS son estimadores consistentes. Sin embargo, una cualidad fundamental de los estimadores ML es la de que, son eficientes («óptimos») y con distribución normal. Un aspecto interesante de la utilización de F_{ULS} es que, a diferencia de las otras funciones criterio, no requiere que S sea una matriz no singular.

La minimización de las funciones criterio F_{ULS}, F_{GLS} y F_{ML} se efectuará mediante procedimientos iterativos de optimización numérica. En general, partiendo de una estimación inicial $\theta^{(1)}$ se generará una secuencia $\theta^{(1)}$, $\theta^{(2)}$, $\theta^{(3)}$, $\theta^{(4)}$... de valores de los parámetros tal que $F(\theta^{(1)}) < F(\theta^{(t-1)})$ hasta conseguir la convergencia. El método numérico de minimización utilizará el vector de derivadas primeras de la función $F(\theta)$ así como una aproximación de la matriz de derivadas segundas de dicha función. En el caso de la estimación ML la aproximación utilizada de la matriz de derivadas segundas es la matriz de información que servirá también para calcular los errores estándars de las estimaciones ML (para la estimación de la matriz de varianzas y covarianzas de los estimadores ULS y GLS véase Browne, 1982).

Un programa de ordenador que operativiza los métodos de estimación MC2E, ULS, GLS y ML mencionados en este apartado, en el marco de un modelo más general incluso que el de ecuaciones simultáneas presentado en este capítulo, es el programa LISREL de Joreskog y Sorbom (1984).

9.6. Contraste Gi-cuadrado de la bondad del ajuste

Cuando el modelo está sobreidentificado, es decir, cuando el valor d dado por [11] es mayor que 0, la matriz Σ de varianzas y covarianzas poblaciones está sometida a restricciones (de sobreidentificación) que deben verificarse si la especificación del modelo es correcta («verdadera»). Si la estimación se efectúa utilizando el método de la máxima verosimilitud (de sistema completo) considerado en el apartado anterior,

dichas restricciones de sobreidentificación sobre Σ pueden contrastarse estadísticamente, de manera que la especificación del modelo (que implica dichas restricciones) puede ser falseada. El contraste de estas restricciones de sobreidentificación equivaldrá a un contraste de la «bondad del ajuste» del modelo especificado.

El estadístico de contraste utilizado es un estadístico de contraste de la *razón de verosimilitud,* cuya hipótesis nula sostiene que Σ satisface las restricciones impuestas por la especificación del modelo, mientras que la hipótesis mantenida sostiene, simplemente, que Σ es una matriz definida positiva. El valor del estadístico de contraste se obtiene multiplicando por $(N - 1)$ el valor mínimo de la función criterio $F_{ML}(\theta)$. Si la hipótesis nula es cierta (o sea, si la especificación del modelo es «verdadera») la distribución asimptótica (para muestra grande) de dicho estadístico de contraste en Gi-cuadrado con grados de libertad igual al número d (dado por [11]) de restricciones de sobreidentificación sobre Σ impuestas por la especificación del modelo. La problemática de la utilización de dicho estadístico de contraste de la razón de verosimilitud en el contexto de los modelos de ecuaciones estructurales que ahora nos ocupa viene desarrollada en Satorra (1983). A base de considerar la diferencia de los estadísticos Gi-cuadrado de la bondad del ajuste correspondientes a dos modelos anidados, puede contrastarse una gran variedad de hipótesis estructurales relativas al modelo especificado.

9.7. Ilustración

Ilustraremos la metodología estadística propuesta en este capítulo mediante dos ejemplos. En el primero, que tendrá carácter experimental, analizaremos una muestra de datos «artificiales» generada asignando valores concretos a los parámetros del modelo educacional considerado anteriormente en la figura 2. En el segundo ejemplo, situado en el contexto de una investigación empírica actualmente en curso, se efectuará la estimación y contraste, utilizando datos empíricos, de un modelo de fertilidad o de *cohort analysis.*

Ilustración con muestra arficial

Un estudio en el ámbito educacional que intente evaluar la incidencia recíproca que tienen entre sí variables como las «aspiraciones educacionales» (AE) y el «rendimiento académico» (RA) referidas a los individuos de una determinada población estudiantil (por ejemplo, la población de alumnos de 3.er curso de B.U.P.), puede conducir a la especificación de un modelo como el presentado anteriormente en la figura 2[11]. Considerando valores concretos para los parámetros de dicho modelo, concretamente los valores presentados en la figura 3, se ha «simulado» una muestra de $N = 200$ observaciones de las variables del modelo RA, AE, I y NEP. Para la generación de dichas observaciones, las realizaciones de U_1 y U_2 se han obtenido de distribuciones normales (como corresponde a los supuestos estadísticos habituales) in-

[11] Por ejemplo, dicho modelo es una versión simplificada de Land (1971).

dependientes, mientras que las realizaciones de las variables exógenas I y NEP se han obtenido de sendas distribuciones uniformes (siendo, ahora, esta elección distribucional completamente arbitraria). La matriz S de varianzas y covarianzas muestra muestral calculada a partir de la muestra «artificial» resultante se presenta en la tabla 1.

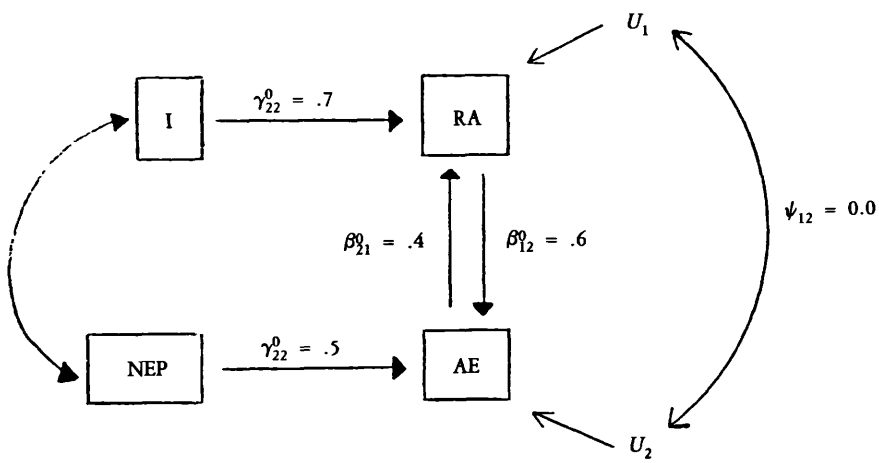


FIGURA 3. Diagrama *path* del modelo educacional y valores «verdaderos» de los parámetros utilizados en la generación de la muestra artificial.

TABLA 1. Matriz de varianzas y covarianzas muestral correspondiente a la muestra «artificial».

	RA	AE	I	NEP
RA	6.461			
AE	6.543	9.006		
I	2.445	1.019	2.799	
NEP	1.060	2.170	-0.179	3.283

En base a la «información muestral» contenida en dicha matriz S de varianzas y covarianzas, y utilizando la metodología propuesta en este capítulo, intentaremos reproducir las características «poblacionales» (valores «verdaderos») de los parámetros del modelo.

Como etapa previa a cualquier análisis estadístico inferencial, consideraremos la expresión analítica del modelo e investigaremos el estado de identificación del mismo.

Las ecuaciones del modelo, que corresponden al diagrama *path* presentado en la figura 2, son las siguientes:

$$\begin{aligned} RA &= \beta_{12}AE + \gamma_{11}I + U_1 \\ AE &= \beta_{21}RA + \gamma_{22}NEP + U_2 \end{aligned} \qquad [13]$$

y la matrices, B, Γ y Ψ de coeficientes del modelo, bajo las correspondientes restricciones a priori, serán:

$$B = \begin{bmatrix} 0 & \beta_{12} \\ \beta_{21} & 0 \end{bmatrix} \qquad \Gamma = \begin{bmatrix} \gamma_{11} & 0 \\ 0 & \gamma_{22} \end{bmatrix}$$

$$\text{y} \qquad \Psi = \begin{bmatrix} \Psi_{11} & \Psi_{12} \\ \Psi_{21} & \Psi_{22} \end{bmatrix}$$

en donde $\Psi_{12} = \Psi_{21}$. Por tanto, el conjunto θ de parámetros del modelo, o coeficientes «libres» de las matrices B, Γ y Ψ, será el vector de parámetros siguiente:

$$\theta = (\beta_{12}, \beta_{21}, \gamma_{11}, \gamma_{22}, \Psi_{11}, \Psi_{12}, \Psi_{22})$$

Dado que el modelo no impone a priori restricciones sobre la matriz Ψ de varianzas y covarianzas de los términos de perturbación del modelo, y que las restricciones sobre B, Γ y Ψ son de «exclusión», el problema de la identificación del modelo se resolverá averiguando si cada ecuación del modelo satisface la condición de rango (y la de orden). La matriz A correspondiente en este caso será la siguiente:

$$A = \left[\begin{array}{cc|cc} 1 & -\beta_{12} & \gamma_{11} & 0 \\ -\beta_{21} & 1 & 0 & \gamma_{22} \end{array}\right]$$

Puesto que el rango de la matriz $[0 \quad \gamma_{22}]$ es igual a 1 (con tal de que $\gamma_{22} \neq 0$), la primera ecuación del modelo está identificada y, además, dado que la condición de orden se verifica con igualdad estricta, dicha ecuación estará exactamente identificada. De la misma manera concluiríamos que la segunda ecuación está también exactamente identificada. Por tanto, el modelo está exactamente identificado y tiene sentido el intentar «recuperar» los valores «verdaderos» de los parámetros a partir de la información muestral».

Si el investigador ignorase la problemática de la interdependencia entre las ecuaciones planteadas en [12] estimaría, seguramente, los parámetros del modelo aplicando MCO por separado en cada ecuación. Dicho procedimiento de estimación, aplicado a la muestra «artificial» mencionada anteriormente, conduce a los resultados que presentamos en la tabla 2.

Puede observarse en dicha tabla 2 que las estimaciones MCO de los parámetros estructurales del modelo obtenidas difieren sustancialmente de los valores «verdaderos» de los mismos. Se observa, por ejemplo, que la estimación correspondiente al «efecto directo» de RA (Y_1) sobre AE (Y_2) es .95, un valor mayor que la estimación correspondiente al «efecto directo» de AE sobre RA (.65), contrariamente a lo que puede observarse en la figura 2 entre los valores verdaderos de dichos «efectos».

En cambio, si aplicamos uno cualquiera de los métodos de estimación de ecuaciones simultáneas, propuestos en este capítulo, observaremos resultados mucho más sa-

TABLA 2. Resultados de la estimación MCO de las dos ecuaciones del modelo educacional (*output* de BMDP).

```
 REGRESSION TITLE IS
ESTIMACION MCO DE LA PRIMERA ECUACION : β̂12 , γ̂11

 DEPENDENT VARIABLE. . . . . 1 Y1

MULTIPLE R              0.9503          STD. ERROR OF EST.          0.7931
MULTIPLE R-SQUARE       0.9031

                                                  STD. REG
   VARIABLE        COEFFICIENT   STD. ERROR        COEFF          T    P(2 TAIL)

Y2          2        0.65462 β̂12  0.01913          0.773      34.212      0.0
X1          3       .0.63501 γ̂11  0.03432          0.418      18.502      0.0

 REGRESSION TITLE IS
ESTIMACION MCO DE LA SEGUNDA ECUACION : β̂21 , γ̂22

 DEPENDENT VARIABLE. . . . . 2 Y2
 TOLERANCE . . . . . 0.0100

MULTIPLE R              0.8824          STD. ERROR OF EST.          1.4155
MULTIPLE R-SQUARE       0.7786

                                                  STD. REG
   VARIABLE        COEFFICIENT   STD. ERROR        COEFF          T    P(2 TAIL)
                        β̂21
Y1          1        0.95470 γ̂22  0.04056          0.809      23.537      0.0
X2          4        0.53291      0.05690          0.213       6.202      0.0000
```

tisfactorios. La tabla 3 presenta las estimaciones MC2E y ML de los parámetros estructurales del modelo. Puede observarse que en este caso los dos métodos de estimación conducen a estimaciones idénticas, siendo ello debido a que las dos ecuaciones del modelo están exactamente identificadas. Cabe señalar que los errores estándars presentados en la tabla 3 corresponden solamente a las estimaciones ML y que (como es habitual en el programa LISREL) las estimaciones MC2E han sido utilizadas como valores iniciales del proceso iterativo que conduce a las estimaciones ML.

Por otra parte, la tabla 4 muestra la ejecución de las dos etapas de los MC2E. En dicha tabla pueden observarse los *outputs* de BMDP correspondientes a la estimación MCO de las regresiones siguientes:

Primera etapa MC2E :	Regresión de Y_1 sobre X_1 y X_2
	Regresión de Y_2 sobre X_1 y X_2
Segunda etapa MC2E:	Regresión de Y_1 sobre $\hat{Y}_2$ y X_1
	Regresión de Y_2 sobre $\hat{Y}_1$ y X_2

siendo respectivamente $\hat{Y}_1(IY_1$, en el *output*) y $\hat{Y}_1(IY_2)$ los ajustes de Y_1 e Y_2 obtenidas en la primera etapa de los MC2E. Conviene observar, no obstante, que los errores estándars de las estimaciones MCO de la segunda etapa no coinciden con los errores estándars MC2E.

TABLA 3. Estimaciones MC2E y ML de los parámetros del modelo educacional obtenidas mediante LISREL. Nótese que las dos estimaciones coinciden (y coinciden también con las estimaciones ULS y GLS).

A) ESTIMACIONES MC2E:

BETA

	RA	AE
RA	0.0	0.544 $\hat{\beta}_{12}$
AE	0.454 $\hat{\beta}_{21}$	0.0

GAMMA

	I	NEP
RA	0.675 $\hat{\gamma}_{11}$	0.0
AE	0.0	0.514 $\hat{\gamma}_{22}$

PSI

	RA	AE
RA	0.730	
AE	0.386	3.525

B) ESTIMACIONES ML

BETA

	RA	AE
RA	0.0	0.544 $\hat{\beta}_{12}$
AE	0.454 $\hat{\beta}_{21}$	0.0

GAMMA

	I	NEP
RA	0.675 $\hat{\gamma}_{11}$	0.0
AE	0.0	0.514 $\hat{\gamma}_{22}$

PSI

	RA	AE
RA	0.730	
AE	0.386	3.525

MEASURES OF GOODNESS OF FIT FOR THE WHOLE MODEL :

CHI-SQUARE WITH 0 DEGREES OF FREEDOM IS 0.00 (PROB. LEVEL = 1.000)

C) ERRORES ESTANDARS DE LAS ESTIMACIONES ML

STANDARD ERRORS

BETA

	RA	AE
RA	0.0	0.049
AE	0.089	0.0

GAMMA

	I	NEP
RA	0.041	0.0
AE	0.0	0.079

TABLA 4. Resultados de las estimaciones MCO correspondientes a la ejecución de las dos etapas MC2E en la estimación del modelo educacional (*output* de BMD).

REGRESSION TITLE IS
PRIMERA ETAPA MC2E: OBTENCION INSTRUMENTO (IY1) DE Y1

DEPENDENT VARIABLE. 1 Y1

MULTIPLE R	0.6328	STD. ERROR OF EST.	1.9732
MULTIPLE R-SQUARE	0.4004		

VARIABLE		COEFFICIENT	STD. ERROR	COEFF	T	P(2 TAIL)
X1	3	-0.89719	0.08375	0.591	10.713	0.0
X2	4	0.37173	0.07733	0.265	4.807	0.0000

REGRESSION TITLE IS
PRIMERA ETAPA MC2E:OBTENCION INSTRUMENTO (IY2) DE Y2

DEPENDENT VARIABLE. 2 Y2

MULTIPLE R	0.4592	STD. ERROR OF EST.	2.6726
MULTIPLE R-SQUARE	0.2109		

VARIABLE		COEFFICIENT	STD. ERROR	COEFF	T	P(2 TAIL)
X1	3	0.40790	0.11344	0.227	3.596	0.0004
X2	4	0.68333	0.10474	0.413	6.524	0.0000

REGRESSION TITLE IS
SEGUNDA ETAPA MC2E PARA LA PRIMERA ECUACION

DEPENDENT VARIABLE. 1 Y1

MULTIPLE R	0.6328	STD. ERROR OF EST.	1.9732
MULTIPLE R-SQUARE	0.4004		

VARIABLE		COEFFICIENT	STD. ERROR	COEFF	T	P(2 TAIL)
X1	3	0.67530 $\hat{\gamma}_{11}$	0.09321	0.444	7.245	0.0000
IY2	6	0.54397 $\hat{\beta}_{12}$	0.11317	0.295	4.807	0.0000

REGRESSION TITLE IS
SEGUNDA ETAPA MC2E PARA LA SEGUNDA ECUACION

DEPENDENT VARIABLE. 2 Y2

MULTIPLE R	0.4592	STD. ERROR OF EST.	2.6726
MULTIPLE R-SQUARE	0.2109		

VARIABLE		COEFFICIENT	STD. ERROR	COEFF	T	P(2 TAIL)
X2	4	0.51432 $\hat{\gamma}_{22}$	0.11224	0.311	4.582	0.0000
IY1	5	0.45464 $\hat{\beta}_{21}$	0.12644	0.244	3.596	0.0004

Puesto que el modelo está exactamente identificado (nótese que, atendiendo a [11], $d = ((2 \times 2 + 1/2(3 \times 2)) - 7 = 0)$ no existen restricciones de sobreidentificación para contrastar. Ello se refleja en el hecho que, en la estimación ML, el estadístico Gi-cuadrado de la bondad del ajuste tiene 0 grados de libertad y un valor igual a cero.

En el supuesto que el investigador especificase incorrelación entre los términos de perturbación U_1 y U_2 (especificación que, en nuestro ejemplo, sabemos que es cierta) entonces el modelo especificado sería sobreidentificado con una restricción de sobreidentificación. La estimación ML de dicho modelo sobreidentificado se presenta en la tabla 5.

TABLA 5. Estimaciones ML de los parámetros del modelo educacional sobreidentificado (matriz PSI diagonal).

ESTIMACIONES ML

MODELO DE ASPIRACIONES EDUCACIONALES: PSI DIAGONAL

LISREL ESTIMATES (MAXIMUM LIKELIHOOD)

BETA

	RA	AE
RA	0.0	0.602 $\hat{\beta}_{12}$
AE	0.494 $\hat{\beta}_{21}$	0.0

GAMMA

	I	NEP
RA	0.654 $\hat{\gamma}_{11}$	0.0 $\hat{\gamma}_{12}$
AE	0.0	0.502 $\hat{\gamma}_{22}$

PSI

	RA	AE
RA	0.649	
AE	0.0	3.293

MEASURES OF GOODNESS OF FIT FOR THE WHOLE MODEL :

CHI-SQUARE WITH 1 DEGREES OF FREEDOM IS 2.16 (PROB. LEVE

En dicha tabla se observa también el valor del estadístico Gi-cuadrado de la «bondad del ajuste» que, para un grado de libertad, es igual a 2.16. El nivel de probabilidad correspondiente es .141, valor que es superior a los niveles de significación habituales y, por tanto, cabe «aceptar» la restricción de sobreidentificación (o el modelo sobreidentificado). Por otra parte, cabe señalar que la incorporación en la especificación del modelo de una restricción «cierta» conduce a que la estimación de los parámetros restantes sea más eficiente. Ello queda reflejado en el descenso de la magnitud del error estándar cuando, para un mismo parámetro, pasamos de la tabla 3 a la tabla 5.

Cuando en el modelo presentado en la figura 2 especificamos incorrelación entre los términos de perturbación (matriz PSI diagonal) e incluimos un «efecto directo» de NEP sobre RA se obtiene un modelo exactamente identificado. Ello es consecuencia de que las ecuaciones [8] y [10] permiten la obtención de los parámetros «libres» del modelo en función de los elementos de Π, y de que el valor de d, en dicho caso, es igual a 0. La estimación ML de dicho modelo viene presentada en la tabla 6.

TABLA 6. Estimación ML y valores «t» del modelo educacional con PSI diagonal e inclusión de un efecto directo de NEP sobre RA.

A) ESTIMACIONES ML

MODELO DE ASPIRACIONES EDUCACIONALES: PSI DIAGONAL Y GA(1,2) NO CERO

LISREL ESTIMATES (MAXIMUM LIKELIHOOD)

BETA

	RA	AE
RA	0.0	0.623 $\hat{\beta}_{12}$
AE	0.454 $\hat{\beta}_{21}$	0.0

GAMMA

	I	NEP
RA	0.643 $\hat{\gamma}_{11}$	-0.054 $\hat{\gamma}_{12}$
AE	0.0	0.514 $\hat{\gamma}_{22}$

PSI

	RA	AE
RA	0.624	
AE	0.0	3.525

MEASURES OF GOODNESS OF FIT FOR THE WHOLE MODEL :

CHI-SQUARE WITH 0 DEGREES OF FREEDOM IS 0.00 (PROB. LEVEL = 1.000)

B) VALOR «t» PARA CADA PARAMETRO ESTIMACO (COCIENTE ENTRE LA ESTIMACION PUNTUAL Y EL ERROR ESTANDAR)

T-VALUES

BETA

	RA	AE
RA	0.0	23.345
AE	5.091	0.0

GAMMA

	I	NEP
RA	18.169	-1.487
AE	0.0	6.489

PSI

	RA	AE
RA	9.409	
AE	0.0	5.408

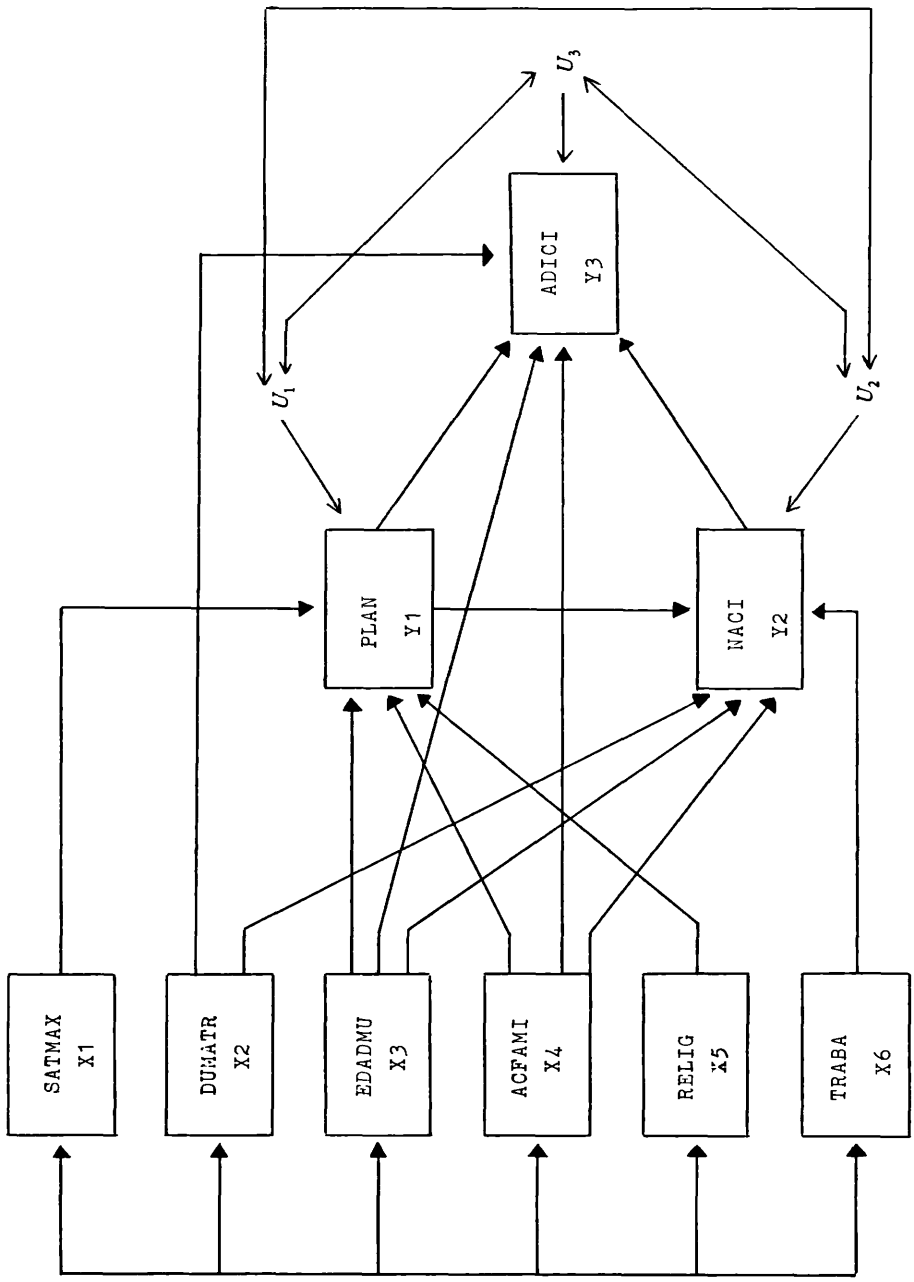


FIGURA 4. Diagrama *path* correspondientes al modelo de fertilidad.

Como era de esperar, a juzgar por el valor «t» correspondiente (cociente entre la estimación puntual y el error estándar), dicho «efecto» de NEP sobre RA no es significativo.

Ilustración con datos empíricos

Con objeto de investigar las «causas» del declive en fertilidad (número de hijos) de las parejas observado en Holanda en años recientes, el Netherland Central Bureau of Statistics realizó una encuesta en la que fueron entrevistadas 4.221 parejas (véase, *Netherland Survey on Fertility and Parenthood Motivation*, 1975). Con referencia a los datos de dicha encuesta, y dentro de una investigación actualmente en curso (véase Stronkhorst y Satorra, 1984), se consideran las siguientes variables:

PLAN: Número de hijos planeados inicialmente por la pareja.
NACI: Número de hijos nacidos hasta el momento de la entrevista.
ADICI: Número de hijos adicionales que la pareja espera tener.
SATMAX: Número ideal de hijos que conducirían a satisfacción máxima en la pareja (opinión subjetiva de la pareja).
DUMATR: Tiempo de duración del matrimonio hasta el momento de la entrevista.
EDAD: Edad de la mujer al inicio de la pareja.
ACFAMI: Indicador de actitud de la pareja ante el hecho familiar (valores bajos = actitud negativa, valores altos = actitud positiva).
RELIG: Frecuencia de asistencia de la pareja a actos religiosos.
TRAB: Número de meses que la mujer ha dedicado a un trabajo remunerado fuera del hogar.

En base a dichas variables se especifica un modelo de ecuaciones simultáneas en el que PLAN, NACI y ADICI son las variables endógenas y SATMAX, DUMATR, EDAD, RELIG y TRAB la variables exógenas. La consideración teórica de las interrelaciones entre variables conduce a la especificación (tentativa) presentada en el diagrama *path* de la figura 4.

Dicha especificación queda caracterizada también mediante la correspondiente matriz *A* presentada en la tabla 7.

TABLA 7. Matriz *A* correspondiente a la especificación del modelo de fertilidad considerado.

	Y_1	Y_2	Y_3	X_1	X_2	X_3	X_4	X_5	X_6
Y_1	1	0	0	γ_{11}	0	γ_{13}	γ_{14}	γ_{15}	0
Y_2	β_{21}	1	0	0	γ_{22}	γ_{23}	γ_{24}	0.	γ_{26}
Y_3	β_{31}	β_{32}	1	0	γ_{32}	γ_{33}	γ_{34}	0	0

Dado que el modelo no impone restricciones sobre la matriz Ψ de varianzas y covarianzas de los términos de perturbación de las ecuaciones, el problema de la identificación lo resolveremos investigando si se verifican o no las condiciones de rango y orden. De la matriz A presentada en la tabla 7 se desprende fácilmente que cada ecuación verifica la condición de rango y, por tanto, que el modelo está identificado. Por ejemplo, para la tercera ecuación, la comprobación de la condición de rango conduce a comprobar que de la matriz siguiente:

$$\begin{bmatrix} 0 & \gamma_{11} & \gamma_{15} \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

se puede extraer un determinante de orden 2 distinto de cero. Ello será así con tal de que uno de los coeficientes γ_{11} y γ_{15} sea distinto de cero.

Considerando una submuestra de 1.741[12] individuos (parejas) se ha obtenido la matriz de varianzas y covarianzas muestral que presentamos en la tabla 8[13, 14].

TABLA 8. Matriz de varianzas y covarianzas muestral de las variables del modelo de fertilidad (tamaño de la muestra N = 1.741).

COVARIANCE MATRIX TO BE ANALYZED

	PLAN	NACI	ADICI	SATMAX	DUMATR	EDADMU	ACFAMI	RELIG	TRABA
PLAN	1.191								
NACI	0.358	0.997							
ADICI	0.110	-0.508	0.804						
SATMAX	0.487	0.132	0.244	1.180					
DUMATR	0.515	1.862	-1.750	-0.035	9.839				
EDADMU	-0.381	-0.425	-0.470	0.147	1.127	16.553			
ACFAMI	0.168	0.287	-0.093	0.098	0.377	-0.287	1.186		
RELIG	0.389	0.168	0.101	0.406	0.482	1.278	0.362	2.343	
TRABA	-0.219	-0.533	0.179	-0.189	0.503	0.907	-0.411	-0.122	3.757

Las estimaciones MC23, ULS y ML obtenidas mediante LISREL a partir de dicha matriz de varianzas y covarianzas muestral vienen presentadas en la tabla 9.

---

[12] Utilizamos aquí una submuestra con fines «exploratorios». El resto de la muestra se reserva para un análisis «confirmatorio» posterior.

[13] Los datos aquí utilizados han sido cedidos por el Netherland Central Bureau of Statistics.

[14] Los puntos de vista expresados en este capítulo son los de los autores y no coinciden necesariamente con los del Central Bureau of Statistics o ninguna otra institución.

TABLA 9. Estimaciones MC2E, ULS y ML de los parámetros del modelo de fertilidad especificado en la figura 5.

A) ESTIMACIONES MCZE

INITIAL ESTIMATES (TSLS)

BETA

	PLAN	NACI	ADICI
PLAN	0.0	0.0	0.0
NACI	0.206	0.0	0.0
ADICI	0.711	-0.756	0.0

GAMMA

	SAIMAX	DUMAIR	EDADMU	ALFAMI	RELIG	IRABA
PLAN	0.374	0.0	-0.034	0.069	0.109	0.0
NACI	0.0	0.184	-0.024	0.100	0.0	-0.138
ADICI	0.0	-0.070	-0.026	0.019	0.0	0.0

PSI

	PLAN	NACI	ADICI
PLAN	0.943		
NACI	-0.038	0.476	
ADICI	-0.444	0.135	0.592

B) ESTIMACIONES MAXIMO VEROSIMILES (ML)

MODELO DE FERTILIDAD: DATOS ENCUESTA NCBS

LISREL ESTIMATES (MAXIMUM LIKELIHOOD)

BETA

	PLAN	NACI	ADICI
PLAN	0.0	0.0	0.0
NACI	0.204	0.0	0.0
ADICI	0.685	-0.649	0.0

GAMMA

	SAIMAX	DUMAIR	EDADMU	ALFAMI	RELIG	IRABA
PLAN	0.371	0.0	-0.034	0.068	0.113	0.0
NACI	0.0	0.187	-0.024	0.099	0.0	-0.139
ADICI	0.0	-0.064	-0.025	-0.004	0.0	0.0

PSI

	PLAN	NACI	ADICI
PLAN	0.943		
NACI	-0.037	0.475	
ADICI	-0.450	0.084	0.572

SQUARED MULTIPLE CORRELATIONS FOR STRUCTURAL EQUATIONS

PLAN	NACI	ADICI
0.209	0.507	0.367

TOTAL COEFFICIENT OF DETERMINATION FOR STRUCTURAL EQUATIONS IS 0.6

MEASURES OF GOODNESS OF FIT FOR THE WHOLE MODEL :

CHI-SQUARE WITH 4 DEGREES OF FREEDOM IS 53.98 (PROB. LEVEL = 0.0)

GOODNESS OF FIT INDEX IS 0.993

ADJUSTED GOODNESS OF FIT INDEX IS 0.924

ROOT MEAN SQUARE RESIDUAL IS 0.078

TABLA 9 (Continuación)

C) ESTIMACIONES ULS

MODELO DE FERTILIDAD: DATOS ENCUESTA NCBS

LISREL ESTIMATES (UNWEIGHTED LEAST SQUARES)

BETA

	PLAN	NACI	ADICI
PLAN	0.0	0.0	0.0
NACI	0.197	0.0	0.0
ADICI	0.718	-0.600	0.0

GAMMA

	SATMAX	DUMATR	EDADMU	ACFAMI	RELIG	TRABA
PLAN	0.326	0.0	-0.050	0.190	0.121	0.0
NACI	0.0	0.195	-0.026	0.067	0.0	-0.145
ADICI	0.0	-0.063	-0.026	-0.110	0.0	0.0

PHI

	SATMAX	DUMATR	EDADMU	ACFAMI	RELIG	TRABA
SATMAX	1.180					
DUMATR	-0.035	9.039				
EDADMU	0.147	1.127	16.553			
ACFAMI	0.098	0.377	-0.287	1.186		
RELIG	0.406	0.482	1.278	0.362	2.343	
TRABA	-0.169	0.503	0.907	-0.411	-0.122	3.757

PSI

	PLAN	NACI	ADICI
PLAN	0.921		
NACI	0.050	0.446	
ADICI	-0.501	0.009	0.636

SQUARED MULTIPLE CORRELATIONS FOR STRUCTURAL EQUATIONS

PLAN	NACI	ADICI
0.227	0.553	0.280

TOTAL COEFFICIENT OF DETERMINATION FOR STRUCTURAL EQUATIONS IS 0.702

MEASURES OF GOODNESS OF FIT FOR THE WHOLE MODEL :

GOODNESS OF FIT INDEX IS 0.999

ADJUSTED GOODNESS OF FIT INDEX IS 0.989

ROOT MEAN SQUARE RESIDUAL IS 0.069

Atendiendo a la estimación ML, la consideración de la significación individual de los parámetros conduce a los cocientes entre estimaciones y errores estándars (valores «t») que presentamos en la tabla 10.

TABLA 10. Valores «t» asociados a las estimaciones ML de los parámetros del modelo de fertilidad.

MODELO DE FERTILIDAD: DATOS ENCUESTA NCBS

T-VALUES

BETA

	PLAN	NACI	ADICI
PLAN	0.0	0.0	0.0
NACI	5.733	0.0	0.0
ADICI	10.625	-11.786	0.0

GAMMA

	SAIMAX	QUMAIR	EDADMU	ALFAMI	MELIG	IRADA
PLAN	17.760	0.0	-5.783	3.104	8.387	0.0
NACI	0.0	34.858	-5.824	6.095	0.0	-15.772
ADICI	0.0	-5.926	-5.154	-0.217	0.0	0.0

PSI

	PLAN	NACI	ADICI
PLAN	29.445		
NACI	-1.001	29.050	
ADICI	-10.625	2.494	13.748

Cuando un parámetro estimado («efecto directo») es igual a cero, el valor «t» correspondiente tiene, aproximadamente, distribución normal estándar, de manera que, a partir de dicha tabla 10, cabe observar la significación de los distintos «efectos» especificados por el modelo. Los valores estandarizados de los parámetros correspondientes a aquellos efectos altamente significativos (valores «t» superiores a 3) se presentan en la figura 5.

Referente al contraste Gi-cuadrado de las restricciones de sobreidentificación implicadas por el modelo, conviene alertar al lector sobre el valor del estadístico de contraste observado. Dicho valor, que viene representado en la tabla 10, es 53.98, y con respecto a 4 grados de libertad (4 restricciones de sobreidentificación) es altamente significativo (nivel de probabilidad 0.00). Por tanto, un constraste al nivel de significación habitual (.05 ó .01), rechazaría la hipótesis de que las restricciones de sobreidentificación implicadas por el modelo son correctas; o sea, rechazaríamos la hipótesis de que el modelo especificado es correcto. Nos llevaría fuera de los límites de este capítulo la resolución, aquí, de este conflicto entre los «datos» y la especificación del modelo («teoría») suscitado a raíz de la observación del estadístico de contraste Gi-cuadrado. El lector interesado en esta problemática puede consultar, con respecto a este modelo de fertilidad concreto, Stronkhorst y Satorra (1984) o, referente a la problemática general de utilización del contraste de la razón de verosimilitud, Saris, den Ronden y Satorra (1984).

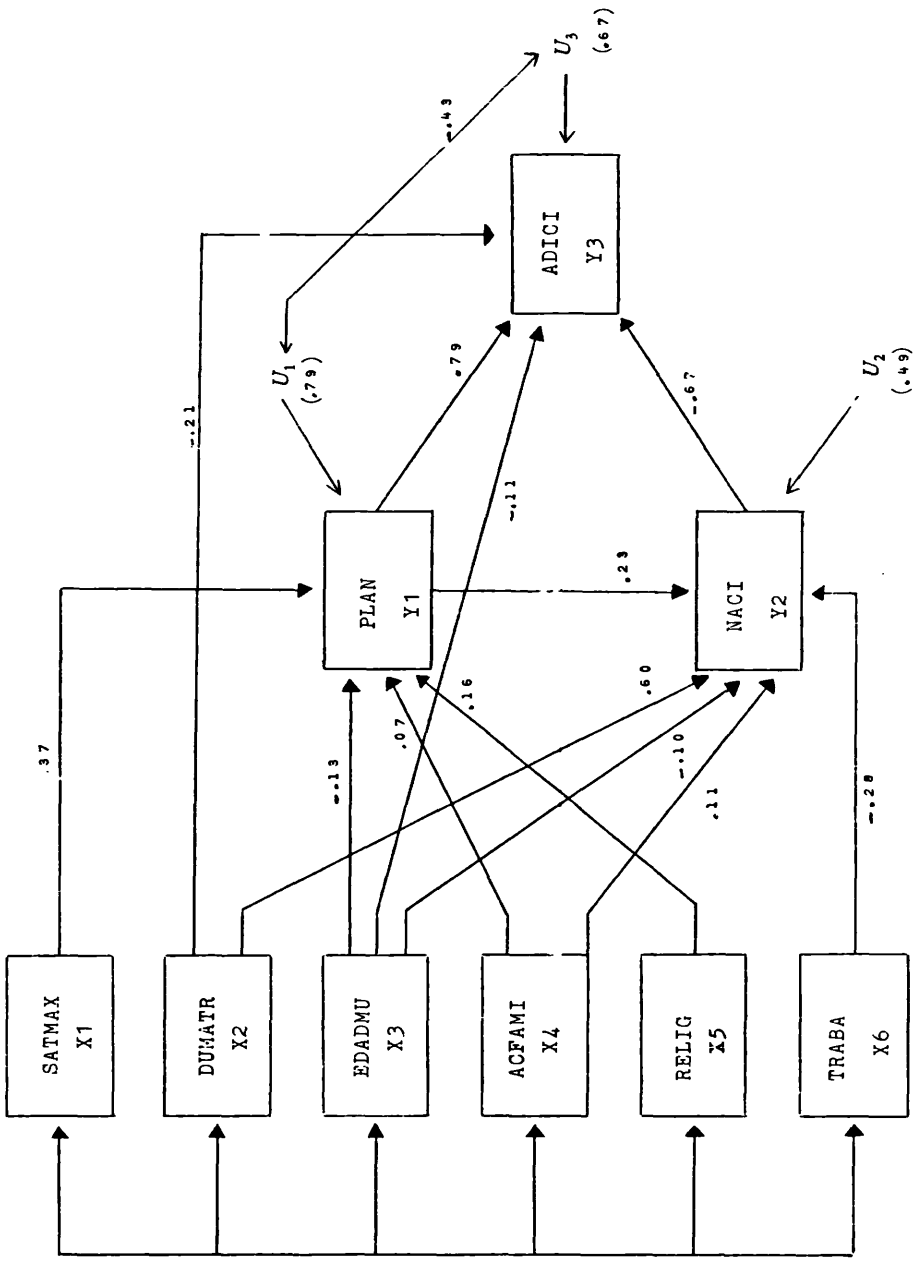


FIGURA 5. Diagrama *path* y valores estandarizados de los parámetros significativos.

9.8 Conclusión

En este capítulo se han introducido los métodos estadístico adecuados para el análisis de los modelos de causalidad en el contexto de la investigación empírica en las ciencias sociales. Un ejemplo con datos «artificiales» y otro con datos empíricos han ilustrado la metodología propuesta.

Antes de concluir el capítulo, queremos señalar, sin embargo, algunas limitaciones de las técnicas propuestas. Quizá la limitación más fundamental es la que proviene de uno de los supuestos asumidos al empezar el capítulo; concretamente, del supuesto según el cual las variables «teóricas» y relevantes en el modelo son directamente observables u observables sin error. Muchas veces, el investigador estará interesado en incluir en el modelo variables que, siendo conceptualmente bien definidas, no se corresponden con ninguna característica medible (observable) de los individuos. Afortunadamente en la actualidad existen métodos adecuados para hacer frente a estas complicaciones. Precisamente, dicha problemática del error de medida, o de la naturaleza no observable de las variables teóricas, es la que vendrá tratada en el próximo capítulo.

Otra de las limitaciones del tratamiento dado en este capítulo es la que proviene del supuesto relativo a la naturaleza interval de las variables manejadas. A menudo, en estudios empíricos en las ciencias sociales, las variables son de escala ordinal siendo necesario en dicho caso introducir modificaciones en los procedimientos de estimación propuestos. Finalmente, no queremos dejar de mencionar las limitaciones inherentes a los supuestos (a) de normalidad de los términos de perturbación y (b) linealidad de los «efectos» entre variables. La violación de estos supuestos puede implicar también que las distribuciones de los estadísticos utilizados difieran sensiblemente de sus distribuciones teóricas.

Por último, cabe decir que el lector interesado puede completar la presente introducción a los modelos de causalidad con referencias en el ámbito de la Sociología como las de Blalock (1964), Duncan (1975), Asher (1976), Hanushek y Jackson (1977), Bagozzi (1983) y Saris y Stronkhorst (1984). Un *overview* del tratamiento econométrico dado a los modelos de ecuaciones estructurales puede encontrarse en Hausman (1983).

10. Tres enfoques diferentes para resolver el problema del error aleatorio de medida en los modelos de ecuaciones lineales estructurales*

por *Willem E. Saris*
Traducción: *Juan Javier Sánchez Carrión*

10.1. Introducción

En general, las variables que aparecen en la teoría de las ciencias sociales son no-observables o solamente observables con error. Esto significa que hay una diferencia entre las variables teóricas y las variables observadas.

Si representamos las varaibles teóricas por η_i, las variables observadas por y_i y el error de medida por e_i, la forma más simple de la relación entre estas variables queda representada en la ecuación [1].

$$y_i = \eta_i + e_i \qquad [1]$$

Se ha hecho el supuesto, sin pérdida de generalidad, de que todas las variables están expresadas en desviaciones de sus medias, lo que lleva al resultado de que las medias de estas variables a las que se les ha cambiado de escala son igual a cero

$$E(y_i) = 0 \; ; \; E(\eta_i) = 0, \; E(e_i) = 0 \qquad [2]$$

También se hace el supuesto de que los errores de medida no están relacionados con las variables teóricas ni tampoco están relacionados entre sí, o que

$$E(\eta_i e_i) = 0 \quad \text{para todos } i, j \qquad [3]$$

$$E(e_i e_j) = 0 \quad \text{para todos } i \neq j \qquad [4]$$

En la ecuación [1] las variables no observadas se expresan en la misma escala que las variables observadas, dado que el coeficiente de rescalonamiento que elegimos es igual a uno. Esta es una de las posibilidades para evitar la indeterminación de la unidad de la escala de las variables teóricas no observadas.

En el análisis de caminos[1] se utiliza la ecuación [5] para describir la misma relación:

$$y_i = \lambda_{ii}\eta_i + e_{ii} \qquad [5]$$

---

* El autor agradece a G. Mallenbergh sus útiles comentarios a una versión previa de este artículo.

[1] Véase, por ejemplo, Costner (1969) o Blalock (1969).

con los supuestos [2], [3], [4] más el [6] que afirma que la varianza de las variables no observadas es uno, es decir, se supone que las variables están estandarizadas.

$$E(\eta_i^2) = 1 \qquad [6]$$

Esta es otra forma de evitar la misma indeterminación de la unidad de la escala de las variables no observadas.

Ambas formulaciones se han utilizado en la teoría clásica del test[2]. Aunque las dos formas tienen sus ventajas haremos uso de la forma de la ecuación [1] con los supuestos [2], [3], [4] por razones que resultarán claras después.

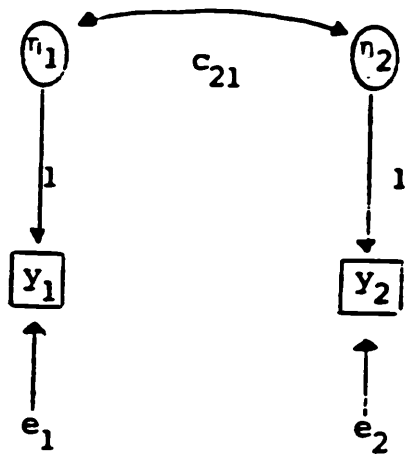


FIGURA 1. Modelo *path* para el modelo de medida de dos variables teóricas con una observación por variable.

En el caso de dos variables teóricas, para las que se ha hecho una observación, el modelo tiene una forma semejante al que aparece en la figura 1.

Si indicamos las varianzas y covarianzas entre las variables observadas mediante σ_{ij}, las varianzas y covarianzas de las variables teóricas mediante c_{ij} y las varianzas y covarianzas de los términos de error por θ_{eij} entonces las varianzas y covarianzas de las variables observadas se pueden expresar con los parámetros del modelo

$$\begin{aligned}
\sigma_{11} &= E(\eta_1 + e_1)(\eta_1 + e_1) = c_{11} + \theta_{e11} \\
\sigma_{22} &= E(\eta_2 + e_2)(\eta_2 + e_2) = c_{22} + \theta_{e22} \qquad [7] \\
\sigma_{21} &= E(\eta_2 + e_2)(\eta_1 + e_1) = c_{21}
\end{aligned}$$

[2] Véase, por ejemplo, Lord y Novick (1968) o Jöreskog (1971).

donde hemos dejado fuera algunos elementos de acuerdo con los supuestos que hicimos bajo [3] y [4].

Según [7] está claro que las varianzas de las variables observadas no son las mismas que las varianzas de las variables teóricas.

Tal como podía esperarse, la diferencia surge debido a los términos de error. Pero esto significa que el uso de variables observadas en lugar de variables teóricas para estimar la correlación o la regresión entre las variables teóricas lleva a estimaciones sesgadas e inconsistentes[3]. Por lo tanto, hay que introducir alguna corrección para el error de medida. Sin embargo esta corrección sólo es posible si se conocen las varianzas de los términos de error. Según las tres ecuaciones de [7] está claro que no se pueden deducir estas varianzas —incluso aunque se conocieran perfectamente las varianzas y covarianzas de las variables observadas—, dado que sólo hay tres ecuaciones y tenemos que estimar cinco parámetros. Esto significa que hay un problema de identificación.

En la literatura se han mencionado tres enfoques diferentes para estimar la varianza del error de medida.

1) introducción de los indicadores múltiples,
2) replicación de la misma observación
3) replicación con indicadores múltiples.

Los mismos enfoques también se pueden utilizar para estimar las varianzas y covarianzas de las variables teóricas.

A continuación vamos a discutir los tres enfoques. Para ello introduciremos algunos modelos que se utilizan en psicología para especificar el error de medida aleatorio de los modelos (Lord y Novick, 1968; Jöreskog, 1970, 1971, 1974; Jöreskog y Sörbon, 1977). Discutiremos la estimación y la verificación de los modelos ofreciendo un ejemplo.

10.2. Indicadores múltiples

El primer enfoque consiste en la introducción de más de una variable observada para cada variable teórica. Estos indicadores múltiples para cada variable se pueden elegir de diferentes formas. Se puede hacer una distinción entre indicadores paralelos, τ-equivalentes y congenéricos (Jöreskog, 1971, 1975; Werts, Jöreskog, Linn, 1972; Van de Kemp, Mallenbergh, 1976).

Si dos indicadores representan la misma variable teórica, se expresan en la misma unidad de medida y las varianzas de los errores son las mismas, en ese caso decimos que los indicadores son «paralelos».

Tales indicadores se pueden formular mediante redacciones diferentes de la misma pregunta o test, o dividiendo aleatoriamente los items en diferentes tests.

En tales casos se puede formular el siguiente modelo de medida:

$$\begin{aligned} y_1 &= \eta + e_1 \\ y_2 &= \eta + e_2 \end{aligned} \qquad [8]$$

[3] Véase, por ejemplo, Johnston (1963) o Godberger (1973).

donde

$$\theta_{e11} = \theta_{e22}$$

y los supuestos [2], [3] y [4].

Si dos indicadores representan las mismas variables teóricas, se expresan en la misma unidad de medida, pero los errores de medida no tienen las mismas varianzas, estos indicadores se llaman τ-equivalentes.

Un ejemplo de tales indicadores podrían ser dos tests de aritmética de iguales longitudes, de los cuales uno se ha construido con más cuidado que el otro. Ambos vendrán expresados en la misma unidad pero las varianzas de sus errores serán diferentes. En el caso de indicadores τ-equivalentes el modelo es semejante al descrito en [8], pero sin que las varianzas de los errores sean iguales.

Si dos indicadores representan la misma variable teórica, pero no se expresan en la misma unidad de medida ni tienen iguales las varianzas de los errores, estos indicadores se llaman «congenéricos»[4].

Un ejemplo de tales indicadores son los utilizados para medir el estatus socioeconómico; esto es, los ingresos, los estudios y la ocupación. Cada uno de estos indicadores se expresa en diferentes unidades de medida y la varianza del error también será diferente en los tres indicadores. En tales casos no está claro en qué unidades se debería de expresar la variable teórica. Como consecuencia, la variable teórica generalmente se normaliza. Por razones que serán evidentes más adelante elegimos expresar la variable teórica en las unidades de medida de uno de los indicadores. Esto significa que hay que introducir en la ecuación una constante de rescalonamiento para los otros indicadores. Indicaremos esta constante mediante λ. En este caso, limitándonos de nuevo a dos indicadores, el modelo es como sigue:

$$\begin{aligned} y_1 &= \eta + e_1 \\ y_2 &= \lambda\eta + e_2 \end{aligned} \qquad [9]$$

A partir de estas definiciones queda claro que el modelo para los instrumentos τ-equivalentes y paralelos es un caso específico del modelo para indicadores congenéricos. En el caso de indicadores τ-equivalentes el valor de $\lambda = 1$ y en el caso de medidas paralelas $\lambda = 1$ y $\theta_{e11} = \theta_{e22}$.

Volviendo de nuevo a nuestro problema original, relativo a la estimación de las matrices de varianza-covarianza de las variables teóricas, la figura 2 representa un modelo con dos variables teóricas y dos conjunto de indicadores congenéricos.

Este modelo se puede formular mediante las ecuaciones siguientes:

$$\begin{aligned} y_1 &= \eta_1 + e_1 \\ y_2 &= \lambda_{21}\eta_1 + e_2 \\ y_3 &= \eta_2 + e_3 \\ y_4 &= \lambda_{42}\eta_2 + e_4 \end{aligned} \qquad [10]$$

[4] Con esta formulación utilizamos la misma notación que Werts, Jöreskog y Linn (1972). En otras publicaciones se utiliza la ecuación [5], lo que lleva a formas equivalentes pero con una notación diferente.

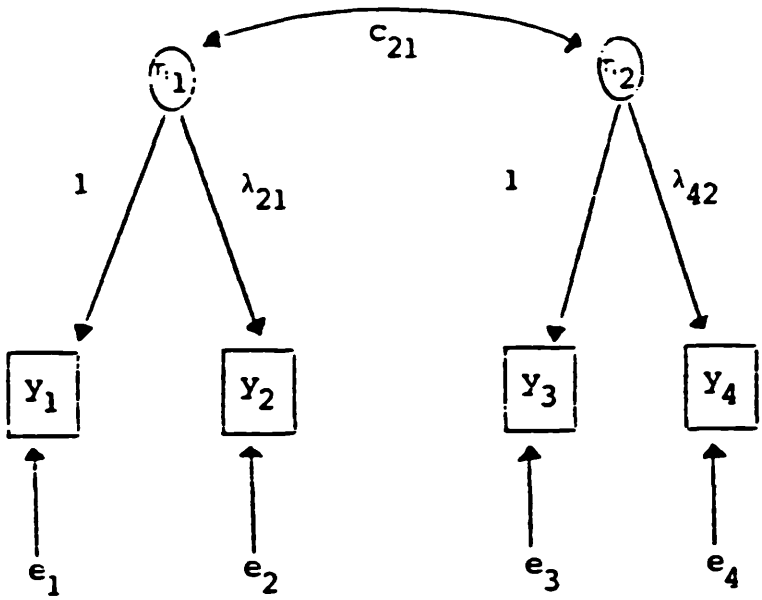


FIGURA 2. Modelo *path* de una teoría con dos variables teóricas y dos conjuntos de instrumentos congenéricos.

donde, tal como es costumbre:

$$E(\eta_i e_j) = 0 \quad \text{para todos } i, j$$

$$E(e_i e_j) = 0 \quad \text{para todos } i \neq j \qquad [11]$$

$$E(y_i) = E(\eta_i) = E(e_i) = 0 \quad \text{para todos } i$$

En este caso hay diez varianzas y covarianzas de las variables observadas que son función de nueve parámetros: c_{11}, c_{21}, c_{22}, θ_{e11}, θ_{e22}, θ_{e33}, θ_{e44}, λ_{21}, λ_{42}. Diferentes autores han mostrado que este modelo es identificable y que puede ser verificado parcialmente. Así, si el modelo es correcto, las varianzas y las covarianzas de las variables teóricas se pueden estimar directamente de los datos. Lo mismo se puede decir cuando los instrumentos son τ-equivalentes o paralelos, puesto que en estos casos hay que estimar menos parámetros.

Un problema con este enfoque es que los dos instrumentos pueden tener alguna varianza común que no se origina en la variable teórica sino, por ejemplo, en variables específicas a un test. Este problema ha sido discutido por Campbell y Fiske (1964) y Wers, Linn y Jöreskog (1974). Cuando se utilizan sólo *dos* indicadores tal hipótesis no se puede contrastar, puesto que conduce a modelos con once parámetros, mientras que Costner (1969) y Costner y Schoenberg (1973) han mostrado que tales tests son posibles siempre y cuando se utilizen tres o más indicadores. Otro problema es que este modelo no se puede distinguir del modelo de análisis factorial donde no se asume que las va-

riables observadas sólo midan la misma variable teórica (Alwin y Jackson, 1980). Para ver una solución a este problema podemos acudir a Saris (1983).

En la medida en que las varianzas y las covarianzas se pueden estimar directamente de modelos con dos variables teóricas, no necesitamos mostrar este modelo con más variables teóricas e indicadores múltiples, dado que el número de varianzas y covarianzas aumenta más deprisa que el número de parámetros desconocidos. Por lo tanto, no habrá más problema de identificación.

10.3. Replicación

Otra manera de tratar el problema del error de medida es mediante la replicación de la misma observación en diferentes puntos del tiempo. Cuando indicamos la variable teórica en el tiempo t_i, por η_i, la variable observada en el tiempo t_i por y_i y el término del error en el tiempo t_i por e_i esta replicación lleva al modelo representado en la figura 3.

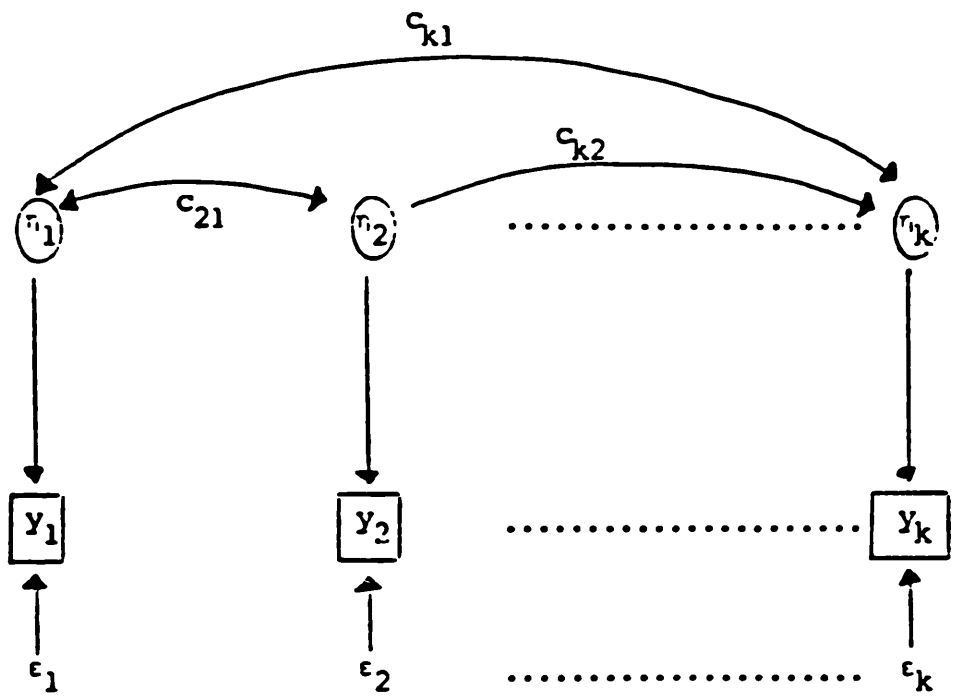


FIGURA 3. Modelo *path* de observación replicada para una variable teórica.

Podemos formular este modelo mediante k ecuaciones con la forma de la ecuación [1], utilizando los mismos supuestos hechos en [2], [3] y [4].

En la medida en que se utiliza el mismo instrumento en cada ocasión, es razonable suponer que las varianzas de los errores en los diferentes puntos del tiempo son iguales, excepto fluctuaciones aleatorias:

$$\theta_{e11} = \theta_{e22} = \ldots = \theta_{ekk} = \theta_e \qquad [12]$$

La varianza de las variables observadas se puede expresar como:

$$\sigma_{ii} = c_{ii} + \theta_e \qquad [13]$$

Las covarianzas son:

$$\sigma_{ij} = c_{ij} \qquad [14]$$

Las ecuaciones [13] y [14] muestran que hay un parámetro más que varianzas y covarianzas para las variables observadas. Esto significa que sin simplificar las restricciones la replicación no conduce a estimaciones únicas de las varianzas y covarianzas de las variables teóricas.

En la literatura se han mencionado diferentes clases de restricciones (Lord y Novick, 1968; Heise, 1969; Wiley y Wiley, 1970; Jöreskog, 1970, 1975). Nosotros sólo discutiremos dos posibilidades. La primera consiste en que el valor de las variables teóricas no se modifique durante el período de observación[6].

En este caso, las diferentes observaciones se pueden ver como instrumentos paralelos para la misma variable, de forma que sean aplicables las derivaciones discutidas en la última sección. Esto significa que es posible estimar la matriz de varianza-covarianza de las variables teóricas si al menos se pueden hacer dos observaciones por cada variable.

Un problema de este supuesto es que sólo es válido si el período de tiempo entre las observaciones es relativamente corto. Pero esto significa que otras variables causantes del error de medida podrían permanecer estables a lo largo del tiempo o que el entrevistado recordase su respuesta de la última ocasión y tratase de ser consistente. En este caso la correlación entre las variables observadas no está causada sólo por el hecho de que midan la misma variable teórica sino también por el efecto atribuible a la memoria. En tal caso no se mantiene el supuesto de que los errores no están correlacionados. De este modo, si el período de tiempo entre las observaciones es corto se puede asumir que las variables teóricas no cambian, pero hay que verificar la correlación de los errores. Tal como indicamos al final del apartado 10.2, ésto sólo es posible en los casos donde hay tres o más observaciones para cada variable teórica.

Un enfoque alternativo consiste en extender las observaciones sobre un periodo más largo, con el fin de que el efecto de la memoria sea desdeñable. Sin embargo, en este caso parece poco realista asumir que la variable teórica todavía tenga el mismo valor, y, por lo tanto, hay que hacer supuestos relativos a las relaciones entre las variables teóricas.

Una posibilidad es el «modelo de retardo -1», discutido por Wiley y Wiley. Los autores suponen que el valor de la variable teórica en t_i es una función de la misma va-

[5] Véase, por ejemplo, Costner (1969) y Costner y Schoenberg (1973).
[6] Véase nota 2.

riable en t_{i-1} y alguna perturbación (disturbance) que es independiente de la variable en momentos anteriores y de otras perturbaciones. Si indicamos el efecto de η_j sobre η_i mediante β_{ij} y el término perturbación por ζ_i, y si nos limitamos a tres observaciones, el modelo tiene la forma del modelo en la figura 4.

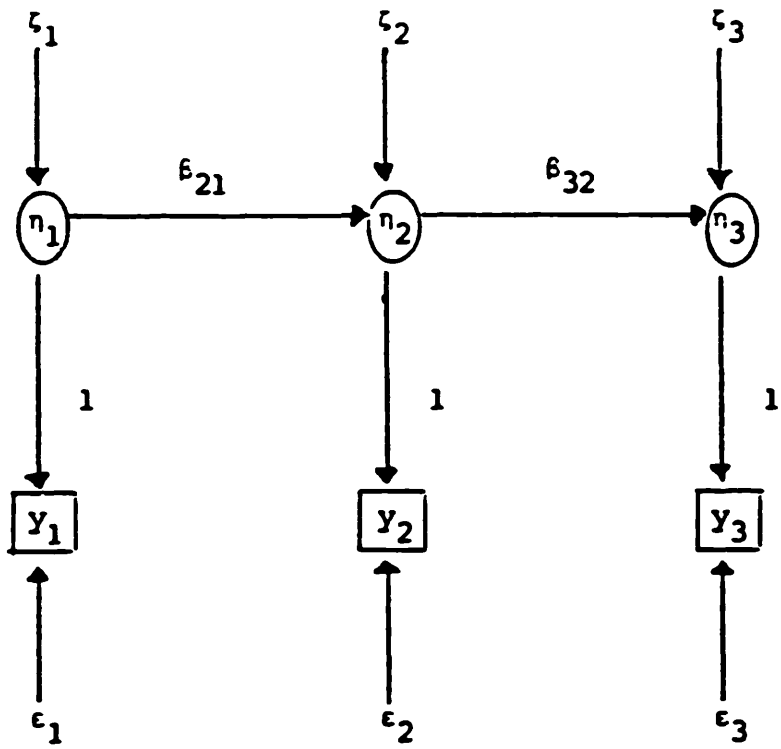


FIGURA 4. Modelo *path* de una observación replicada para una variable teórica, según Wiley y Wiley.

En este modelo las relaciones entre las variables teóricas se pueden especificar como en la ecuación [15]:

$$\begin{aligned} \eta_1 &= \zeta_1 \\ \eta_2 &= \beta_{21}\eta_1 + \zeta_2 \\ \eta_3 &= \beta_{32}\eta_2 + \zeta_3 \end{aligned} \qquad [15]$$

donde

$$E(\eta_i) = E(\zeta_i) = 0 \text{ para todos } i$$

$$E(\eta_i \zeta_j) = 0 \qquad \text{para todos } i, j \qquad [16]$$

$$E(\zeta_i \zeta_j) = 0 \qquad \text{para todos i} \neq j$$

Las relaciones entre las variables observadas y las teóricas son de la forma de la ecuación [1], con los supuestos [2], [3], [4] y que las varianzas de los errores sean las mismas. Bajo estas condiciones el modelo queda determinado por 6 parámetros, c_{11}, β_{21}, β_{32}, θ_e, y las varianzas de los términos de perturbación, indicados por Ψ_{22} y Ψ_{33}. Se pueden obtener seis varianzas y covarianzas de los datos y Wiley y Wiley han mostrado que los parámetros están verdaderamente identificados (1970: 114). Dadas estas estimaciones, las varianzas y covarianzas de las variables teóricas se pueden estimar tal como sigue:

$$\begin{aligned} c_{21} &= \beta_{21} c_{21} \\ c_{31} &= \beta_{32} c_{21} \\ c_{22} &= \beta_{21}^2 c_{11} + \Psi_{22} \\ c_{32} &= \beta_{32} c_{22} \\ c_{33} &= \beta_{32}^2 c_{22} + \Psi_{33} \end{aligned} \qquad [17]$$

Volviendo al problema de este artículo, podemos plantear la cuestión acerca de si es posible estimar la matriz de varianza-covarianza de las variables teóricas que se han observado tres veces. Indicamos una variable teórica como antes, mediante η_1, η_2, η_3 para t_1, t_2, t_3, las observaciones por y_1, y_2, y_3 y lo errores por e_1, e_2, e_3. La segunda variable teórica se indica mediante η_4, η_5, η_6, las observaciones por y_4, y_5, y_6 y los errores por e_4, e_5, e_6. Asumiendo el modelo previo para dos variables, las varianzas y covarianzas c_{11}, c_{21}, c_{31}, c_{32}, c_{33} y c_{44}, c_{54}, c_{64}, c_{55}, c_{65} y c_{66} se pueden encontrar tal como indicamos arriba. Pero nosotros también sabemos que:

$$\sigma_{ij} = E(y_i y_j) = E(\eta_i + e_i)(\eta_j + e_j) = c_{ij}$$

$$\text{para todos } \begin{matrix} i = 1, 2, 3 \\ j = 4, 5, 6 \end{matrix} \qquad [18]$$

de lo cual se desprende que usando este enfoque se pueden obtener a partir de los datos todas las varianzas y covarianzas de las variables no observadas.

La ventaja de este enfoque es que se puede probar si las variables teóricas cambian. Aún más, lleva a tres estimaciones de la matriz de varianza-covarianza para las variables teóricas y en tres puntos del tiempo. Desde luego, estos datos son mucho más ricos que

los obtenidos mediante el primer procedimiento, donde sólo se estimaba una matriz. La desventaja obvia del procedimiento es que el enfoque de la replicación conduce a una investigación más larga. De igual forma, y en relación con este punto, aparece la falta de representatividad debida a la probable pérdida de contestaciones. Otra desventaja es que se deben introducir restricciones en la estructura que ha producido la matriz de varianza-covarianza. Tales restricciones sólo se pueden verificar cuando se tienen más de tres observaciones, lo cual alarga de nuevo la investigación.

10.4. Replicación con indicadores múltiples

Se puede utilizar la replicación con indicadores múltiples para superar el problema del enfoque de la replicación simple, sin que se pierdan las ventajas.

Cuando se observan dos variables teóricas en dos puntos del tiempo mediante dos indicadores, asumiendo que estos indicadores son congenéricos, el modelo se puede representar como en la figura 5 donde η_1 y η_2 representan una variable teórica en el tiempo t_1 y t_2, y η_3 y η_4 representan la otra variable teórica en t_1 y t_2.

Este modelo se puede representar con las ecuaciones siguientes:

$$
\begin{aligned}
y_1 &= \eta_1 + e_1 \\
y_2 &= \lambda_{21}\eta_1 + e_2 \\
y_3 &= \eta_2 + e_3 \\
y_4 &= \lambda_{42}\eta_2 + e_4 \\
y_5 &= \eta_3 + e_5 \\
y_6 &= \lambda_{63}\eta_3 + e_6 \\
y_7 &= \eta_4 + e_7 \\
y_8 &= \lambda_{84}\eta_4 + e_8
\end{aligned}
\qquad [19]
$$

donde

$$
\begin{aligned}
&E(y_i) = E(\eta_i) = E(e_i) = 0 && \text{para todos } i \\
&E(\eta_i e_j) = 0 && \text{para todos } i, j \\
&E(e_i e_j) = 0 && \text{para todos } i \neq j
\end{aligned}
\qquad [20]
$$

A partir de [19] y de [20] se puede derivar la expresión para las varianzas y covarianzas de las variables observadas en los parámetros del modelo. Esta matriz tiene 36 elementos distintos, en tanto que el modelo se caracteriza por 10 varianzas y covarianzas

para las variables teóricas, 8 varianzas para los términos de error y 4 constantes de rescalonamiento. De esta manera hay 14 ecuaciones más que parámetros desconocidos.

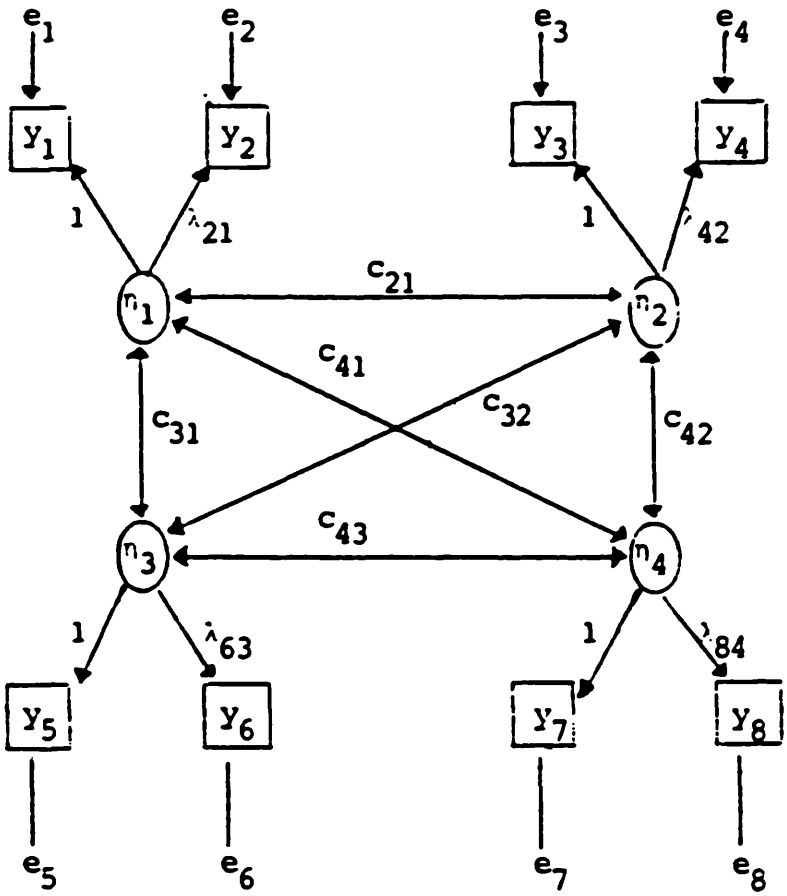


FIGURA 5. Modelo *path* para observación replicada de dos variables teóricas con dos indicadores.

Se puede mostrar que estos parámetros están identificados y que hay muchas posibilidades para formular pruebas del modelo (Blalock, 1970). Por ejemplo, se podrían hacer tests para:

1) la igualdad de las variables teóricas a lo largo del tiempo,
2) la igualdad de la varianza de los errores a lo largo del tiempo,
3) paralelismo y τ-equivalencia entre los indicadores,
4) errores correlacionados a lo largo del tiempo.

Si en lugar de dos se utilizan tres indicadores, también se puede probar para cada conjunto de observaciones la posibilidad de que existan errores correlacionados[7].

Desde luego, una de las ventajas importantes de este diseño de investigación es la posibilidad de hacer los diferentes tests que acabamos de mencionar, mientras que la desventaja que encontrábamos en el estudio de la replicación simple no es tan severa porque sólo se necesitan observaciones en dos puntos del tiempo.

Otra ventaja es la riqueza de los datos comparados con los del estudio del simple indicador múltiple; y ello en la medida en que se tiene información sobre las variables teóricas en dos puntos del tiempo.

10.5. Estimación y verificación

Todos los modelos discutidos hasta ahora son casos especiales de un sistema general de dos ecuaciones matriciales:

$$\begin{aligned} \underset{\sim}{\beta}\underset{\sim}{\eta} &= \underset{\sim}{\zeta} \\ \underset{\sim}{y} &= \underset{\sim}{\Lambda}\underset{\sim}{\eta} + \underset{\sim}{e} \end{aligned} \qquad [21]$$

donde

$$\begin{aligned} E(\underset{\sim}{y}) = 0,\ E(\underset{\sim}{\eta}) = 0 \text{ y } E(\underset{\sim}{e}) = 0 \text{ y } E(\underset{\sim}{\zeta}) = 0 \\ E(\underset{\sim}{\eta}\underset{\sim}{e}') = 0 \text{ y } E(\underset{\sim}{\zeta}\underset{\sim}{e}') = 0 \end{aligned} \qquad [22]$$

A partir de [21] y [22] se desprende que la matriz de varianza-covarianza (C) de las variables teóricas, indicada por η, es

$$\underset{\sim}{C} = \underset{\sim}{B}^{-1}\underset{\sim}{\Psi}\underset{\sim}{B}^{-1'} \qquad \text{donde} \qquad \underset{\sim}{\Psi} = E(\underset{\sim}{\zeta}\underset{\sim}{\zeta}') \qquad [23]$$

y la matriz de varianza-covarianza de las variables observadas en y es

$$\underset{\sim}{\Sigma} = \underset{\sim}{\Lambda}\underset{\sim}{C}\underset{\sim}{\Lambda}' + \underset{\sim}{\theta}_e \qquad \text{donde} \quad \underset{\sim}{\theta}_e = E(\underset{\sim}{e}\underset{\sim}{e}') \qquad [24]$$

Los parámetros en las matrices $\underset{\sim}{B}$, $\underset{\sim}{\Psi}$, $\underset{\sim}{\Lambda}$, $\underset{\sim}{\theta}_e$ se pueden estimar mediante la minimización de la función F para todos los parámetros que hay que estimar, donde

$$F = \log|\Sigma| + \text{tr}(S\Sigma^{-1}) - \log|S| - P \qquad [25]$$

S es la matriz muestral de covarianza y P es igual al número de elementos en $\underset{\sim}{y}$. Si la distribución de $\underset{\sim}{y}$ es multinormal, las estimaciones que minimizan F son estimaciones

[7] Blalock ha indicado que no se pueden aflojar simultáneamente todas las restricciones mencionadas aquí, dado que esto lleva a modelos no identificables. De esto se deriva que no todos los errores se pueden contrastar simultáneamente, pero hay más espacio para la verificación en este caso que en cualquiera de los otros enfoques anteriormente mencionados.

máximo verosimilitud, eficientes en grandes muestras. En la medida en que este sistema es de nuevo un sistema específico del sistema general estimado por el programa LISREL, todos los modelos que hemos discutido aquí pueden estimarse utilizando el programa LISREL (Jöreskog y Sorbon, 1973). Después de calcular las estimaciones de los parámetros, el programa calcula las estimaciones de los errores típicos, la matriz *C*, la solución estandarizada y las diferencias entre la matriz de varianza-covarianza observada y la matriz de varianza-covarianza que reproducimos basándonos en los parámetros estimados. Esos residuos pueden dar una indicación de la bondad del ajuste del modelo a los datos.

El programa también da una medida χ^2 de la bondad global del ajuste, que se puede ver como un test del modelo especificado frente a la alternativa más general, una matriz definida positiva sin limitaciones. Los grados de libertad para esta medida χ^2 son igual a la diferencia entre el número de elementos distintos en *S* y el número de parámetros que hay que estimar.

Se pueden contrastar las restricciones de un modelo estimando primero el modelo sin restricciones y después el modelo restringido. La diferencia en χ^2 es asintóticamente una χ^2 que tiene como grados de libertad la diferencia de grados de libertad entre los dos modelos.

10.6. Un ejemplo

En la medida en que el modelo de indicadores múltiples ha sido discutido por Costner (1969), Costner y Schoenberg (1973) y muchos otros y que el modelo de replicación también lo han estudiado Heise (1969), Wiley y Wiley (1970) y otros, nosotros sólo vamos a ejemplificar el modelo de la replicación con indicadores múltiples. Analizaremos los datos para dos variables teóricas del estudio de Zaal. En ese estudio se midió un número elevado de variables para una muestra de escolares holandeses; la medición se hizo en dos puntos del tiempo. La primera observación se hizo en la guardería y la segunda dos años después, en la escuela primaria. Hemos elegido la inteligencia y la actitud hacia el trabajo de entre el gran número de variables. La primera se midió utilizando un test de inteligencia (Drent e.a., 1968). La medida de la «actitud hacia el trabajo» estaba basada en las puntuaciones del profesor (Zaal, 1978).

Los indicadores múltiples para ambas variables se desarrollaron por el método de la partición en mitades *(split-half)*, lo que significa que para cada variable teórica en los dos puntos del tiempo se obtenía una puntuación basada en los ítems numerados impares y otra puntuación basada en los pares.

Si la inteligencia en t_1 es η_1 y en el tiempo $t_2\eta_2$, donde las observaciones están representadas por y_1 e y_2 respectivamente, y_3 e y_4, respectivamente, mientras que la actitud hacia el trabajo en t_1 se representa por η_3 y η_4 en t_2, donde y_5 e y_6 son las observaciones en t_1 e y_7 e y_8 las observaciones en t_2, el modelo de medida se puede formular tal como se ha hecho en la figura 5 y las ecuaciones [19] y [20]. Con el fin de probar la posibilidad de errores correlacionados en el tiempo, hay que aflojar la restricción de que $E(e_i e_j) = 0$.

Permitiremos que haya la posibilidad de que los errores entre observaciones de las mismas variables, en diferentes puntos del tiempo, estén correlacionados. Este nuevo modelo se representa en la figura 6.

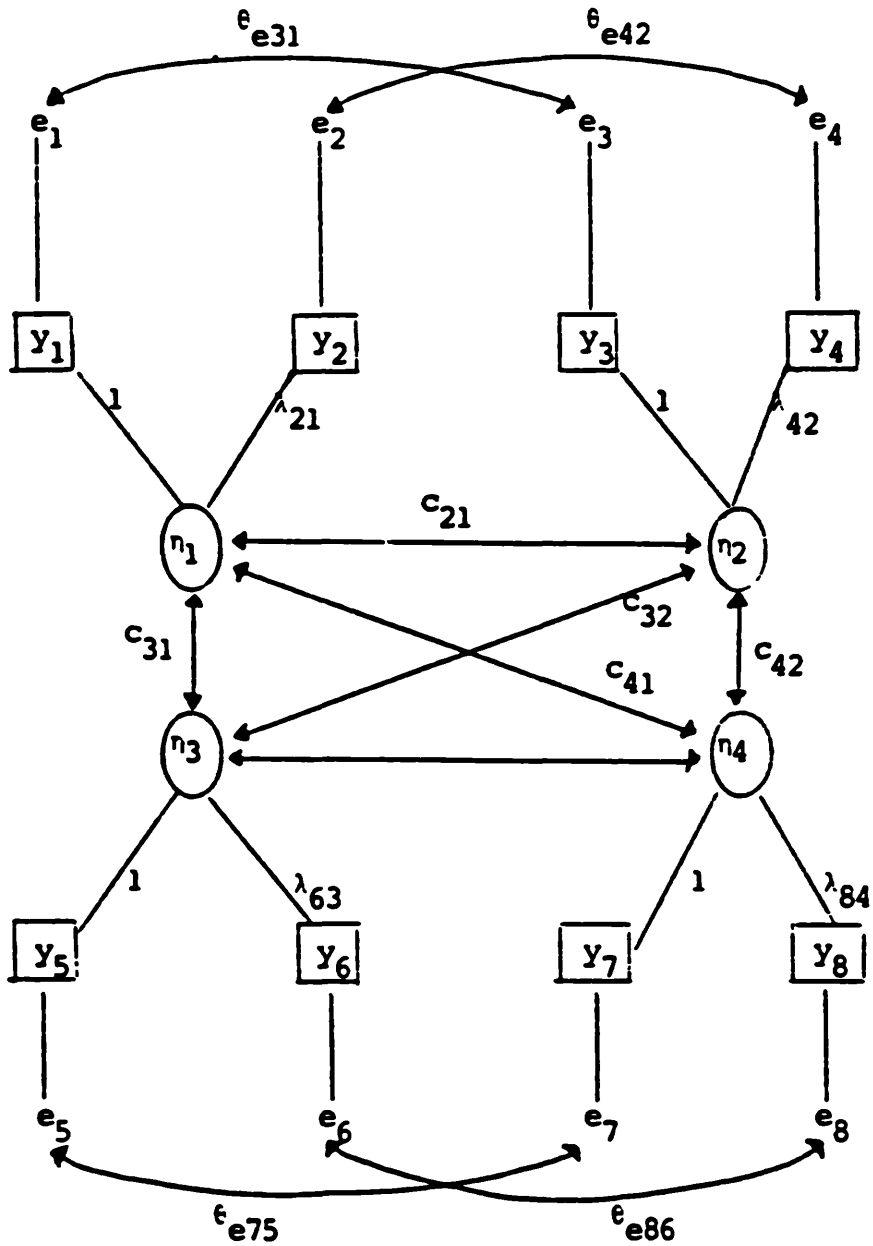


FIGURA 6. Modelo *path* para dos variables teóricas con indicadores congenéricos y términos de error correlacionados a lo largo del tiempo.

Este modelo, al que llamaremos modelo 1, contiene 26 parámetros que hay que estimar. Si el modelo se especifica como un caso especial del sistema más general formulado en [21] y [22], estos parámetros se pueden estimar usando LISREL.

La parte de medida del modelo queda especificada si se formulan $\underset{\sim}{\Lambda}$ y $\underset{\sim}{\theta}_{\varepsilon}$. Para el modelo 1 estas matrices son:

$$\underset{\sim}{\Lambda} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \lambda_{21} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & \lambda_{42} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & \lambda_{53} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & \lambda_{84} \end{pmatrix} \qquad [26]$$

$$\underset{\sim}{\theta}_{\varepsilon} = \begin{pmatrix} \theta_{\varepsilon 11} & & & & & & & \\ 0 & \theta_{\varepsilon 22} & & & & & & \\ \theta_{\varepsilon 31} & 0 & \theta_{\varepsilon 33} & & & & & \\ 0 & \theta_{\varepsilon 42} & 0 & \theta_{\varepsilon 44} & & & & \\ 0 & 0 & 0 & 0 & \theta_{\varepsilon 55} & & & \\ 0 & 0 & 0 & 0 & 0 & \theta_{\varepsilon 66} & & \\ 0 & 0 & 0 & 0 & \theta_{\varepsilon 75} & 0 & \theta_{\varepsilon 77} & \\ 0 & 0 & 0 & 0 & 0 & \theta_{\varepsilon 86} & 0 & \theta\varepsilon_{88} \end{pmatrix}$$

Las relaciones entre las variables teóricas no estaban restringidas; esto se puede hacer especificando que $\underset{\sim}{B} = \underset{\sim}{I}$. En este caso $\underset{\sim}{C} = \underset{\sim}{\Psi}$ y el modelo se reduce a un modelo de análisis factorial con factores correlacionados sin restringir y algunos términos de error correlacionados.

Aunque esperábamos que el modelo 1 de la figura 6 ajustase los datos, pensábamos que esto también lo podía hacer un modelo más simple con menos parámetros. Sin embargo, comenzamos con este modelo dado que todos los modelos restantes son casos específicos del mismo, así podríamos contrastar las restricciones introducidas en el modelo por comparación con los valores χ^2 de los modelos.

En el modelo 2 se suponía que cuando el período de tiempo entre las observaciones era de dos años, las correlaciones entre los términos de error a lo largo del tiempo eran cero:

$$\theta_{\varepsilon 31} = \theta_{\varepsilon 42} = \theta_{\varepsilon 75} = \theta_{\varepsilon 86} = 0 \qquad [27]$$

El modelo 3 era el mismo que el 2, excepto por el hecho de que se asumía que los instrumentos de medida eran τ-equivalente y las unidades de la escala eran las mismas

$$\lambda_{21} = \lambda_{42} = \lambda_{63} = \lambda_{84} = 1 \qquad [28]$$

El modelo 4 era igual que el 3, excepto que aquí se asumía que los instrumentos de medida en cada punto del tiempo eran paralelos, dado que se construían por el procedimiento de la partición en mitades.

$$\theta_{e11} = \theta_{e22}, \theta_{e33} = \theta_{e44}, \theta_{e55} = \theta_{e66}, \theta_{e77} = \theta_{e88} \qquad [29]$$

El modelo 5 era igual que el 4, excepto que aquí también se asumía que los errores de medida en los diferentes puntos del tiempo, para las mismas variables, eran iguales puesto que se utilizaban los mismos instrumentos.

$$\theta_{e11} = \theta_{e33} \text{ y } \theta_{e55} = \theta_{e77} \qquad [30]$$

TABLA 1. Coeficientes de correlación entre 8 vars. observadas en un estudio de Zaal (n = 72).

Variable teórica	Variable observada	Correlaciones							
$1Qt_1$	y_1	1.000							
	y_2	.768	1.000						
$1Qt_2$	y_3	.542	.590	1.000					
	y_4	.446	.523	.784	1.000				
WAt_1	y_5	.427	.500	.348	.484	1.000			
	y_6	.496	.512	.340	.428	.870	1.000		
WAt_2	y_7	.236	.250	.238	.316	.464	.386	1.000	
	y_8	.222	.220	.206	.245	.386	.338	.849	1.000
		y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8

TABLA 2. Las comparaciones de la bondad del ajuste de los 5 modelos para los datos de la tabla 1 y un test de las diferentes hipótesis.

Modelo	χ^2	g.l.	Prob	Diferencia en χ^2	Diferencia en grados de libertad	Probabilidad	Conclusión
1	9.8	10	.46	–	–	–	No rechazado
2	14.9	14	.39	5.1	4	.25	No rechazado
3	18.9	18	.39	4.0	4	.40	No rechazado
4	20.7	22	.54	1.8	4	.75	No rechazado
5	22.1	24	.57	1.4	2	.40	No rechazado

La matriz de correlaciones de las variables observadas se presenta en la tabla 1. Los parámetros de todos los modelos se estimaron utilizando LISREL. La tabla 2 muestra la bondad del ajuste de los diferentes modelos con los grados de libertad y la probabilidad del valor de χ^2 o un valor mayor si el modelo es correcto, las diferencias de valor de χ^2 y de los grados de libertad con el modelo previo y la probabilidad de este valor o de uno mayor si el conjunto especificado de restricciones es correcto.

La tabla 2 muestra que el modelo 1 realmente tenía un ajuste aceptable, pero todas las correlaciones entre los términos de error no eran significativamente diferentes de cero al nivel del 5%. Al restringir estos valores a cero en el modelo 2, el valor de la χ^2 subía pero no significativamente. De este modo el supuesto [27] no se puede rechazar. Cuando introducimos en el modelo 3 la restricción [28] tampoco se reduce dramáticamente el ajuste. Por lo tanto los instrumentos de medida se pueden ver como τ-equivalentes. Al pasar del modelo 3 al 4 se puede contrastar la hipótesis de que las varianzas de los errores son iguales, lo que significa que los instrumentos de medida son paralelos. Tampoco esta hipótesis se pudo rechazar.

Finalmente, tampoco daña significativamente el ajuste el supuesto de iguales varianzas de los errores a lo largo del tiempo.

Dado que al efectuar el test no se pudo rechazar ninguna de las restricciones que manteníamos aceptamos el modelo 5 como nuestro modelo final de medida[8].

La tabla 3 resume las características del modelo, tal como se estimaron por el programa.

La tabla 3 muestra que las varianzas de los errores eran relativamente pequeñas, por lo que consecuentemente el «coeficiente de validez» y la fiabilidad eran relativamente altas.

Puesto que no se podía rechazar el modelo 5, vale la pena mirar la matriz de varianza-covarianza de las variables teóricas estimada por el programa simultáneamente con todos los restantes parámetros del modelo. La tabla 4 da estas varianzas y covarianzas, sus errores típicos y los coeficientes de correlación (encima de la diagonal principal).

Dada esta matriz, se pueden hacer muchos análisis de las correlaciones entre las variables teóricas. Por ejemplo, utilizando un enfoque dinámico se pueden estudiar las relaciones entre estas varibles en dos puntos del tiempo. Estas posibilidades son una ventaja típica del enfoque de la replicación. No iremos más allá en este análisis dado que no es el tema de este artículo (Zaal, 1978).

Finalmente, queremos discutir las ventajas que se obtienen al utilizar la ecuación [1] con las variables teóricas sin estandarizar frente a la utilización de la ecuación [5] con las variables teóricas estandarizadas. Cuando usamos [1] existe la posibilidad de comparar las varianzas de las variables teóricas a lo largo del tiempo y de estimar los coeficientes sin normalizar con el fin de describir las relaciones entre las variables teóricas. Esto tiene considerables ventajas cuando se quieren comparar coeficientes en diferentes poblaciones (Blalock, 1968; Duncan, 1977).

También se pueden calcular los coeficientes estandarizados, que son útiles para comparar los efectos de diferentes variables dentro de una población. Cuando se utiliza la

[8] Un supuesto importante que no se puede verificar con dos instrumentos paralelos es la posibilidad de que haya errores correlacionados entre las dos formas paralelas. Esto deja abierta la discusión acerca de si mediante estas variables observadas se ha representado la variable teórica correcta.

TABLA 3. Resumen de las características del modelo de medida.

Variable teórica	Variable observada	Medias de las variables observadas	Varianza de las variables observadas	Varianzas estimadas (1) de los términos de error	Varianzas estimadas (1) de las variables teóricas	Pendiente tipificada (2) λ^*_{ij}	Correlaciones de los tests paralelos o fiabilidad (3)
$1Qt_1$	y_1	32.8	47.06	10.20 (1.2)	38.13 (7.3)	.89	.79
	y_2	31.3	50.69	10.20 (1.2)			
$1Qt_2$	y_3	48.4	42.90	10.20 (1.2)	32.13 (6.3)	.87	.76
	y_4	37.1	40.70	10.20 (1.2)			
WAt_1	y_5	35.6	153.51	20.65 (2.5)	121.50 (22.2)	.92	.86
	y_6	37.1	129.05	20.65 (2.5)			
WAt_2	y_7	36.2	146.89	20.65 (2.5)	126.88 (23.1)	.92	.86
	y_8		150.31	20.65 (2.5)			

(1) Las estimaciones están calculadas por LISREL, los errores típicos se incluyen entre paréntesis.
(2) $\lambda^*_{ij} = \lambda_{ij}\sqrt{\hat{c}_{ii}}\,/\sqrt{\hat{\theta}_{ii}}$.
(3) La correlación es igual a λ^{*2}_{ij} (Lord y Novick o Werts, Linn, Jöreskog, 1972).

TABLA 4. Varianzas y covarianzas de las variables teóricas, sus errores típicos y correlaciones tal como han sido estimados por LISREL con el modelo 5.

$1Qt_1$	38.125 (7.28)	.589	.679	.284
WAt_1	40.073 (10.14)	121.495 (22.16)	.491	.460
$1Qt_2$	23.760 (5.53)	30.681 (9.08)	32.135 (6.28)	.309
WAt_2	19.729 (9.44)	51.122 (17.34)	19.710 (8.80)	126.884 (23.06)
	$1Qt_1$	WAt_1	$1Qt_2$	WAt_2

ecuación [5], hay que hacer un trabajo extra con el fin de obtener estos resultados, que de otra manera están disponibles automáticamente.

10.7. Resumen

Con el fin de obtener los coeficientes de correlación y de regresión entre las variables teóricas hay que tener en cuenta el error de medida de las variables observadas. En la literatura psicométrica y sociológica se han discutido diferentes enfoques para resolver este problema. En este artículo hemos presentado tres enfoques diferentes para estimar las varianzas y covarianzas de las variables teóricas. El uso de los indicadores múltiples lleva a modelos simples relativamente sobreidentificados, que permiten tests sobre las correlaciones entre los términos de error; sin embargo, la información que se obtiene relativa a las variables teóricas es menos interesante que la obtenida con los otros enfoques.

Si se hace la replicación en un corto período de tiempo, el uso de estudios de replicación es comparable con el enfoque de los indicadores múltiples; sin embargo, en este caso hay que hacer un test para los errores correlacionados debidos a los efectos de la memoria.

Si se hace la replicación en un período de tiempo largo, es necesario utilizar modelos más complejos, con restricciones extras para la identificación. En este caso no se puede hacer ningún test, pero los datos son más interesantes. Una desventaja importante es que la investigación consume más tiempo.

El uso de la replicación con indicadores múltiples tiene la ventaja de los modelos sobreidentificados, lo que hace posible tests sobre los errores correlacionados; la información que se obtiene para las variables teóricas es más interesante que en los casos de los simples estudios de indicadores múltiples; y para cada variable sólo se necesitan observaciones en dos puntos del tiempo, reduciendo así considerablemente el tiempo de investigación.

Además, hemos visto que generalmente es aconsejable utilizar al menos tres observaciones o indicadores, dado que así hay más posibilidades para verificar los modelos de medida.

Finalmente, parece aconsejable expresar las variables no observadas en las unidades de la escala de uno de los indicadores. De esta manera se pueden utilizar coeficientes no estandarizados, mientras que si se prefieren las comparaciones los coeficientes estandarizados siempre se pueden calcular después.

11. Análisis de Tablas de Contingencia: Modelos Lineales Logarítmicos

por *Juan Javier Sánchez Carrión*

11.1. Introducción

El tratamiento tradicional que se ha seguido en el análisis de las relaciones entre dos o más variables de tipo cualitativo —variables nominales— ha sido el de construir una tabla llamada de contingencia, a partir de la cual se procedía a ver 1) si existía relación entre las dos variables; y 2) en caso afirmativo, cuál era la intensidad de la asociación. Para el primero de los propósitos lo normal es hacer un test de la Gi-cuadrado; para estudiar la intensidad de la asociación se recurre a una amplia gama de medidas de asociación, que nos resumen en un número la fuerza y la dirección de la relación entre las variables[1].

Dada la limitación y la confusión que se pueden producir en el análisis cuando sólo se analizan dos variables a la vez (Simpson, 1951) es también común introducir otras nuevas variables de control que permitan «cualificar» la relación entre las variables originales. Lazarsfeld (1955) primero y posteriormente Rosenberg (1968) explicaron el tipo de variables de control existentes y el «proceso de elaboración» a seguir cuando se quiere ver la influencia que tiene una tercera variable sobre la relación original entre otras dos.

El segundo procedimiento mencionado tiene una fecundidad extraordinaria en el análisis de datos cualitativos. Sin embargo su utilidad se puede ver ampliada mediante el recurso a una serie de técnicas. En un caso, siguiendo las ideas de James Davis se pueden construir modelos causales en los que los coeficientes de regresión estandarizados que se utilizan cuando trabajamos con variables cuantitativas se sustituyan por diferencias entre proporciones. La técnica sirve tanto para variables dicotómicas como para politomías, y tiene la propiedad de que, también como ocurre en los modelos causales con variables intervales, se pueden descomponer los efectos entre las variables (Davis, 1975, 1980, 1982; Sánchez Carrión, cap. 12 de este libro).

---

[1] Cualquier manual de estadística para las ciencias sociales hace referencia al test de la Gi-cuadrado y a los diferentes coeficientes que se pueden utilizar. Entre los libros específicos de tablas véase Hildebrand y otros (1977) o Reynolds (1977). Ambos sirven de libros introductorios al tema de las tablas de contingencia. Sin ser libros específicos sobre análisis tabular, en castellano se pueden consultar García Ferrando (1984) y Blalock (1978).

Otra técnica, ésta derivada de las razones *(odds)* y no de la diferencia de proporciones o porcentajes, son los modelos lineales logarítmicos. Si bien la técnica del control por una tercera variable es útil cuando el tamaño de las tablas y el número de categorías de las variables es pequeño, su utilidad es menor cuando nos enfrentamos a situaciones complejas donde se aborda el estudio de las relaciones entre un número elevado de variables. En esta situación es conveniente disponer de una técnica estadística que permita medir y contrastar las complejas asociaciones e interacciones que aparecen en tablas multidimensionales. De igual manera, el análisis desarrollado mediante el proceso de elaboración se puede ver mejorado mediante el uso de los modelos lineales logarítmicos en la medida en que, además de descubrir las pautas de asociación entre las variables, podemos medir los efectos de unas variables sobre otras. En concreto, la técnica de los modelos lineales logarítmicos, desarrollada a partir de los trabajos de Goodman (1972, 1972b, 1973 y 1979) y ampliamente difundida en el análisis tabular (Bishop et al., 1975; Fieneberg, 1977; Upton, 1978, 1980a), permite el estudio de:

1. Las pautas de asociación entre variables categóricas, sin distinción entre dependientes e independientes, mediante el análisis de las frecuencias esperadas de las casillas (modelo general logarítmico lineal).
2. Las relaciones entre variables dependientes e independientes, mediante el análisis de las razones esperadas de una variable dependiente en función de una serie de variables independientes (modelo *logit*).
3. La construcción de modelos causales a partir de los modelos *logit*.

Con objeto de explicar los modelos lineales logarítmicos y su utilización en las tres situaciones de investigación descritas vamos a usar los datos procedentes de una investigación sobre emigrantes iberoamericanos en España. La investigación se llevó a cabo en 1981 por Gloria Lutz y Miguel Roiz, quienes amablemente me cedieron los datos. En la medida en que no estamos interesados en conclusiones de tipo sustantivo sino en la exposición de una técnica de investigación, vamos a asumir que la muestra era aleatoria simple, al tiempo que recodificamos algunas variables.

Tenemos datos de 4 variables: Educación, recodificada en nivel bajo (menos de título de grado medio) y alto (título de nivel medio o más); Antigüedad en el país, recodificada en aquellos que llegaron a España de 1960 a 1970, 71-77 y 1978-80; Permiso de Trabajo, con las categorías No y Sí; e Ingresos, recodificada en ingresos bajos (la mediana de la distribución o menos) e ingresos altos (más de la mediana). Los datos aparecen en la tabla 1.

11.1.1. *Nomenclatura de las tablas de contingencia*

Tomando el ejemplo de una tabla de tres dimensiones, mostraremos la notación que vamos a seguir a lo largo del artículo. Supongamos que tenemos una tabla en la que una variable A tiene I categorías, la variable B tiene J categorías y la varaible C, K categorías. Por simplicidad $I = J = K = 2$, tal como mostramos en la tabla 2.

TABLA 1. Datos para el Análisis.

Estudios	Antigüedad	Permiso	Ingresos	
			Bajos	Altos
Inferiores	1960-70	No	10	7
Inferiores	1960-70	Sí	2	3
Inferiores	1971-77	No	39	9
Inferiores	1971-77	Sí	15	14
Inferiores	1978-80	No	75	10
Inferiores	1978-80	Sí	18	8
Superiores	1960-70	No	7	6
Superiores	1960-70	Sí	2	12
Superiores	1971-77	No	23	19
Superiores	1971-77	Sí	22	22
Superiores	1978-80	No	43	6
Superiores	1978-80	Sí	12	12
			268	128

El número de personas que contestan a la categoría 2 de las variables A y B y a la 1 de la C son f_{221}. En general llamaremos f_{ijk} a la frecuencia de la casilla determinada por el cruce de las categorías i, j, k de las variables A, B, C, respectivamente.

Junto a la frecuencia de la casilla es útil calcular los *marginales* de la tabla. En una tabla de tres dimensiones cabe hablar de los marginales de una variable, de dos variables y el gran total. El marginal de una variable se calcula sumando las casillas de las categorías de esa variable.

$$f_{0jk} = \sum_{i=1}^{I} f_{ijk}$$

El marginal de dos variables se obtiene sumando, igualmente, las casillas de las categorías de las dos variables

$$f_{00k} = \sum_{i=1}^{I} \sum_{j=1}^{J} f_{ijk}$$

TABLA 2. Cruce de tres variables.

	C_1			C_2		
	B_1	B_2		B_1	B_2	
A_1	f_{111}	f_{121}	f_{101}	f_{112}	f_{122}	f_{102}
A_2	f_{211}	f_{221}	f_{201}	f_{212}	f_{222}	f_{202}
	f_{011}	f_{021}	f_{001}	f_{012}	f_{022}	f_{002}

Y, por último, el gran total se calcula a partir de la suma de todas las casillas

$$f_{000} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} f_{ijk}$$

En la tabla de 2 × 2 × 2 hay un total de 4 marginales de una variable y dos marginales de dos variables. En el caso de la variable *C* éstos serían:

— marginales de una variable: f_{110}, f_{120}, f_{210}, f_{220}
— marginales de dos variables: f_{001}, f_{002}

11.1.2. *Asociación entre variables*

Tomando como referencia los datos de la tabla 3, supongamos que se quiere estudiar la relación entre las dos variables. A tal fin, la simple constatación de los datos que aparecen en la tabla resulta poco informativa.

TABLA 3. Cruce de Permiso de Trabajo con Ingresos.

Ingresos	Permiso No	Sí	
Bajos	197	71	268
Altos	57	71	128
	254	142	396

Si queremos estudiar la posible existencia de asociación entre Ingresos y Permiso de trabajo es conveniente comparar las frecuencias de las casillas con alguna medida o norma. Dos posibilidades son: comparar las frecuencias de las casillas con los marginales o compararlas entre sí. En el primero de los casos lo que se hace es calcular proporciones o porcentajes. Supongamos que se quiera estudiar la posible influencia del Permiso en el hecho de tener ingresos altos. Entre los individuos con permiso, 71/142 = .5 tienen ingresos altos; entre los individuos sin permiso, la proporción es 57/254 = .224. La diferencia entre ambas cantidades (.500 − .224) = .276 puede considerarse como una medida de asociación. Sobre este concepto, diferencia de proporciones, se construye la metodología para el análisis tabular desarrollada por el profesor Davis, a la que tuvimos ocasión de referirnos previamente (véase capítulo 12 de este mismo libro.

Alternativamente al cálculo de proporciones o de porcentajes podemos comparar las frecuencias de las categorías entre sí. Procediendo de esta manera, en lugar de proporciones obtenemos razones *(odds)*. Por ejemplo, la razón de ingresos bajos a altos es igual a 268/128 = 2.09; es decir, por cada individuo con ingresos altos hay 2.09 con

ingresos bajos. A esta razón la llamamos *marginal.* De igual manera podemos calcular las razones *condicionales.* Así, entre los emigrantes sin permiso la razón de ingresos bajos a altos es de 197/57 = = 3.46; esta misma razón entre aquellos que tienen permiso es de 71/71 = 1.00. Vemos, pues, que las razones condicionales difieren, lo cual nos indica que para tener ingresos bajos no es lo mismo la situación legal de cara al trabajo —o, lo que es lo mismo, que están asociadas las variables Ingreso y Permiso. Con objeto de elaborar una medida de la intensidad y de la dirección de la asociación podemos dividir las razones condicionales, calculando la razón de razones *(odds ratio):* 3.46/1.00 = 3.46. Podemos interpretar esta medida diciendo que la razón de ingresos bajos a altos para los que no tienen permiso es casi 3.5 veces superior a la de aquellos con permiso.

La razón de razones será igual a 1.00 cuando las variables sean independientes, y mayor o menor a 1.00, según que la asociación sea positiva o negativa, respectivamente. Esta medida tendrá un límite inferior, el cero, pero carece de límite superior. Este hecho es el que hace que la escala sea diferente según estemos encima o debajo del 1.00 (la independencia). Supongamos que calculásemos ahora la razón de razones de los que Sí tienen permiso versus los que No lo tienen; su valor es 1.00/3.46 = .289. Igual asociación hay si calculamos esta razón que la anterior (aquella de los que No tienen permiso *vs.* los que Sí lo tienen); sin embargo, los números son diferentes (.289 *vs.* 3.46), y también lo es la distancia que les separa del 1.00, pudiendo dar la impresión de que es mayor la intensidad de la razón de No a Sí permiso (3.46) que la de Sí a No (.289). Una forma de resolver este problema es transformar ambos números en sus logaritmos naturales. Así, el $\log_e$ de 3.476 es igual a 1.24; y el $\log_e$ de .289 es -1.24; es decir, los números son iguales sólo que cambian los signos.

Esta medida y su transformación logarítmica van a ser la base sobre las que se construya el método de los modelos lineales logarítmicos (una más amplia exposición de la medida en el libro ya citado de Reynolds, 1977 y en Davis, 1971).

11.2. Dos cuestiones a plantearse

Siempre que observamos los datos de una tabla nos podemos plantear dos preguntas:

1. ¿Qué pasaría si sacásemos otra muestra?, ¿se obtendrían los mismos resultados? ¿Se pueden extrapolar los resultados muestrales al conjunto de la población? La respuesta se encuentra en la realización de un test.

2. ¿Por qué se obtienen estos resultados y no otros? En particular, ¿por qué no son iguales todas las casillas? En este caso la respuesta está en el hecho de que hay diferentes efectos (influencias de cada variable aislada y de sus asociaciones) que causan que los datos sean unos y no otros. Habremos de determinar qué efectos son ésos y cuál es su importancia. Veamos cada uno de los dos problemas.

Con relación a la segunda pregunta, partamos de los datos de la tabla 3, expresados en proporciones sobre el total de casos de la muestra (tabla 4).

TABLA 4. Cruce de Permiso de Trabajo con Ingresos (proporciones sobre el total).

Ingresos	Permiso No	Sí	
Bajos	.50	.18	.68
Altos	.14	.18	.32
	.64	.36	1.00

Frente a estos resultados tenemos la tabla de equiprobabilidad, donde las casillas tienen el mismo valor (tabla 5).

TABLA 5. Cruce de Permiso de Trabajo con Ingresos, en el supuesto de equiprobabilidad.

Ingresos	Permiso No	Sí	
Bajos	.25	.25	.50
Altos	.25	.25	.50
	.50	.50	1.00

Buscando la razón de la variabilidad de nuestra tabla original, diremos que parte de la explicación se encuentra en el hecho de que es más frecuente tener ingresos bajos que altos (.68 *vs.* .32). Si ésta fuera toda la explicación, al multiplicar las filas 1 y 2 de la tabla de equiprobabilidad por .68/.50 y .32/.50, respectivamente, esta tabla debería quedar igual a la original (tabla 4). Veamos en la tabla 6 qué ocurre cuando procedemos así.

TABLA 6. Tabla de equiprobabilidad, ajustando por el efecto de Ingresos.

Ingresos	Permiso No	Sí	
Bajos	.25(.68/.50) = .34	.25(.68/.50) = .34	.68
Altos	.25(.32/.50) = .16	.25(.32/.50) = .16	.32
	.50	.50	1.00

Hecha la multiplicación indicada se observa que todavía siguen siendo diferentes ambas tablas (la 4 y la 6).

Una segunda explicación la hallamos en el hecho de que tampoco es igual de frecuente tener que no tener permiso de trabajo (.64 *vs.* .36). Con objeto de tener en cuenta este efecto multiplicamos las columnas 1 y 2 de la tabla 6 por .64/.50 y .36/.50, respectivamente. En la tabla 7 se muestran los resultados.

TABLA 7. Tabla de equiprobabilidad, ajustando por los efectos de Ingresos y Permiso.

Ingresos	Permiso: No	Permiso: Sí	
Bajos	.34(.64/.50) = .435	.34(.36/.50) = .245	.68
Altos	.16(.64/.50) = .205	.16(.36/.50) = .115	.32
	.640	.360	1.00

Aun así, los datos de esta tabla difieren de los obtenidos en la tabla original (tabla 4). Ello es debido a que las probabilidades de tener (ingresos bajos - sin permiso) e (ingresos altos - con permiso) son mayores de lo que cabría esperar en el supuesto de que las dos variables fueran independientes; siendo esto así, habrá que añadir el efecto que tiene la asociación entre las dos variables para terminar de explicar por qué, aún después de considerar los efectos de filas y columnas, las tablas 4 y 7 siguen siendo diferentes.

Resumiendo, vemos que hay 4 efectos, los ya señalados más otro que tiene en cuenta el tamaño de las casillas, que explican los resultados de nuestra tabla original:

1. Efecto de las filas.
2. Efecto de las columnas.
3. Efecto debido a la asociación entre las variables.
4. Efecto debido al número medio de casos en cada casilla.

Estos 4 efectos los denominaremos, respectivamente, λ_i^A, λ_j^B, λ_{ij}^{AB}, μ, y serán los que expliquen, en lo que vamos a llamar el *modelo aditivo logarítmico lineal,* el logaritmo en base *e* de la frecuencia *esperada* de cada casilla.

$$\log_e F_{ij} = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB} \qquad [1]$$

La primera pregunta tiene contestación cuando se realiza un test, en este caso de la Gi-cuadrado, que nos indica si la relación que se observa en la muestra es estadísticamente significativa o, por el contrario, cabe atribuirla al azar. No vamos a detenernos ahora a explicar el test de la Gi-cuadrado —para repasar el tema véase, por ejemplo, Wonnacott y Wonnacott (1977)—; lo que sí vamos a hacer es introducir algunas ideas que no son tan comunes como la realización del propio test en la si-

tuación standar en la que tenemos dos variables y hacemos el supuesto de independencia. En primer lugar veamos los posibles tests a realizar. Dada una tabla de dos dimensiones, además del supuesto de independencia se pueden hacer las hipótesis de que no hay efectos de las filas y/o de las columnas —en general se pueden hacer hipótesis que pongan en cuestión la significatividad de todos los efectos que influyen en la tabla, cualquiera que sea su dimensión.

La segunda aclaración tiene que ver con el cálculo de las frecuencias esperadas —y a partir de ellas la estimación de los parámetros del modelo bajo la hipótesis hecha en el test—. Cuando la tabla tiene más de dos dimensiones no hay una fórmula que permita calcular directamente el valor de las frecuencias esperadas; debemos de utilizar algún algoritmo. Una posibilidad es la utilización del «escalamiento proporcional iterativo» (Haberman, 1972). El algoritmo y su ilustración con un ejemplo no lo vamos a mostrar aquí, y remitimos a las obras de Bishop y otros (1975), Davis (1974) y Nigel Gilbert (1981). Digamos simplemente que el método genera una tabla en la que los marginales correspondientes a las relaciones incluidas en el modelo se hacen iguales a los de la tabla observada, mientras que los otros quedan libres, siendo las frecuencias estimadas para cada casilla máximo-verosimilitud. Este es el algoritmo que utilizan los programas informáticos BMDP (Dixon, 1978) y ECTA (Everyman's Contingency Table Analysis) Goodman y Fay, 1973). Otro procedimiento que también genera estimaciones máximo-verosimilitud se incluye en GLIM (Generalized Linear Interactive Modelling). El procedimiento ha sido desarrollado por Nelder y Wedderburn (1972), y se conoce como «mínimos cuadrados ponderados iterativos». (Se puede ver una revisión del problema de la estimación en los modelos lineales logarítmicos en Payne, 1977 y Upton, 1978).

Otro aspecto a explicar son los grados de libertad de las tablas que se contrastan en el test. Los grados de libertad son igual al número de casillas menos el número de efectos que requieren estimación. Cuando se trata de contrastar la hipótesis de independencia en la tabla de I × J, el número de efectos a estimar es de $1 + (I - 1) + (J - 1) = I - J - 1$. Veamos por qué:

— en el cálculo del efecto medio del tamaño de las casillas hay que estimar un parámetro independiente

— dada la constricción de que la suma de los efectos de las filas ha de ser igual a cero, el número de efectos fila independientes a estimar será de $(I - 1)$,

— por la misma razón anterior, el número de efectos columna independientes a estimar será igual a $(J - 1)$.

Si la tabla es de 2 × 2, el número de efectos será igual a 3, y los grados de libertad $(4 - 3) = 1$. En el caso de la hipótesis de asociación, a los efectos anteriores habría que añadir el efecto asociación. El número de efectos independientes a estimar debido a la asociación es de $(I - 1)(J - 1)$. En la tabla 2 × 2 el número de efectos total sería de 4 y los grados de libertad $4 - 4 = 0$.

Por último, señalemos que vamos a utilizar como alternativa al estadístico Gi-cuadrado (χ^2) el ratio de verosimilitud (Y^2):

$$Y^2 = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} f_{ij} \log_e (f_{ij} / F_{ij})$$

donde f_{ij} es la frecuencia observada en la fila i y la columna j, y F_{ij} es la frecuencia esperada bajo la hipótesis del test, en las mismas fila y columna.

La distribución muestral de la Y^2 es la misma que la de la X^2, pero la primera tiene algunas ventajas de las que la última carece (ver infra). Su valor es muy semejante. Para los datos de la tabla 3, después de calcular las frecuencias esperadas en el supuesto de independencia (véase tabla 8) vemos cómo χ^2 es aproximadamente igual a Y^2:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} (f_{ij} - F_{ij})^2/F_{ij} = (197 - 171.9)^2/171.9 + (71 - 96.1)^2/96.1 + \\ + (57 - 82.1)^2/82.1 + (71 - 45.9)^2/45.9 = 31.62 \qquad [2]$$

$$Y^2 = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} f_{ij}\log_e(f_{ij}/F_{ij}) = 2(197(.136) + 71(-.303) + \\ + 57(-.364) + 71(.436)) = 2(15.487) = 30.974 \qquad [3]$$

11.3. Concepto de modelo

Siguiendo en este punto la excelente y clara explicación que hace del tema Nigel Gilbert (1981), vamos a llamar modelo a un conjunto de hipótesis que intentan explicar las interrelaciones que se producen entre los fenómenos sociales. Los modelos se componen de conceptos y de sus relaciones. Si estuviéramos estudiando los Ingresos de los Sudamericanos en España, partiríamos de la variabilidad que hay entre el dinero que ganan unos y otros y trataríamos de encontrar una explicación a este fenómeno. Un modelo muy simple podría señalar que Ingreso y Permiso de Trabajo están relacionados, formulando la hipótesis, con la confianza de poder rechazarla, de que ambas variables son independientes. Este sería nuestro modelo, que incluiría dos conceptos (Ingresos y Permiso) y un supuesto de no asociación entre ambos. De acuerdo con el modelo construiríamos un mundo a su semejanza (un mundo en el que ambas variables fueran independientes) y estudiaríamos qué cabe esperar que les ocurriera a los individuos en esta situación. Si el modelo fuera correcto, los resultados observados en la realidad y los derivados de nuestro modelo —generados mediante alguna técnica analítica, por ejemplo el escalonamiento proporcional iterativo— deberían de coincidir. Caso contrario habremos de buscar otro modelo que sí produzca esta similaridad —para este ejemplo, todos los modelos alternativos, contrastados, están recogidos en el apartado siguiente. La búsqueda se puede realizar mediante un proceso de abstracción guiado por la teoría o mediante el recurso a algún procedimiento analítico (véase *infra*). Seleccionado el modelo correcto, sólo restaría calcular las lambdas estimadas, con objeto de ver la importancia relativa de los diferentes efectos que influyen en la tabla. En el gráfico 1, tomado de Nigel Gilbert, se esquematiza el procedimiento que acabamos de explicar.

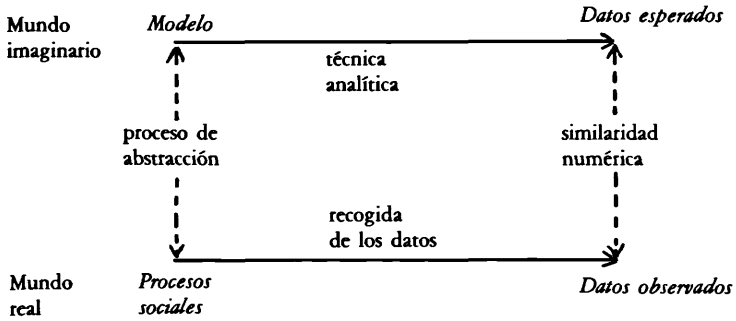


GRÁFICO 1. Diagrama esquemático que ilustra la relación entre los mundos «real» e «imaginario» (tomado de Nigel Gilbert, 1981, pág. 5).

De este gráfico se derivan las etapas a seguir en el ajuste de los modelos y los problemas que se plantean. Resumiendo muy brevemente:

a) Selección del modelo.

b) Generación de los datos esperados, en función del modelo seleccionado.

c) Comparación de los datos esperados con los datos observados, mediante la realización de un test.

d) Si los dos conjuntos de datos no son suficientemente similares se repiten los tres pasos previos con otro modelo diferente.

e) Si los dos conjuntos de datos son similares, es decir, si el modelo es correcto, se puede ver la posibilidad de simplificarlo; ello supone seguir los pasos previos buscando ese modelo más parsimonioso.

f) Llegados al modelo final, calculamos los parámetros del modelo para ver su importancia relativa.

11.3.1. *El modelo saturado para tablas 2 × 2*

Llamamos modelo saturado a aquel que incluye tantos parámetros como efectos influyen en la tabla. Si cuando hablábamos de la tabla de dos dimensiones veíamos que había 4 posibles efectos que podían explicar la variabilidad de sus casillas, el modelo que incluya los 4 parámetros correspondientes a estos efectos le llamaremos saturado. Tal como explicábamos (eq. 1), los parámetros a calcular en este modelo son μ, λ_i^A, λ_j^B, λ_{ij}^{AB}. Veamos el cálculo de cada uno de ellos utilizando los datos de la tabla 3.

a) Efecto medio

$$\mu = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \log_e F_{ij}$$

y sustituyendo los valores correspondientes,

$$\mu = 1/4(\log_e 197 + \log_e 71 + \log_e 57 + \log_e 71) = 1/4(17.84) = 4.46$$

b) Efecto de las filas

$$\lambda_i^A = 1/IJ \sum_{j=1}^{J} \log_e(F_{ij}/F_{mo\ i,j})$$

Tomando el efecto de la fila 1 (variable Ingresos)

$$\lambda_1^A = 1/4(\log_e 197/57 + \log_e 71/71) = 1/4(1.24) = .31$$

c) Efectos de las columnas

$$\lambda_j^B = 1/IJ \sum_{i=1}^{I} (\log_e F_{ij} + \log_e F_{ij}/F_{i,mo\ j})$$

Para la columna 2 (variable Estudios):

$$\lambda_2^B = 1/4(\log_e 71/197 + \log_e 71/57) = 1/4(-.8) = -.2$$

Las lambdas están sometidas a la siguiente constricción:

$$\sum_{i=1}^{I} \lambda_i^A = 0 \quad ; \quad \sum_{j=1}^{J} \lambda_j^B = 0$$

d) Efecto debido a la asociación

$$\lambda_{ij}^{AB} = 1/IJ\log_e(F_{ij}F_{mo\ i,mo\ j}/F_{i,mo\ j}F_{mo\ i,j})$$

para la casilla (1,1)

$$\lambda_{11}^{AB} = 1/4\log_e(197(71)/57(71)) = 1/4\log_e 3.456 = .310$$

En este caso la constricción es que

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \lambda_{ij}^{AB} = 0$$

lo que implica que

$$\lambda_{22}^{AB} = -\lambda_{12}^{AB} = -\lambda_{21}^{AB} = \lambda_{11}^{AB}$$

Veamos ahora si nuestro modelo permite predecir los valores obtenidos, tomando como ejemplo la casilla (1,1)

$$\log_e F_{11} = \mu + \lambda_1^A + \lambda_1^B + \lambda_{11}^{AB}$$
$$\log_e 197 = 4.46 + .31 + .2 + .31 = 5.28$$

Si calculamos el $\log_e$ de 197 (frecuencia de la casilla 1,1) vemos que su valor coincide con el 5.28 previamente obtenido —lo cual era lo previsto, puesto que el modelo saturado siempre reproduce con exactitud los datos originales.

Los resultados obtenidos en el cálculo de cada efecto nos indican que las influencias más importantes en la casilla (1,1) se deben por igual al hecho de que los individuos se distribuyen desigualmente en la variable Ingresos y a la asociación entre Ingresos y Permiso. Menor importancia tiene la desigual distribución de emigrantes con y sin permiso de trabajo.

Antes de pasar a otros modelos alternativos queremos mencionar un punto práctico de interés. Caso de que haya casillas con frecuencia cero se plantea la imposibilidad de ajustar el modelo saturado y se crean problemas con el resto de los modelos. Tal como hemos mostrado, las lambdas son combinaciones lineales de los logaritmos de las frecuencias de las casillas. Dado que el logaritmo de cero es $-\infty$, es recomendable añadir a cada casilla 1/2 antes de ajustar el modelo saturado (Goodman, 1970). Esta solución es válida siempre y cuando no haya una razón teórica que justifique la existencia de los ceros —si se cruzan el sexo y el tipo de enfermedades padecidas por los individuos, lógicamente la casilla varón y cáncer de mama estará vacía. En este caso el tipo de análisis que estamos explicando hay que complementarlo con los modelos de casi-independencia y las tablas incompletas (Upton, 1978: 102-117; Fienberg, 1977: 108-129; Knoke y Burke, 1980: 63-67).

11.3.2. *Otros modelos para las tablas de 2 × 2*

Hemos visto que el modelo saturado reproduce exactamente el valor de las casillas de la tabla, pero ello era a costa de incluir todos los posibles efectos que la afectan. En la medida en que nuestro interés consiste en encontrar la pauta de relaciones más sencilla que explique los valores obtenidos, lo que haremos es buscar otros modelos más parsimoniosos que cumplan esta función. El primer modelo que probaremos será el modelo de independencia. Este modelo incluye los parámetros:

$$\log_e F_{ij} = \mu + \lambda_i^A + \lambda_j^B$$

Si este modelo fuese correcto, las frecuencias esperadas (F_{ij}) de cada casilla serían $F_{ij} = F_{i0}F_{0j}/F_{00}$. De esta forma estimamos las frecuencias esperadas y a partir de ellas calculamos los parámetros de nuestro modelo —parámetros estimados por tratarse de datos muestrales. En la tabla 8 se ofrecen las frecuencias esperadas de éste y de los otros modelos alternativos.

Antes de proceder al cálculo de los parámetros es necesario contrastar la bondad del modelo. Para ello calculamos el ratio de verosimilitud (Y^2), haciendo un test con

1 grado de libertad y en el que la hipótesis nula es que hay independencia entre ambas variables. La ecuación 3 nos da el resultado del cálculo de la Y^2. La probabilidad de obtener al azar una $Y^2 = 30.97$, con 1 grado de libertad, bajo el supuesto de independencia, es de .0000, por lo que rechazamos este modelo.

Rechazada la bondad del modelo anterior resulta improcedente calcular el valor de sus parámetros. También resulta improcedente estudiar la posible bondad de modelos alternativos con menos efectos. A pesar de ello, con objeto de poder contrastar los resultados que se pueden obtener, incluimos en la tabla 8 los parámetros, frecuencias esperadas y ratios de verosimilitud correspondientes a los diferentes modelos a ajustar en una tabla de dos dimensiones.

TABLA 8. Resultados obtenidos al ajustar todos los modelos posibles para los datos de la tabla 3.

Modelos	F_{11}	F_{12}	F_{21}	F_{22}	λ_1^A	λ_1^B	λ_{11}^{AB}	μ	$g.l.$	Y^2
Saturado	197	71	57	71	.31	.2	.31	4.46	0	0
Independencia	171.9	96.1	82.1	45.9	.369	.291	—	4.486	1	30.97
Efecto B nulo	134	134	64	64	.369	—	—	4.529	2	63.18
Efecto A nulo	127	71	127	71	—	.291	—	4.553	2	81.63
Equiprobabilidad	99	99	99	99	—	—	—	4.595	3	113.66

Según cada modelo, las frecuencias esperadas son distintas, por lo que también los parámetros tienen valores distintos. A medida que suprimimos efectos la Y^2 aumenta, como indicador de que la bondad del modelo disminuye.

11.3.3. *Test de la importancia de los parámetros*

Al hablar del ratio de verosimilitud decíamos que este estadístico tiene algunas ventajas de las que carece la χ^2. En concreto, el ratio de verosimilitud es aditivo ante su partición para modelos anidados *(nested models)*. Diremos que dos modelos M_1 y M_2 son anidados si todos los efectos (la lambdas) contenidos en M_1 son subconjuntos de los efectos contenidos en M_2. La diferencia en Y^2 entre los dos modelos sirve como test de los efectos adicionales en M_2, condicional de los efectos en M_1. Esta diferencia también tiene una distribución χ^2, siendo sus grados de libertad igual a la diferencia en el número de parámetros ajustados en los dos modelos. Esta propiedad no se mantiene para la χ^2 de Pearson; es por ello que elegimos el ratio de verosimilitud para los tests.

Partiendo de esta propiedad de la Y^2 podemos ver la importancia de cualquier parámetro haciendo el test correspondiente. Supongamos que tenemos los dos modelos siguientes:

Modelos	g.l.	Y^2
M_1	g_1	Y_1^2
M_2	g_2	Y_2^2
Parámetro extra en M_1	$(g_2 - g_1)$	$(Y_2^2 - Y_1^2)$

Podemos hacer un test para ver si la diferencia entre los dos modelos es significativa —es decir, ver si el parámetro extra en M_1 es significativo. Para ello hacemos un test en el que la hipótesis nula (H_0) sea que el parámetro extra es igual a cero, para una Y^2 igual a $(Y_2^2 - Y_1^2)$, con $(g_2 - g_1)$ grados de libertad. Veamos un ejemplo en el que M_1 es el modelo de independencia y M_2 el de efecto Permiso nulo.

Modelos	g.l.	Y^2
Independencia (μ, λ_i^A, λ_j^B)	1	30.97
Efecto Permiso nulo (μ, λ_i^A)	2	63.18
λ_j^B	1	32.21

En este caso haríamos un test para ver si el efecto λ_j^B es significativo. Una $Y^2 = 32.21$, con 1 grado de libertad, es significativa al nivel del .0000, por lo que concluimos que no se puede omitir λ_j^B sin que el modelo sea peor.

11.3.4. *Modelos jerárquicos*

Todos los modelos que vamos a considerar son miembros de una clase de modelos llamados jerárquicos. Con palabras de G. Upton (1978: 57) estos modelos obedecen la siguiente regla. Supongamos que el parámetro relacionado con un conjunto de variables S se incluye en el modelo; entonces, el modelo debe de incluir todos los parámetros relacionados con cualquier subconjunto de S. Esto es lo que hemos hecho en la tabla 2 × 2. Cuando incluíamos el parámetro relacionado con el conjunto de variables A y B (λ_{ij}^{AB}) también incluíamos los parámetros relacionados con los subconjuntos de S, A (λ_i^A) y B (λ_j^B).

11.4. Tablas multidimensionales

En la tabla de dos dimensiones tan sólo cabía la posibilidad de asociación o de independencia entre las dos variables. Cuando introducimos nuevas variables los tipos de posibles relaciones aumentan. Para ilustrar los nuevos tipos de hipótesis vamos a utilizar una tabla de tres dimensiones, que incluye las variables A, B y C.

a) Hipótesis de Independencia Mutua

Las tres variables son independientes: A es independiente de B y C, y B es independiente de C. Bajo esta hipótesis,

$$P_{ijk} = P_{ioo}P_{ojo}P_{ook} \qquad \text{donde} \qquad P = \text{probabilidad}$$

El modelo correspondiente a esta situación sería aquel que solo incluyera los efectos de cada variable más el efecto medio

$$\log_e F_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C$$

b) Hipótesis de Independencia Condicional

Existe independencia condicional entre A y B cuando ambas variables son independientes para cada categoría de C, aun cuando A y B estén asociada cuando plegamos C.

En general A y B son condicionalmente independientes, dadas las categorías de C, si

$$P_{ijk} = P_{iok}P_{ojk}/P_{ook}$$

Esta hipótesis corresponde al modelo lineal logarítmico

$$\log_e F_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ik}^{AC} + \lambda_{jk}^{BC}$$

c) Hipótesis de Independencia Múltiple

Si A y B tienen la misma pauta de asociación para cada categoría de C, se puede decir que A y B son independientes de C. Bajo esta hipótesis tenemos que,

$$P_{ijk} = P_{ijo}P_{ook}$$

El modelo correspondiente a esta hipótesis es

$$\log_e F_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB}$$

d) Hipótesis de Interacción

Decimos que hay interacción cuando las tres variables están relacionadas de tal manera que la asociación entre dos de ellas cambia con cada nivel de la tercera. El modelo lineal logarítmico correspondiente a esta situación es el siguiente:

$$\log_e F_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}$$

Este sería el modelo saturado para la tabla de tres dimensiones. Añadiendo las asociaciones e interacciones correspondientes se pueden generalizar estas hipótesis para

tablas de cualquier número de variables. Por ejemplo, el modelo saturado para la tabla de cuatro dimensiones incluiría los efectos atribuibles a los marginales de cada variable, las asociaciones mutuas entre las 4 variables, las 4 interacciones de tercer orden y la interacción de cuarto orden. Así,

$$\log_e F_{ijkl} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_l^D + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{il}^{AD} + \lambda_{jk}^{BC} + \lambda_{jl}^{BD} + \lambda_{kl}^{CD} +$$

$$+ \lambda_{ijk}^{ABC} + \lambda_{ijl}^{ABD} + \lambda_{ikl}^{ACD} + \lambda_{jkl}^{BCD} + \lambda_{ijkl}^{ABCD}$$

11.4.1. *Selección del modelo*

En el caso de que la tabla sea de dos dimensiones, la selección del modelo no resulta complicada, dado que sólo hay que elegir entre 5 modelos distintos. Cuando las variables son tres, el número de modelos alternativos sube hasta 18, por lo que ya la selección se hace más compleja. En el caso de una tabla de 4 dimensiones, al existir 144 modelos no saturados distintos, la selección se haría imposible si no tuviéramos alguna estrategia que, sin necesidad de probarlos todos, nos llevara a aquel que mejor se ajusta a los datos, manteniendo el principio de parsimonia —por supuesto, más difícil será la selección cuando el número de variables se incremente a 5, 6, etc. Hay diferentes estrategias en la selección del modelo adecuado. Brown (1976) sugiere el uso de un procedimiento que denomina *screening*. Alternativamente, otro procedimiento ampliamente utilizado consiste en seleccionar nuestro modelo partiendo del saturado. Este segundo procedimiento es el que vamos a explicar aquí, remitiendo al autor citado o a otros trabajos como el de G. Upton (1980) para ver el procedimiento del *screening*.

En el segundo procedimiento partimos del ajuste del modelo saturado para los datos objeto de estudio, buscando las lambdas estandarizadas cuyo valor sea igual o superior a ± 2. Una primera aproximación al modelo que estamos buscando será aquel que incluya todos estos efectos. La justificación de este procedimiento se encuentra en el hecho de que las lambdas que nosotros obtenemos son estimadores de las lambdas poblacionales. Cuando el tamaño de la muestra sea grande y bajo la hipótesis de que la lambda verdadera es igual a cero, la razón (lambda estimada/error estándar) tiene una distribución aproximadamente normal con media cero y varianza uno. En esta situación un valor de la lambda estimada mayor que (± 2) es significativa al nivel del .05.

Una vez obtenido este modelo de partida procedemos a buscar otro que mejore el ajuste previamente obtenido. A semejanza con la regresión de pasos procedemos, 1) a eliminar aquellos parámetros que nos lleven a un modelo que ajuste los datos que sea más parsimonioso que el anterior, y 2) a añadir parámetros que cuando se incluyan en el modelo de partida proporcionen una mejora significativa del ajuste.

A continuación incluimos un ejemplo que ilustra el procedimiento mencionado, junto a los principios discutidos hasta este momento; para ello utilizamos los datos de la tabla 1. En este ejemplo ilustrativo analizamos los datos en cuestión, estudiando la pauta de relaciones entre las variables (modelo general lineal logarítmico); la influencia de Estudios, Antigüedad y Permiso de Trabajo (variables independientes) sobre los

Ingresos (variable dependiente) (modelo logit); y el estudio de las relaciones causales entre todas las variables, mediante la construcción de un modelo causal.

11.5. Modelo general lineal logarítmico

Con objeto de encontrar el mejor modelo para nuestros datos hemos seguido la estrategia de ajustar el modelo saturado, buscando las lambdas estandarizadas cuyos valores sean iguales o superiores a ± 2. Procediendo de tal manera y utilizando el programa BMDP obtenemos los resultados de la tabla 9.

TABLA 9. Valores estandarizados de las estimaciones.

I	2.755	EP	2.039
P	3.225	EA_1	− .588
A_1	−6.561	EA_2	−1.088
A_2	5.617	EA_3	1.811
A_3	3.477	A_1PI	.218
E	− .491	A_2PI	−1.280
PI	4.255	A_3PI	.914
A_1I	−3.300	EPI	− .423
A_2I	.163	EA_1I	.326
A_3I	4.257	EA_2I	− .820
EI	2.159	EA_3I	− .544
A_1P	1.130	EA_1P	.704
A_2P	−2.094	EA_2P	− .820
A_3P	.463	EA_3P	− .169
		EA_1PI	− .937
		EA_2PI	2.064
		EA_3PI	− .692

De acuerdo con los resultados de la tabla 9 elegimos como candidato a mejor modelo aquel que incluye las lambdas correspondientes a las asociaciones Educación-Permiso de Trabajo (a partir de aquí EP), Antigüedad-Permiso (AP), Educación-Ingresos (EI), Antigüedad-Ingresos (AI) y Permiso-Ingresos (PI). Siguiendo nuestro criterio de selección deberíamos de incluir también la interacción entre las 4 variables (IPAE), dado que su valor estandarizado es 2.064. No procedemos así, asumiendo que ésta es una de cada 20 veces que se obtiene este valor al azar. De acuerdo con la argumentación de Upton en uno de los ejemplos de su libro (1978, p. 77), suponemos que si IPAE fuera una interacción verdadera habríamos encontrado interacciones de tercer orden significativas; mirando a los valores correspondientes vemos que no es éste el caso, las 4 interacciones de tercer orden están lejos de ser significativas.

El próximo paso en nuestro análisis es ajustar el modelo EP, AP. EI, AI, PI, para ver si se trata de un modelo correcto. La tabla siguiente ofrece los resultados:

Parámetros	*g.l.*	Y^2	*P*	χ^2	*P*
EP, AP, EI, AI, PI	11	12.19	.3495	12.27	.3435

En función de los resultados del test vemos que el modelo ajusta bien: $Y^2 = 12.19$, con 11 grados de libertad, es un valor típico para una observación de una distribución χ^2. Hemos llegado a un modelo simplificado que explica nuestros datos sin considerar la asociación entre Educación y Antigüedad, las 4 interacciones de tercer orden y la interacción de cuarto orden. Sin embargo, todavía quedan 9 parámetros y vamos a ver si podemos encontrar un modelo más simple. La tabla 10 muestra los resultados al suprimir de una en una cada una de las 5 asociaciones del modelo.

TABLA 10. Eliminación de las 5 asociaciones.

Modelo	Parámetros	*g.l.*	Y^2	Parámetro a prueba	Resultado de la prueba
1	EP, AP, EI, AI, PI	11	12.19	Modelo	Modelo ajusta bien
2	AP, EI, AI, PI	12	17.53	EP	Significativo ($P = .0208$)
3	EP, EI, AI, PI	13	18.53	AP	Significativo ($P = .0421$)
4	EP, AP, AI, PI	12	20.43	EI	Significativo ($P = .0041$)
5	EP, AP, EI, PI	13	37.73	AI	Significativo ($P = .0000$)
6	EP, AP, EI, AI	12	33.46	PI	Significativo ($P = .0000$)

Si suprimimos el parámetro EP, ajustando el modelo 2, vemos que el modelo no es significativo —la probabilidad de obtener al azar una $Y^2 = 17.53$, con 12 grados de libertad, bajo la hipótesis de que el modelo es correcto, es .1307—; sin embargo, éste no es el caso del parámetro bajo test, que es significativo al nivel del .05 ($P = 0.208$). Esto significa que no podemos suprimir la asociación entre Educación y Estudios. A continuación suprimimos el resto de los parámetros incluidos en el modelo 1. Tal como vemos en la tabla 10 todos los parámetros que contrastamos son significativos, especialmente aquellos que relacionan los Ingresos de los emigrantes con las otras tres variables. Debemos concluir que no se puede suprimir ningún parámetro de nuestro modelo original, sin que haya un empeoramiento significativo del ajuste.

Después de considerar la eliminación de algunos parámetros estudiamos la posibilidad de mejorar el ajuste mediante la inclusión de nuevos efectos. La tabla 11 muestra la selección de los parámetros.

TABLA 11. Selección de los parámetros.

Modelo	Parámetros	*g.l.*	Y^2	Parámetro a prueba	Resultado de la prueba
7	AE, EP, AP, EI, AI, PI	9	9.46	AE	No significativo (P = .3962)
8	PAE, EI, AI, PI	7	8.11	EP	No significativo (P = .3952)
9	EPI, AP, AI	10	12.11	EPI	No significativo (P = .7913)
10	API, PE, IE	9	9.12	IPA	No significativo (P = .2554)
11	EAI, EP, AP, PI			EAI	No significativo

El primer candidato a ser elegido es la asociación entre Antigüedad y Educación. Hacemos un test de la diferencia entre los dos modelos (1 y 7), con 2 grados de libertad y una Y^2 = 2.73, bajo la hipótesis de que AE = 0, y vemos que este parámetro no es significativo (P = .3962), llegando a la conclusión de que no hay una mejora significativa del ajuste por el hecho de incluir este parámetro. Obviamente, los siguientes candidatos son las 4 interacciones de tercer orden. A partir de los resultados de la tabla 11 concluimos que ninguno de ellos es significativo. Poniendo juntos los resultados de la eliminación y selección de parámetros concluimos que el modelo 1 es el más parsimonioso. Para darse una idea de la bondad del modelo ofrecemos en la tabla 12 las frecuencias estimadas de cada casilla.

TABLA 12. Frecuencias esperadas según el modelo EP, AP, EI, AI, PI.

			Ingresos	
Estudios	Antigüedad	Permiso	Bajos	Altos
Inferiores	1960-70	No	10.0	6.5
Inferiores	1960-70	Sí	2.5	4.8
Inferiores	1971-77	No	41.2	11.3
Inferiores	1971-77	Sí	16.7	7.0
Inferiores	1978-80	No	72.2	8.7
Inferiores	1978-80	Sí	16.4	5.9
Superiores	1960-70	No	6.0	7.5
Superiores	1960-70	Sí	2.5	9.2
Superiores	1971-77	No	24.6	13.0
Superiores	1971-77	Sí	16.6	26.1
Superiores	1978-80	No	43.0	10.0
Superiores	1978-80	Sí	16.3	11.3

Tal como se puede ver, no hay grandes diferencias entre los valores observados y los estimados —tal como sabíamos a partir de la realización del test de la χ^2.

Una vez que hemos llegado al modelo final hay que calcular los valores de los parámetros con objeto de estudiar su importancia relativa. En la tabla 13 incluimos los parámetros estimados bajo el modelo EP, EI, AP, AI, PI. Damos las lambdas y sus exponenciales, las taus.

TABLA 13. Parámetros estimados bajo el modelo EP, AP, EI, AI, PI.

Efecto	λ	τ
Gran media	2.471	11.84
I	.203	1.225
P	.229	1.257
A_1	− .772	.462
A_2	.454	1.575
A_3	.318	1.374
E	− .032	.968
PE	.128	1.136
PA_1	.067	1.069
PA_2	− .177	.837
PA_3	.111	1.117
IE	.163	1.117
IA_1	− .423	.655
IA_2	.007	1.007
IA_3	.416	1.516
IP	.273	1.314

De acuerdo con los parámetros estimados es posible predecir las frecuencias esperadas de las casillas. Los parámetros incluidos en el modelo son aquellos necesarios para explicar los valores de las casillas, y sus tamaños indican la importancia de cada uno de ellos. El modelo 1 señala que el logaritmo de la probabilidad de la casilla (*i, j, k, l*) viene dado por

$$\log_e F_{ijkl}^{PAE} = \mu + \lambda_i^I + \lambda_j^P + \lambda_k^A + \lambda_l^E + \lambda_{ij}^{IP} + \lambda_{ik}^{IA} + \lambda_{il}^{IE} + \lambda_{jk}^{PA} + \lambda_{jl}^{PE}$$

Esta es la forma aditiva de nuestro modelo general logarítmico lineal[2]. Si queremos estimar el logaritmo de la probabilidad de las casillas que requieren parámetros no incluidos en la tabla 13 (columna de las lambdas), su valor es el mismo que el de aquellos que sí aparecen en la tabla, viniendo su signo determinado por los

[2] Si quisiéramos expresar nuestro modelo en función de las frecuencias de las casillas, en vez de sus logaritmos, se pueden cambiar las lambdas por sus exponenciales. Llamando η (eta) al exponencial de μ y τ(tau) a los exponenciales de las lambdas, tendríamos el siguiente modelo multiplicativo:

$$F_{ijkl}^{PAE} = \eta\tau_i^I\tau_j^P\tau_k^A\tau_l^E\tau_{ij}^{IP}\tau_{ik}^{IA}\tau_{il}^{IE}\tau_{jk}^{PA}\tau_{jl}^{PE}$$

subíndices. Por cada número 2 que aparece en el subíndice hay que multiplicar el valor dado por (−1). Por ejemplo $\lambda_{13}^{IA} = .416$, mientras que $\lambda_{23}^{IA} = .416(-1) = -.416$. De acuerdo con esta explicación, el logaritmo de la probabilidad de la casilla (1, 2, 3, 1) se estima de la forma que sigue:

$$\log_e F_{1231}^{IPAE} = 2.471 + .203 + (-.229) + .318 + (-.032) + (-.273) +$$

$$+ .416 + .163 + (-.111) + (-.128) = 2.798$$

Y la frecuencia estimada es exp(2.798) = 16.41, que es el número de la casilla (1, 3, 2, 1) en la tabla 12.

Hasta el momento hemos explicado la forma de seleccionar el modelo, veamos ahora su interpretación. En función de nuestros resultados podemos decir que no hay interacciones de tercer o de cuarto orden. Todo lo que tenemos en nuestro modelo son las asociaciones entre (en orden decreciente de importancia):

— IA: Ingresos y Antigüedad en el país. Los sudamericanos que llegaron a España en el período 1978-80 es más probable que tengan ingresos bajos (.416). Aquellos que llegaron en los años 60 son los que tienen mayores ingresos (−.423).

— IP: Ingreso y Permiso de trabajo. Aquellos que no tienen permiso de trabajo tienen ingresos bajos (.273).

— IE: Ingresos y Educación. Tener estudios inferiores significa tener ingresos bajos (.163).

— PE: Permiso y Educación. Es más fácil obtener permiso de trabajo si se dispone de un nivel superior de estudios (.128).

— PA: Permiso y Antigüedad. Los que llegaron entre 1978-80 es más probable que no tengan permiso de trabajo (.111).

Junto con las 5 asociaciones debemos considerar los marginales de cada variable. Los marginales se interpretan como el efecto de una distribución desigual de los entrevistados en cada variable. Y vemos que la influencia más importante proviene de los marginales de la variable Antigüedad.

11.6. Modelo Logit

Una vez que hemos visto la pauta de relaciones podemos estar interesados en el estudio de los efectos de un conjunto de *factores* (variable independientes) sobre otra variable considerada como *respuesta* (variable dependiente), tal como hacemos en la regresión múltiple —Goodman denomina al modelo logit «enfoque modificado de la regresión» (Goodman, 1972). Vamos a considerar el Ingreso de los emigrantes como respuesta, estudiando la influencia de Antigüedad, Educación y Permiso. En este caso, en lugar de explicar los logaritmos de las casillas lo que se explica son las razones condicionales (o los logaritmos de las razones condicionales) de la variable dependiente.

Cuando se ajusta un modelo *logit* se consideran todas las asociaciones e interacciones significativas que incluyen la variable dependiente, junto con la interacción

entre las variables independientes. Esta inclusión es la diferencia principal entre el procedimiento de estimación de los modelos logarítmicos lineales y los modelos logit. Tal como Knoke y Burke (1980: 26) explican, lo que interesa es encontrar qué efectos de los que incluyen la variable dependiente son significativos; y para llegar a este punto se utiliza un modelo base que asume que todos estos efectos son igual a cero —éste es el caso cuando se tiene un modelo con dos parámetros: la variable dependiente y la interacción entre las independientes.

En el ejemplo que estamos siguiendo, el modelo logit saturado tendría esta forma:

$$\log_e \frac{F_{1jkl}}{F_{2jkl}} = 2\lambda^{I} + 2\lambda^{IP} + 2\lambda^{IA} + 2\lambda^{IE} + 2\lambda^{IPA} + 2\lambda^{IPE} + 2\lambda^{IAE} + 2\lambda^{IPAE}$$

y suprimiendo los logaritmos y utilizando la notación de Goodman (1972), tenemos

$$\Phi_{ijkl} = \beta^{I} + \beta^{IP}_{ij} + \beta^{IA}_{ik} + \beta^{IE}_{il} + \beta^{IPA}_{ijk} + \beta^{IPE}_{ijk} + \beta^{IAE}_{ikl} + \lambda^{IPAE}_{ijkl}$$

donde Φ_{ijkl} es el logaritmo de la razón condicional de Ingresos, y las β corresponden a las λ (ejemplo, $\beta^{IP}_{ij} = 2\lambda^{IP}_{ij}$).

Con objeto de estimar los parámetros para la ecuación procedemos de igual manera que en el modelo general logarítmico lineal. Ajustamos el modelo saturado, buscando las lambdas que incluyan la variable dependiente (Ingresos) con un valor estandarizado igual o mayor a ± 2. Si miramos la tabla 9 vemos que éstos son λ^{I}_{i}, λ^{IP}_{ij}, λ^{IA}_{ik}, λ^{IE}_{il}. Por la misma razón ofrecida anteriormente no incluimos λ^{IPAE}_{ijkl}. Ahora se ajusta el modelo IP, IA, IE, PAE, obteniendo una $Y^2 = 8.11$, con 7 grados de libertad. Este valor no es significativo ($P = .3226$). Este modelo da las frecuencias esperadas de la tabla 14.

TABLA 14. Frecuencias esperadas según el modelo IP, IA, IE, PAE.

			Ingresos	
Estudios	Antigüedad	Permiso	Bajos	Altos
Inferiores	1960-70	No	10.2	6.8
Inferiores	1960-70	Sí	1.7	3.3
Inferiores	1971-77	No	37.5	10.5
Inferiores	1971-77	Sí	15.7	13.3
Inferiores	1978-80	No	75.2	9.8
Inferiores	1978-80	Sí	18.7	7.3
Superiores	1960-70	No	6.0	7.0
Superiores	1960-70	Sí	3.1	10.9
Superiores	1971-77	No	28.1	13.9
Superiores	1971-77	Sí	17.7	26.3
Superiores	1978-80	No	39.9	9.1
Superiores	1978-80	Sí	14.2	9.8

A partir de los datos de la tabla 14 vemos el amplio recorrido de las razones. La mayor corresponde a las condiciones bajo nivel de educación, recién llegados (1978-80) y sin permiso para trabajar: la razón de ingresos bajos a altos para un emigrante en esas circunstancias es de 7.67 a 1.00 (75.2/9.8). En el lado opuesto, la razón de ingresos bajos a altos para emigrantes con estudios superiores, llegados en los 60 y con permiso de trabajo tiene un valor de .284 a 1.000. Ahora bien, ¿qué explicación hay para que se den estas grandes diferencias?, ¿qué influencia tiene cada variable independiente? Podemos encontrar una respuesta cuando estudiamos los parámetros de la ecuación logit. Para el caso general esta ecuación incluye los siguientes parámetros:

$$\Phi_{ijkl} = \beta^{I} + \beta_{ij}^{IP} + \beta_{ik}^{IA} + \beta_{il}^{IE}$$

Y para la primera de las razones que acabamos de poner como ejemplo, mirando los valores correspondientes en la tabla 15 tenemos:

$$\Phi_{i131} = .4074 + .554 + .798 + .282 = 2.0414$$

El exponencial de 2.0414 es igual a 7.70 —es decir, excepto errores de redondeo, la razón esperada de la tabla 13.

TABLA 15. Parámetros estimados según el modelo IP, IA, IE, PAE.

Efecto	λ	β	γ	λ (estandarizada)
I	.2037	.4074	1.503	3.032
IP	.277	.554	1.74	4.645
IA_1	−.415	−.830	.436	−3.759
IA_2	.015	.030	1.030	.182
IA_3	.399	.798	2.221	4.626
IE	.141	.282	1.325	2.397

El mismo valor de 7.67 es el que se obtiene cuando se omiten los logaritmos en toda la ecuación. La razón esperada de Ingresos se calcula así:

$$\Omega_{i131} = \gamma^{I}\gamma_{1}^{IP}\gamma_{3}^{IA}\gamma_{1}^{IE} \quad \text{donde} \quad \Omega_{i131} = \frac{F_{1131}}{F_{2131}} \quad \text{y} \quad \gamma = \exp\beta$$

$$\Omega_{i131} = (1.503)(1.74)(2.222)(1.325) = 7.69$$

Vamos a utilizar esta forma multiplicativa del modelo *logit* para interpretar los resultados obtenidos:

1. La razón de ingresos bajos a altos para un emigrante elegido al azar es de 1.503 a 1.000.

2. Controlando por Educación y Antigüedad, la razón de Ingresos bajos a altos para un individuo sin permiso de trabajo es de 1.74 a 1.00. Suponiendo que todos los emigrantes tuviesen iguales estudios y antigüedad en el país, por cada uno sin permiso que tenga ingresos altos nos vamos a encontrar 1.74 con ingresos bajos.

3. Controlando por Permiso y Educación, la razón de ingresos bajos a altos para un emigrante llegado entre 1978-80 es de 2.22 a 1.00. En el caso de un emigrante llegado entre 1960 y 1970 la razón es de .436 a 1.000 —o, lo que es lo mismo, 1.00 a 2.29. La relación entre Ingresos y la categoría 1971-77 de Antigüedad no es significativa —en la tabla 15 vemos que su valor estandarizado es .182.

4. Finalmente, controlando por Permiso y Antigüedad la razón de ingresos bajos a altos para un emigrante con nivel de estudios inferiores es de 1.325 a 1.000.

Es evidente a partir de los resultados comentados que la influencia más importante sobre Ingresos es aquella debida a la Antigüedad de los emigrantes en el país. Si expresamos esta influencia en porcentajes se puede decir que el 39.1% del logaritmo de la razón condicional (.798/2.0414) está explicado por la relación que hay entre Ingresos y Antigüedad. Le sigue en importancia la influencia de Permiso (27.1%), y finalmente Educación (13.8%). Puesto que no hay interacción significativa, esta pauta de influencias se mantiene en todas las circunstancias. Parece que ser paciente y esperar a llevar muchos años en España es lo más importante para tener ingresos altos. Por otro lado, no tiene gran importancia el nivel de estudios conseguido.

11.7. Modelos Causales

Todo lo que hemos hecho en el punto anterior es ajustar un modelo de la forma «Ingresos está directamente influenciado por Permiso, Antigüedad y Educación». Tal afirmación se puede simbolizar en un grafo como el de la figura 2.

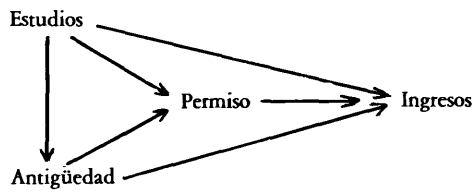
GRÁFICO 2. Figura ilustrativa, sólo con efectos directos.

Pero podemos estar también interesados en las relaciones entre las variables independientes. Si tenemos una teoría que nos dice cuáles son las relaciones, es posible evaluar esta teoría utilizando un modelo causal. Por ejemplo, podemos asumir la pauta de relaciones del gráfico 3, intentando calcular el tamaño de cada asociación, entre las variables.

Con objeto de calcular estos efectos, cuando se trata con variables cuantitativas se puede utilizar la técnica del *path analysis* (Duncan, 1966, 1975; Sanders, 1980) o el sistema más general implementado en el programa LISREL (Joreskorg y Van Thillo,

1973; Saris, 1980). Para variables categóricas, Goodman (1973a, 1973b, 1979) ha desarrollado un método que trata de ser similar al *path analysis*. El paralelismo se rompe en varios aspectos. Primero, de acuerdo con Fienberg (1977: 105) no está clara la posibilidad de usar sistemas no recursivos con los modelos *logit*. Segundo, cuando se tienen politomías no hay un solo valor para el efecto de una variable sobre otra. Y tercero, no es posible descomponer el tamaño de los efectos de las variables, midiendo los efectos causales y los espúreos.

GRÁFICO 3. Figura ilustrativa con efectos indirectos.



Aparte de las diferencias mencionadas, a la hora de calcular los efectos de las variables antecedentes sobre las consecuentes se procede como en el *path analysis*, ajustando una serie de ecuaciones *logit* a las tablas plegadas que determinen la especificación de nuestro modelo. Tenemos tres variables consecuentes, luego se ajustan tres ecuaciones *logit* a las tablas plegadas —según palabras de Goodman (1979: 1084), el orden de prioridad de las variables determina qué modelos son relevantes y los modelos, a su vez, determinan qué tablas son las relevantes.

Primera variable consecuente es Antigüedad, que depende de la Educación. Por lo tanto, fijamos un modelo *logit* que explique la razón (o el logaritmo de la razón) de Antigüedad. Y este modelo se aplica a la tabla de dos dimensiones Educación-Antigüedad, plegando Permiso e Ingresos. La segunda variable consecuente es Permiso de Trabajo, con Educación y Antigüedad como precedentes. Ahora se ajusta el modelo relevante, aplicado a la tabla de tres dimensiones Educación-Antigüedad-Permiso —plegando Ingresos—, que explica la razón de Permiso. Finalmente tenemos Ingresos, con Permiso, Educación y Antigüedad como variables antecedentes. Ajustamos un modelo que explique la razón (o el logaritmo de la razón) de Ingresos bajos a altos a partir de la tabla de 4 dimensiones.

Comenzando con el primer modelo, puesto que el modelo de independencia da una $\chi^2 = 7.4$, con 1 grado de libertad ($P = .0243$), concluimos que ambas variables están relacionadas. Ajustando el modelo saturado al cruce de Educación-Antigüedad obtenemos los parámetros estimados que aparecen en la tabla 16.

Debido a los valores estandarizados de la tabla 16 sólo vamos a considerar significativa la asociación entre Educación y Antigüedad (1978-80). El significado de esta asociación es como sigue: la gente que emigró en los 60 tenía más estudios que aquellos que lo hicieron a partir de 1978 —la razón de educación inferior a superior para un emigrante del 1978 es de 1.433 a 1.000.

A continuación analizamos la tabla de 3 dimensiones AEP, plegando I, y ajustan-

TABLA 16. Parámetros estimados según el modelo saturado.

Efecto	λ	β	γ
EA_1	−.12	−.24	.786
EA_2	−.07	−.14	.869
EA_3	.19	.36	1.433

do el mejor modelo. Después de ajustar diferentes modelos llegamos a AE, AP, EP. Este modelo tiene una $Y^2 = 1.14$, con 2 grados de libertad ($P = .5643$). Los parámetros estimados son los de la tabla 17.

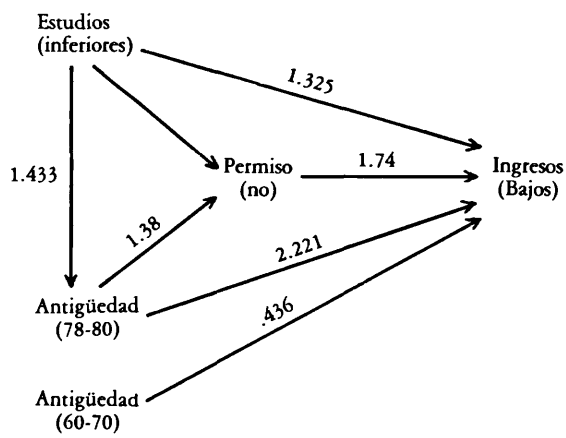
TABLA 17. Parámetros estimados según el modelo AP, EP, AE.

Efecto	λ	β	γ	λ (estandarizada)
A_1P	−.013	−.026	.974	− .122
A_2P	−.148	−.296	.743	−1.911
A_3P	.161	.322	1.380	2.061
EP	.172	.344	1.410	3.218

Nuestro tercer modelo incluye las 4 variables. El modelo que mejor ajusta los datos es el mismo que el obtenido cuando explicábamos el modelo logit IP, IA, IE, PAE, cuyos parámetros estimados están en la tabla 15.

Cuando acumulamos los resultados de los tres modelos precedentes obtenemos el modelo recursivo del gráfico 4. Este modelo ajusta las tablas marginales EA_3, EP, A_3P, EI, PI, A_1I, A_3I y tiene una $Y^2 = (0 + 1.14 + 8.11) = 9.25$, con $(0 + 2 + 7) = 9$ grados de libertad ($P = .414$). De esta manera, este modelo recursivo proporciona un buen ajuste de nuestros datos.

GRÁFICO 4. Modelo lineal logarítmico para los datos de la Tabla 1.



Con respecto a la interpretación de estos resultados, podemos decir que el valor de cada flecha indica el efecto de la variable independiente sobre la razón de la dependiente, controlando por el resto. Vemos así que la influencia más importante sobre Ingresos se debe a la Antigüedad: la razón de Ingresos bajos a altos para un emigrante recién llegado a España es de 2.22 a 1.00. En el caso del emigrante antiguo la razón es negativa. Para este grupo hay más individuos con ingresos altos: por cada persona con ingresos bajos hay 2.3 (1/.436) que son «ricos». Además de esta influencia directa hay otra indirecta a través de Permiso. No podemos calcular su tamaño pero podemos decir que opera en un sentido positivo: los recién llegados no tienen permiso de trabajo y los que no tienen permiso tienen ingresos bajos.

Permiso tiene una influencia positiva sobre Ingresos. La razón de Ingresos para una persona sin permiso, controlando por Educación y Antigüedad, es de 1.74 a 1.00. Esto significa que dos individuos con igual estudios y llegados al mismo tiempo, si uno no tiene permiso para trabajar su razón de ingresos bajos a altos será de 1.74 a 1.00, mientras que la misma razón para el emigrante con permiso sería de .574 a 1.000. Junto a este efecto directo de Permiso sobre los Ingresos debe de haber una influencia espúrea debido al hecho de que ambas variables están asociadas con Educación y Antigüedad.

Finalmente, Educación tiene también una asociación positiva sobre Ingresos (bajos); su valor es de 1.325. Además de este efecto directo hay otro indirecto a través de Permiso y de Antigüedad que aumenta la influencia total de los Estudios sobre los Ingresos —los signos de todos los caminos indirectos son positivos, así que, aun cuando no podamos calcular el efecto indirecto total sí podemos decir que es positivo.

Con esta interpretación de los resultados concluimos la explicación de los modelos lineales logarítmicos.

Programas de ordenador

ECTA (Everyman's Contingency Tables Analyser)

El programa funciona en batch[3]. Como algoritmo utiliza el «escalonamiento proporcional iterativo» (ver texto). Requiere el uso de un gran ordenador, pudiéndose obtener económicamente de

Leo A. Goodman
Department of Sociology
University of Chicago
1126 East 59th Street
Illinois 60637
U.S.A.

BMDP (Biomedical Computer Program)

Dentro de este paquete se incluye un programa (el 4F) para el análisis de modelos lineales logarítmicos. Al igual que ECTA utiliza el escalonamiento proporcional iterativo. Funciona en batch y el paquete se encuentra en la mayoría de los grandes Centros de Cálculo.

[3] Sistema por el cual los datos se le proporcionan al ordenador (normalmente en fichas) para que éste, después de varios minutos, horas o días, devuelva los resultados al usuario.

GLIM (Generalised Linear Modelling)

Se trata de un paquete muy potente, diseñado para todas aquellas técnicas derivadas del modelo lineal. Utiliza como algoritmo los «mínimos cuadrados iterativos» (ver texto). Funciona interactivamente[4] y requiere de un potente ordenador. Se puede obtener en

Numerical Algorithms Group Ltd.
13 Banbury Road
Oxford OX2 6NN
Inglaterra

LOGLIN

De acuerdo con su creador, el programa está pensado para pequeños ordenadores y funciona interactivamente. Utiliza el escalonamiento proporcional iterativo (ver texto). Se puede obtener en

G. N. Gilbert
Department of Sociology
University of Surrey
Guildford GU2 5XH
Inglaterra

[4] En este modo el usuario se relaciona directamente con el ordenador, vía un terminal, obteniendo los resultados inmediatamente.

12. Análisis de Tablas de Contingencia: Sistema de las Diferencias de Proporciones (Exégesis del trabajo de James A. Davis)

por *Juan Javier Sánchez Carrión*

12.1. Introducción

En una serie de artículos (1976, 1979, 1980, 1982), Davis explica la metodología desarrollada por él para analizar las tablas de contingencia. Entendemos que esta metodología resulta especialmente interesante para el tratamiento de variables nominales y ordinales. Todos los artículos son relativamente fáciles de comprender para el estudiante o el investigador con unos conocimientos básicos del análisis tabular. Por tanto, el objetivo de estas páginas es presentar parte de la obra del autor —dispersa en diferentes publicaciones y escrita en inglés— resumida en un solo artículo, facilitando así su divulgación.

En el capítulo de este número donde se explican lo modelos lineales logarítmicos ya se justifica la oportunidad de utilizar esa técnica, así como los sistemas de la *D*, a la hora de estudiar tablas multidimensionales. Los sistemas de la *D* es un método basado en tres ideas principales. En primer lugar utiliza las diferencias de porcentajes (o proporciones) como medida base de asociación. En segundo lugar, esta técnica parte del «paradigma de la elaboración», desarrolado por Lazarsfeld y Rosenberg (1955) y Rosenberg (1968), y va más allá asignando valores a las relaciones conceptuales definidas por esos autores a la vez que diseña un sistema total. Por último, sigue los principios *flow graph,* con el fin de representar los sistemas lineales en forma de grafo (Davis, 1979). En pocas palabras, tal como indica el profesor Davis, el sistema de la *D* es un intento de seguir el trabajo de los económetras sin olvidar que la mayoría de las variables sociales son cualittivas, por lo que se necesita del análisis tabular.

Para explicar los sistemas de la *D* vamos a utilizar los datos procedentes de una investigación sobre emigrantes iberoamericanos en España. La investigación fue llevada a cabo en 1981 por Gloria Lutz y Miguel Roiz, quienes amablemente me permitieron hacer uso de los resultados de su encuesta. En la medida en que no estamos interesados en conclusiones de tipo sustantivo, sino en la exposición de una técnica de análisis, vamos a asumir que la muestra era aleatoria simple, al tiempo que recodificamos las variables objeto de estudio.

Tenemos datos de 4 variables. Estudios, recodificada en nivel inferior (menos de título de grado medio) y nivel superior (título de grado medio o más). Antigüedad en el país, recodificada en aquellos que llegaron a España de 1960 a 1970, del 71 al 77 y de 1978 a 1980. Permiso de Trabajo, con las categorías no y sí. Por último, Ingresos,

recodificada en bajos (la mediana de la distribución o menos) y altos (más de la mediana). Los datos correspondientes aparecen en la tabla 1.

TABLA 1. Datos para el Análisis.

Estudios	Antigüedad	Permiso	Ingresos	
			Bajos	Altos
Inferiores	1960-70	No	10	7
Inferiores	1960-70	Sí	2	3
Inferiores	1971-77	No	39	9
Inferiores	1971-77	Sí	15	14
Inferiores	1978-80	No	75	10
Inferiores	1978-80	Sí	18	8
Superiores	1960-70	No	7	6
Superiores	1960-70	Sí	2	12
Superiores	1971-77	No	23	19
Superiores	1971-77	Sí	22	22
Superiores	1978-80	No	43	6
Superiores	1978-80	Sí	12	12
			268	128

12.2. Diferencia de Proporciones

Tomando como referencia los datos de la tabla 2, supongamos que se quiera estudiar la relación entre las dos variables. A tal fin, la simple constatación de los datos que aparecen en la tabla resulta poco informativa.

TABLA 2. Cruce de Permiso de Trabajo con Ingresos.

Ingresos (X_2)	Permiso (X_1) No	Sí	Total
Bajos	197	71	268
Altos	57	71	128
Total	254	142	396

Si queremos estudiar la asociación entre Ingresos y Permiso de Trabajo es conveniente comparar las frecuencias de las casillas con alguna medida o norma. Dos posibilidades son: comparar las frecuencias de las casillas entre sí o compararlas con los

marginales. En el primero de los casos lo que hacemos es calcular razones *(odds)* (véase el capítulo 11 de este mismo libro «Análisis de Tablas de Contingencia: Modelos lineales logarítmicos»). En el segundo estudiamos los porcentajes o proporciones. Supongamos que se quiera estudiar la posible influencia del Permiso en el hecho de tener ingresos bajos. Entre los individuos con Permiso de trabajo, 71/142 = .5 tienen ingresos bajos; entre los individuos sin Permiso, la proporción es 197/254 = .776. La diferencia entre ambas cantidades .776 − .500 = .276 puede considerarse como una medida de asociación. El cuadro 3 muestra la forma de presentar estos cálculos.

TABLA 3. Proporción «ingresos bajos», según Permiso de Trabajo.

Permiso	Proporción «ingresos bajos»
No	.776 (254)
Sí	.500 (142)
d = (.776 − .500) = .276	

De momento, podemos interpretar estos resultados en términos de colectivos o en términos individuales. Podemos decir que en el colectivo de los individuos sin Permiso de Trabajo, la proporción que tiene ingresos bajos es superior a la que existe en el grupo de los que sí tienen Permiso (.276 superior). Alternativamente se puede hablar de que un individuo sin Permiso es mucho más probable que tenga ingresos bajos que otro con Permiso de Trabajo (.276 más probable).

Entre las propiedades de la diferencia de proporciones (a partir de ahora d) podemos decir que su valor es cero cuando las dos variables son independientes, teniendo un máximo de +1.000 y −1.000, según que la asociación sea positiva o negativa, respectivamente. Otra propiedad es que la d no se ve afectada en su valor absoluto por el cambio de orden de las categorías de la variable independiente, aun cuando sí cambia su signo. Si vemos la d de «con Permiso» menos «sin Permiso», su valor será igual a −.276.

Una desventaja de la d, con relación a las razones, es su asimetría. La tabla 4 nos da los resultados si calculamos la d de la proporción de «sin Permiso».

TABLA 4. Proporción «no Permiso», según Ingresos.

Ingresos	Proporción «no Permiso»
Bajos	.735 (268)
Altos	.445 (128)
d = (.735 − .445) = .290	

La *d* de Permiso es mayor que la *d* de Ingresos. Este hecho es importante y obligará a que en la aplicación que se haga de esta medida a la construcción de modelos causales sea necesario especificar correctamente el orden de las variables. Como regla digamos que la variable independiente debe de dar las categorías, mientras que se calculan las proporciones de estas categorías para una categoría de la variable dependiente. Es decir, entre los cuadros 3 y 4, el primero será el correcto.

12.3. Inferencia estadística utilizando proporciones y diferencias de proporciones

A partir de los conocimientos básicos de estadística sabemos que una proporción *(P)* no es más que la media para una variable codificada 0 y 1; y que la varianza de una proporción *(P)* es igual a $P(1 - P)$ (Wonnacott y Wonnacott, 1977: 167). Cuando queremos estimar la proporción poblacional (π), a partir de la proporción muestral *(P)*, podemos construir un intervalo de confianza utilizando la desviación típica muestral de la proporción: $\sigma_P = \sqrt{\frac{P(1 - P)}{n}}$. Así, $\pi = P \pm Z\sigma_p$.

Con relación a la *d*, ésta no es sino una combinación lineal de proporciones. Así, la desviación típica muestral de la $d(\sigma_d)$ será igual a:

$$\sqrt{\frac{P_i(1 - P_i)}{n_i} + \frac{P_j(1 - P_j)}{n_j}}$$

y la *d* poblacional (D),

$$D = d \pm Z\sigma_d$$

Tomando como ejemplo la *d* de Ingresos, veamos en el siguiente cuadro (tabla 5) el cálculo de la desviación típica muestral.

TABLA 5. Varianza muestral de la *d* de Ingresos y Permiso de Trabajo.

Permiso	Proporción «ingresos bajos»	$(1 - P)$	$P(1 - P)$	n	$P(1 - P)/n = \sigma_d$
No	.776	.224	.174	254	.000684
Sí	.500	.500	.250	142	.001760
				396	.002444

La desviación típica muestral será $\sqrt{.00244}$ y la D poblacional estará comprendida en el intervalo (con un nivel de confianza del 95%) $.276 \pm (1.96\sqrt{.00244})$, que va de

.3728 a .1792[1]. Utilizando los intervalos de confianza para ver la significatividad de la *d* (Wonnacott y Wonnacott, 1977: 241) comprobamos que su valor es distinto de cero, al nivel de significación del .05.

12.4. Ecuaciones lineales y su representación en grafos

Las ecuaciones lineales se componen de variables, constantes y coeficientes. Así, la relación lineal entre dos variables se representa por la ecuación $X_2 = K + aX_1$, donde X_1 y X_2 son las variables; K es la constante, igual al valor de X_2 cuando $X_1 = 0$; y a es un coeficiente que expresa el cambio de valor de X_2 cuando X_1 varía en una unidad.

Cuando se pasa de 2 a más variables, sus relaciones se pueden representar mediante una serie de ecuaciones. Un ejemplo que incluyera las relaciones entre 4 variables podría resumirse con el siguiente conjunto de ecuaciones:

$$X_2 = aX_1 + K_2 \qquad [1]$$

$$X_3 = bX_1 + K_3 \qquad [2]$$

$$X_4 = -cX_1 + dX_2 - eX_3 + K_4 \qquad [3]$$

El sistema tiene 4 variables, 3 constantes y 5 coeficientes. Mediante una serie de manipulaciones algebráicas podemos hallar las propiedades del sistema. Sin embargo resulta más sencillo describir esas propiedades si traducimos este conjunto de ecuaciones a un grafo —especialmente en este caso, cuando se superan las dos variables adentrándonos en sistemas más complejos. Todo lo que necesitamos para pasar de las ecuaciones a los grafos, y viceversa, es una serie de reglas.

Regla 1. Cada coeficiente se asocia con una flecha que va de la variable independiente a la dependiente. Adoptamos la convención de que la línea sea continua cuando el coeficiente sea positivo, discontinua si es negativo y se suprime la línea cuando el coeficiente es cero.

Regla 2. Las constantes se asocian con puntos que tienen flechas sin coeficientes. También aquí acordamos que estas flechas tengan el valor 1.000.

De acuerdo con estas reglas, éste sería el grafo de la ecuación primera, sin numerar (figura 1).

---

[1] Estos resultados son válidos en el supuesto de que la muestra sea aleatoria simple. De forma orientativa digamos que cuando la muestra sea multietápica, situación normal de las encuestas sociológicas, habrá que multiplicar la varianza muestral del supuesto de muestra aleatoria simple por dos (Davis, 1978; Moser y Kalton, 1977).

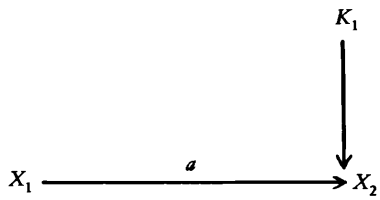


FIGURA 1. Grafo de la relación entre dos variables.

El sistema que incluyera las ecuaciones [1], [2], [3] se representaría en el siguiente grafo (figura 2).

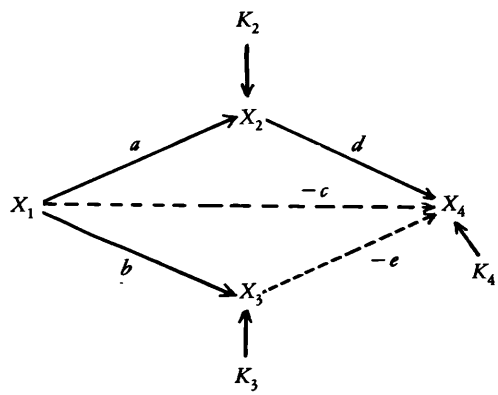
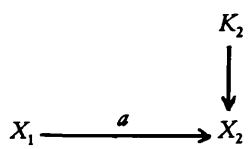


FIGURA 2. Grafo del sistema de ecuaciones 1, 2 y 3.

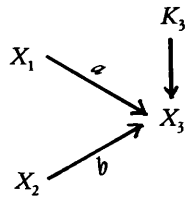
Hasta ahora hemos visto la forma de representar gráficamente un sistema de ecuaciones, pero si queremos obtener de un grafo la misma información contenida en las ecuaciones hemos de ampliar los conceptos ya vistos, explicando cómo se calcula el valor de las variables. Para ello seguiremos otra serie de reglas.

Regla 3. El valor de una variable determinada por una sola fuente es igual al producto del valor de la fuente por el coeficiente, más la constante.



implica que $X_2 = aX_1 + K_2$

Regla 4. El valor de una variable determinada por dos o más fuentes es igual a la suma de los valores de cada fuente multiplicados por sus coeficientes respectivos, más la constante.



implica que $X_3 = aX_1 + bX_2 + K_3$

Junto a las reglas a seguir para la representación gráfica y el cálculo del valor de las variables de un grafo, vamos a añadir una nueva para calcular el efecto que tiene una variable sobre otra. En el supuesto bivariable el efecto es igual a la diferencia de proporciones (d) entre las dos variables X_1 y X_2. Supongamos que el valor de X_1 aumenta en Δ_1 («delta», es la letra griega utilizada para designar la cantidad de cambio en las propiedades de un sistema), entonces X_2 cambiará en $\Delta_1 * d$ unidades.

Cuando se tiene un sistema con tres o más variables, como el de la figura 3, a la hora de calcular el efecto de X_1 sobre X_4 hay que introducir el concepto de «camino» (*path*). Un «camino» es la ruta que va de una variable antecedente a otra consecuente siguiendo las flechas intermedias. En todo camino hay que calcular su signo y su valor.

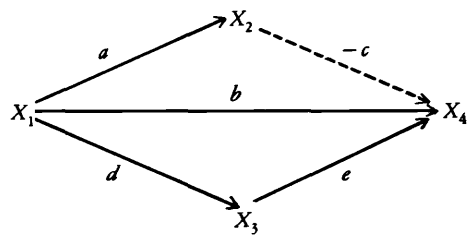


FIGURA 3. Grafo de la relación entre 4 variables.

Regla 5. El valor de un camino es igual al producto de los valores de las flechas que les unen —y el signo es igual al producto de los signos.

Entre X_1 y X_4 se establecen tres caminos: dos indirectos y uno directo. Vía X_2 hay un camino indirecto igual a $(a * -c)$; vía X_3 hay otro camino indirecto $(d * e)$; y luego existe el camino directo (b). El efecto total causal de X_1 sobre X_4 es igual a $(a * -c) + (d * e) + b$. Por lo tanto, un cambio de Δ_2 en X_1 supone un cambio de (Δ_2 * valor de los caminos) en X_4.

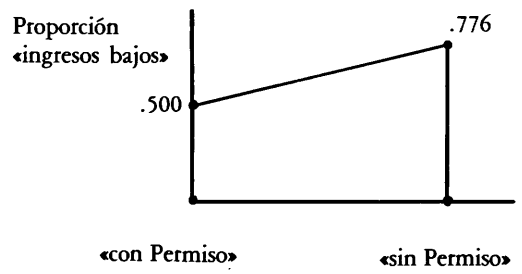
Si analizamos el efecto de X_2 sobre X_4 vemos que sólo existe un camino directo $(-c)$, igual al efecto causal de la primera variable sobre la segunda. Sin embargo, junto al efecto causal hay otro efecto espúreo, que es debido a la relación de ambas variables con aquellas que les preceden (X_1 y X_3). Este efecto espúreo será igual a la

diferencia entre la asociación bruta *(zero order)* de ambas variables menos el efecto causal, también entre las dos[2].

Con estas reglas podemos pasar del sistema de ecuaciones a su representación a un grafo y viceversa. Entendemos que el grafo tiene la ventaja de ser a simple vista más informativo que el sistema de ecuaciones.

Todo lo que necesitamos ahora es mostrar cómo se puede pasar de las tablas a un sistema de ecuaciones, susceptible de ser traducido a un grafo. Comencemos con el caso más simple, cuando tenemos una tabla de dos variables dicotómicas.

De acuerdo con los requisitos de una ecuación, a partir de la tabla 2 (tomemos estos datos como ejemplo) hemos de obtener un coeficiente, una constante y unas variables. Con objeto de calcular el coeficiente y la constante vamos a transformar nuestras variables Ingresos y Permiso, X_2 y X_1 respectivamente, en variables 0-1. En este caso, la d es el coeficiente —la pendiente, en términos de la regresión. Su significado es que cuando el Permiso cambia en una unidad, es decir pasa de «con Permiso» (0) a «sin Permiso» (1), los Ingresos cambian en d unidades (.776 − .500). Veamos gráficamente:



$$\text{Pendiente} = \frac{\text{cambio Ingresos}}{\text{cambio Permiso}} = \frac{d}{1} = d$$

Por lo tanto, la d es la pendiente en un sistema de variables 0-1. Veamos ahora la constante. Según la ecuación $\bar{X}_2 = K_2 + d\bar{X}_1$, K_2 es la constante o intercepción. Cuando $X_1 = 0$, tenemos que $X_2 = K_2$. Según nuestros datos, después de calcular el valor de d, tenemos que $X_2 = K_2 + .276X_1$.

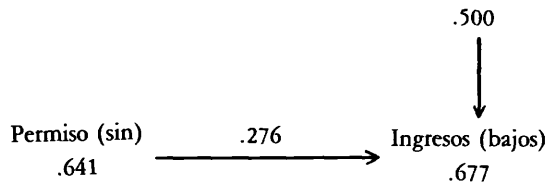
X_1 es igual a cero cuando la categoría de Permiso de Trabajo es «con Permiso», y en este caso Ingresos (X_2) es igual a .500. Por lo tanto, la constante es igual a la proporción de casos en la variable dependiente cuando la independiente es igual a cero.

Por último, ya hemos dicho que cuando se trata de variables 0-1 una proporción es igual a la media. En el caso que estamos estudiando, la media de Ingresos es igual a la proporción del marginal de la categoría 1 de esta variable (268/396 = .677); mientras que la media de Permiso será igual a la proporción del marginal de su categoría 1, (254/396 = .641). Poniendo juntas todas las ideas construimos la siguiente ecuación:

$$\text{Ingresos (bajos)} = .500 + (.276 * .641) = .677$$

[2] Sobre la definición de los efectos en el análisis causal véase la primera parte del artículo de Alwin y Hauser (1975). En este apartado tan sólo se ofrece una breve introducción al tema de las ecuaciones lineales y su representación en grafos, imprescindible para seguir las explicaciones posteriores. Para una mayor información sobre el tema se pueden ver Heise (1975), Stinchcombe (1968) y Davis (1979).

Y siguiendo las reglas de construcción de grafos podemos representar esta ecuación en la forma siguiente:



De igual manera se puede proceder cuando tengamos más de 2 variables; pero antes de pasar a este caso vamos a introducir algunos conceptos nuevos, relacionados con las tablas multidimensionales.

12.5. Tablas Multidimensionales

Al hablar del «proceso de elaboración», Rosenberg (1968) ilustra las ventajas de añadir nuevas variables cuando previamente se ha visto la relación entre otras dos. Básicamente la ventaja radica en poder atribuir con precisión los efectos correspondientes a cada una de las variables que intervienen en el modelo. Si estudiamos la tasa de mortalidad de distintos municipios españoles, agrupándolos en rurales y urbanos, podemos observar que su valor es mayor en el campo. Sin embargo, de esta constatación no se puede deducir que el hecho rural sea un factor que acelere la mortalidad. Cualquier estudiante de demografía sabe que una característica determinante del campo español es el envejecimiento de su población. Y este hecho sí que está asociado con una mayor mortalidad; siendo así que se puede atribuir erróneamente la mortalidad al fenómeno rural cuando es la edad de las personas que viven en ese medio lo que realmente da la explicación de esa mayor mortalidad: campo implica más viejos, y más viejos lleva a más mortalidad. Con objeto de salir de la duda y poder atribuir a cada variable lo que le pertenece se hace necesario controlar por la edad cara a sacar conclusiones sobre la relación entre hábitat y tasa de mortalidad. El mismo estudiante sabe también que la forma de controlar la edad es hacer *como* si las poblaciones de los municipios rurales y de los urbanos tuvieran la misma pirámide de edad —es decir, estandarizar o normalizar las poblaciones. Ahora sí; si en poblaciones de igual edad los municipios rurales siguen teniendo mayor tasa de mortalidad, al menos no se puede encontrar la disculpa de la edad como factor que explique la mayor mortalidad, lo cual no es óbice para que se prueben otras hipótesis relacionadas con las condiciones sanitarias, la dieta, etc., etc., repitiendo la misma operación. Traduciendo lo dicho hasta ahora al lenguaje de las variables y de las asociaciones, todo lo que hemos hecho se puede explicar de dos formas diferentes. Se parte de la asociación entre las dos variables (hábitat y mortalidad) y se estudia su valor para cada nivel de edad —a esta técnica se le conoce con el nombre de *control por una tercera variable.* Alternativamente, se puede partir de la asociación original entre las dos variables calculando de nuevo la asociación en el supuesto de que no hubiera relación entre Edad y Hábitat

—esto es lo mismo que decir que la edad de los individuos que viven en los municipios rurales es la misma que la de los urbanos. La primera de las técnicas es la que se utiliza tradicionalmente cuando se estudian tablas multidimensionales. La segunda parte del principio de estandarización, aplicado al análisis tabular por Rosenberg (1962), y ha sido desarrollada por Davis (1982) para tablas de todo tipo de dimensiones, ampliando igualmente sus aplicaciones sociológicas al terreno de la simulación. Aun cuando el sistema de análisis que estamos explicando, basado en la diferencia de porcentajes o proporciones, permite el uso de las dos técnicas, para la estimación de los coeficientes vamos a seguir en el resto del artículo el principio de la estandarización, frente al cálculo de las d condicionales para cada categoría de la variable de control (sobre esta técnica véase Davis, 1976, 1980 y Sánchez Carrión, 1983).

Al comienzo del artículo veíamos el cálculo de la d cuando teníamos dos variables. Estudiaremos ahora la situación que se produce cuando tenemos tres variables y queremos calcular los coeficientes y las constantes correspondientes a sus interrelaciones. Para ello vamos a utilizar los datos de la tabla 6, donde aparece el cruce de Estudios con Antigüedad en el país y con Permiso de Trabajo.

TABLA 6. Cruce de Estudios con Antigüedad y con Permiso (proporciones).

Estudios	Antigüedad	Permiso		Total (n.os absolutos)
		No	Sí	
Inferiores	1960-70	.77	.23	22
Inferiores	1971-77	.62	.38	77
Inferiores	1978-80	.76	.23	111
Superiores	1960-70	.48	.52	27
Superiores	1971-77	.49	.51	86
Superiores	1978-80	.67	.33	73

Desde el momento en que tenemos más de dos variables lo primero que habrá que hacer es *especificar*, en función del conocimiento sustantivo que tengamos del problema, qué relaciones se establecen entre ellas. Vamos a asumir que los estudios están relacionados con la Antigüedad en España[3]. A su vez los Estudios también están relacionados con el hecho de tener o no Permiso de Trabajo, dependiendo esta última variable del nivel de Educación. Por último también Permiso de Trabajo depende de la Antigüedad que se tenga en el país. Esta especificación queda recogida en la figura 4.

De acuerdo con esta figura tenemos que calcular los coeficientes a, b y c y las constantes K_1 y K_2. Cada coeficiente se puede ver como un número que estima el cambio

[3] No está claro el sentido de la causalidad y se podría interpretar que ambas variables están relacionadas entre sí de forma recíproca. El sistema funcionaría igual en este supuesto. Por comodidad vamos a asumir la relación que aparece en la figura 4.

o la diferencia que se produce en la variable dependiente (cabeza de la flecha) al cambiar la variable independiente (cola de la flecha). Según nuestro gráfico, y siguiendo la lógica del análisis multivariado, tan sólo los Estudios influyen en la Antigüedad; por ello, a la hora de estimar esta influencia es suficiente con calcular la d bruta bivariada, sin necesidad de controlar por Permiso —sin tener en cuenta si los emigrantes tienen o no Permiso de Trabajo. Sin embargo, si queremos calcular el efecto de los Estudios sobre el Permiso, no es suficiente calcular la d bruta bivariada, puesto que los Estudios además de influir directamente sobre el Permiso de Trabajo lo hacen indirectamente a través de la Antigüedad. Y si procediéramos de tal manera no sabríamos qué parte de la influencia sobre Permiso se debe a los Estudios y qué parte a la Antigüedad, atribuyendo erróneamente toda la influencia a la primera variable. Lo mismo ocurre al estudiar la relación entre Antigüedad y Permiso: el cálculo del efecto bruto entre las dos variables oculta que ambas están influidas por los Estudios. La solución en ambos casos radica en descomponer el efecto entre variable dependiente e independiente en la parte que se debe a su relación directa (efecto neto) y aquella que hay que atribuir a la relación indirecta que se crea vía Antigüedad —primer caso— o a la relación de ambas variables con el Estudio —segundo caso.

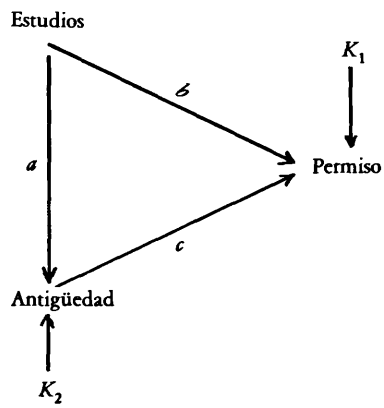


FIGURA 4. Grafo de la relación entre Estudios, Antigüedad y Permiso de Trabajo.

El problema planteado en el párrafo anterior es semejante al que indicábamos al estudiar la relación entre Hábitat y Mortalidad, y también la solución se encuentra en la estandarización. Si queremos estudiar la influencia directa de Estudios sobre Permiso vamos a hacer que Estudio y Antigüedad no estén relacionados —esto es lo mismo que hacer que tanto los emigrantes con estudios inferiores como aquellos con estudios superiores tengan la misma Antigüedad en España. Una vez que los emigrantes están estandarizados en cuanto a la Antigüedad podemos ver cuál es la relación —ahora neta— entre Estudios y Permiso, calculando la diferencia de proporciones y utilizando esta medida como coeficiente b. Veamos el procedimiento estadístico a seguir.

El cruce de las tres variables se puede presentar en la forma de la tabla 6 o también como un conjunto de proporciones consistente en 1) las proporciones marginales

para la variable Estudios; 2) las proporciones de Antigüedad para cada categoría de Estudios; y 3) las proporciones de Permiso para cada combinación de Estudios y Antigüedad. Lo que hacemos, por tanto, es colocar las variables en orden causal y construir una serie de tablas, añadiendo cada vez una nueva variable y calculando sus proporciones. En la tabla 7 se recoge el cálculo de las proporciones siguiendo el procedimiento mencionado.

TABLA 7. Proporciones recurrentes para el sistema de la figura 4.

	A = Estudios				
	Inferiores	Superiores			
1.	.530	.470	(396)		
	B = Antigüedad				
	Estudios	1960-70	1971-77	1978-80	
2.	Inferiores	.105	.367	.529	(210)
3.	Superiores	.145	.462	.392	(186)
		.124	.412	.464	(396)
	C = Permiso				
	Estudios	Antigüedad	No	Sí	
4.	Inferiores	1960-70	.770	.230	(22)
5.	Inferiores	1971-77	.620	.380	(77)
6.	Inferiores	1978-80	.765	.235	(111)
7.	Superiores	1960-70	.480	.520	(27)
8.	Superiores	1971-77	.490	.510	(86)
9.	Superiores	1978-80	.670	.330	(73)
			.640	.360	(396)

Nota: Entre paréntesis se incluyen los totales sobre los que se calculan las proporciones de cada línea.

Según esta tabla, la línea 1 indica que el 53% de los emigrantes tiene estudios inferiores, mientras que el 47% posee estudios superiores. La línea 2 dice que el 10.5% de los emigrantes con estudios inferiores llegaron a España entre 1960 y 1970; el 37% lo hizo del 71 al 77; y el 53% entre 1978 y 1980. Y, por último, la línea 9 muestra que el 67% de los emigrantes con estudios superiores y llegados a España en el período 1978-80 no tienen Permiso de Trabajo, en tanto que el 33% sí lo tienen.

A partir de este cuadro (tabla 7), si queremos calcular la influencia de Estudios sobre Antigüedad utilizaremos las líneas 2 y 3, procediendo tal como mostrábamos al

calcular el efecto de Permiso sobre Ingreso en el ejemplo bivariado. La única diferencia con aquel ejemplo se produce porque Antigüedad tiene tres categorías. Ahora hemos de tomar una categoría de esta variable como base, comparando con ella la variación en las otras dos cuando varía la variable independiente[4]. En este caso tendremos dos coeficientes para la relación entre Estudios y Antigüedad. Veamos su cálculo en la tabla 8.

TABLA 8. Proporción de emigrantes llegados en los períodos 60-70 y 78-80, según Estudios.

Estudios	Proporción 60-70 (*vs.* 71-77)
Inferiores	.105
Superiores	.145
	$d = -.040$

Estudios	Proporción 78-80 (*vs.* 71-77)
Inferiores	.529
Superiores	.392
	$d = .137$

Igualmente habrá dos constantes: .145 para Antigüedad (1960-70) y .392 para Antigüedad (1978-80). La media de Estudios es .530 y las medias para Antigüedad (1978-80) y Antigüedad (1960-70), .464 y .124, respectivamente. La tabla en cuestión se puede expresar en las dos ecuaciones siguientes:

$$\text{Antigüedad (60-70)} = .145 + (-.040 * .530) = .124 \quad [4]$$

$$\text{Antigüedad (78-80)} = .392 + (.137 * .530) = .464 \quad [5]$$

Los resultados de las dos ecuaciones coinciden con las medias de Antigüedad (60-70) y Antigüedad (78-80). Ambas ecuaciones se pueden traducir al grafo de la figura 5.

[4] Cuando las variables son dicotómicas, si cambiamos la base (utilizamos como base «con Permiso» en lugar de «sin Permiso») cambian el signo y la constante (−.276 y .776, *vs.* .276 y .500), lo cual no tiene una importancia sustantiva. Sin embargo, cuando se cambia la base en el caso de politomías, normalmente se modifican todos los coeficientes. Davis (1976: 125) señala que aunque la elección de la base sea arbitraria, es recomendable seguir tres reglas prácticas: 1) que las bases tengan una frecuencia de casos grande; 2) que sean distinguibles en términos sustantivos; y 3) es posible probar con varias bases y presentar los resultados con aquella que ofrezca mayor claridad.

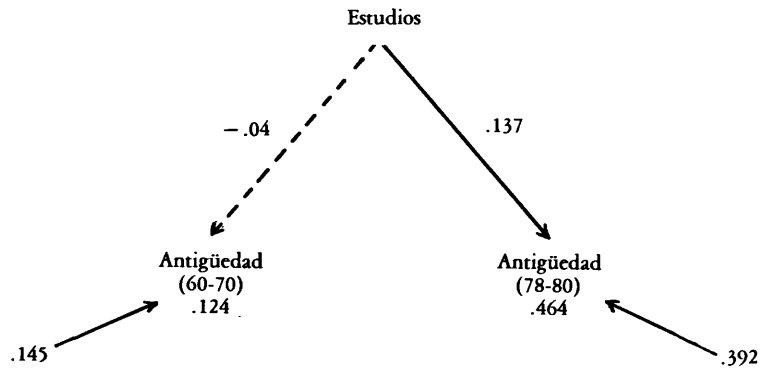


FIGURA 5. Grafo de la relación entre Estudios y Antigüedad.

El siguiente efecto a calcular sería el de Estudios sobre Permiso. Con objeto de calcular este coeficiente es necesario hacer que no haya asociación entre Estudios y Antigüedad. Tal cosa se producirá si hacemos que la proporción de emigrantes para cada categoría de Antigüedad sea la misma entre los amigrantes con estudios inferiores y con estudios superiores. Para ello serviría cualquier proporción, sin embargo, es común utilizar las proporciones de los marginales de una de las variables. La tabla 9 ofrece el supuesto de no asociación entre las variables, utilizando los marginales de Antigüedad.

TABLA 9. Cruce de Estudios y Antigüedad, en el supuesto de no asociación entre variables (proporciones y números absolutos)

	Estudios		
Antigüedad	Inferiores	Superiores	Total
1960-70	.124 (26)	.124 (23)	49
1971-77	.412 (77)	.412 (77)	164
1978-80	.464 (97)	.464 (86)	183
Total	1.000 (210)	1.000 (186)	396

Lo que hemos hecho ha sido dejar fijos los totales y calcular el número de individuos que habría en cada casilla, en el supuesto de que no hubiera asociación entre las variables[5]. Ahora, utilizando estos valores absolutos veamos cuál es la relación entre

[5] Obsérvese que no hay diferencia de la Antigüedad según los niveles de Estudio.

las tres variables. Para ello construimos el cruce de las tres variables que se ve en la tabla 10.

TABLA 10. Cruce de Estudios con Antigüedad y con Permiso, en el supuesto de no asociación entre las dos últimas variables (proporciones y números absolutos).

Estudios	Antigüedad	Permiso		Total
		No	Sí	
Inferiores	1960-70	.77 (20)	.23 (6)	26
Inferiores	1971-77	.62 (54)	.38 (33)	87
Inferiores	1978-80	.76 (74)	.23 (23)	97
Superiores	1960-70	.48 (11)	.52 (12)	23
Superiores	1971-77	.49 (38)	.51 (39)	77
Superiores	1978-80	.67 (58)	.33 (28)	86

Los totales corresponden al número de individuos que habría en cada combinación de las categorías de las variables Estudios y Antigüedad, en el supuesto de no asociación entre ambas variables. Las proporciones de las casillas son las originales (tabla 7) y las frecuencias absolutas indican el número de emigrantes que nos habríamos encontrado según los totales de no asociación y las proporciones originales. Si ahora queremos calcular el efecto neto (directo) de Estudios sobre Permiso no tenemos más que plegar *(collapse)* la variable Antigüedad y ver la diferencia de proporciones de Permiso para las categorías de Estudios. La tabla 11 nos ofrece los cálculos.

TABLA 11. Proporción de «no Permiso», según Estudios.

Estudios	Proporción «no Permiso»
Inferiores	(20 + 54 + 74)/210 = .705
Superiores	(11 + 38 + 58)/186 = .575
d = (.705 − .575) = .130	

Plegando la variable Estudios podemos ver la influencia de Antigüedad sobre Permiso (tabla 12).

Además de los coeficientes ya calculados, si deseamos construir el sistema de ecuaciones vemos que es necesario calcular la constante K_3 (véase figura 6), junto a la media de Permiso.

TABLA 12. Proporción de «no Permiso», según Antigüedad.

Antigüedad	Proporción «no Permiso»
1978-80	(74 + 58)/183 = .721
1971-77	(54 + 38)/163 = .564

d = (.721 − .564) = .157

Antigüedad	Proporción «no Permiso»
1960-70	(20 + 11)/49 = .633
1971-77	(54 + 38)/163 = .564

d = (.633 − .564) = .069

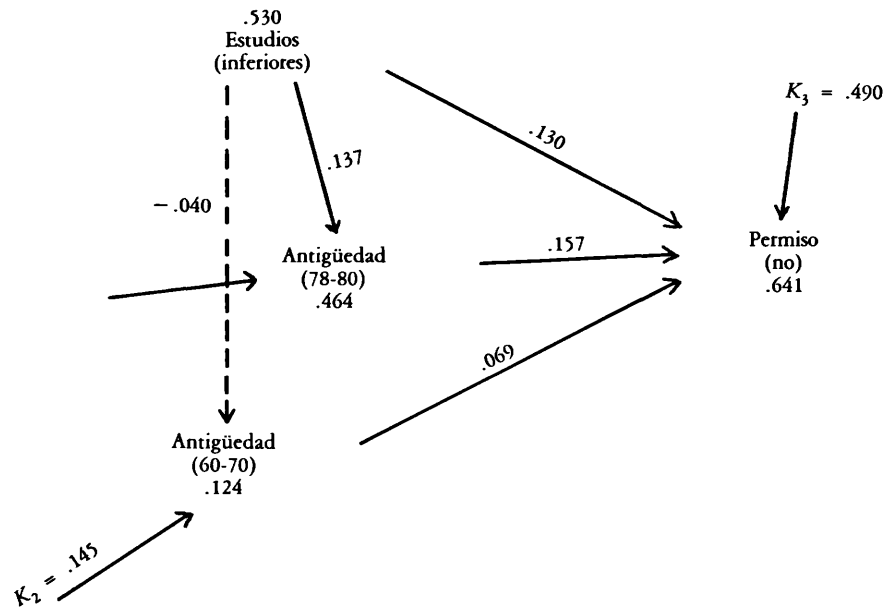


FIGURA 6. Grafo de la relación entre Estudios, Antigüedad y Permiso, con los coeficientes de las relaciones entre las 3 variables.

K_3 es igual a la proporción de emigrantes sin permiso cuando Antigüedad y Estudios son iguales a cero —es decir, la proporción de emigrantes sin permiso llegados entre 1971-77 y con estudios superiores. Según la tabla 7 esta proporción es igual a .490. La

media (proporción) de Permiso es 254/396 = .641. Con los datos que hemos calculado, el sistema de la figura 6 puede traducirse al siguiente sistema de ecuaciones:

$$\text{Permiso (no)} = .490 + (.13 * .53) + (.16 * .464) + (.07 * .124) = .641 \quad [6]$$

$$\text{Antigüedad (60-70)} = .145 + (-.04 * 530) = .129 \quad [7]$$

$$\text{Antigüedad (78-80)} = .392 + (.137 * .530) = .465 \quad [8]$$

Estos resultados pueden interpretarse de la siguiente manera. Los Estudios están relacionados con la Antigüedad, especialmente con la categoría 1978-80: los individuos llegados en ese período es más probable que tuvieran estudios inferiores que aquellos que salieron de Iberoamérica de 1971-77 —esa mayor probabilidad es del .137. Por el contrario, los emigrantes de los 60 tenían una formación académica superior a los del período 71-77, aun cuando la diferencia sea mínima —más adelante veremos si este coeficiente de −.040 es significativo. Dicho de otra forma, los del 60 tenían una probabilidad de tener estudios inferiores menor (P = .040) que los llegados en la etapa 1971-77.

A su vez, los estudios inferiores están relacionados positivamente con el hecho de no tener Permiso de Trabajo: es .130 más probable que un individuo con estudios inferiores no tenga Permiso de Trabajo que otro con estudios superiores. Este sería el efecto directo de Estudios; indirectamente esta variable también influye vía Antigüedad: menos estudios se asocia con menor Antigüedad y menor Antigüedad implica menos posibilidades de conseguir Permiso de Trabajo. ¿Cuál es la importancia de este efecto? Según las reglas de representación de los grafos y cálculo de las transmisiones (véase Regla 5)[6], el efecto indirecto de Estudios sobre Permiso será igual al producto de los efectos de Estudios sobre Antigüedad y de Antigüedad sobre Permiso; en este caso, (.137 * .16) + (−.040 * .070) = .0191. Y el efecto total será igual al directo más el indirecto: .130 + .019 = .150. Este efecto ha de ser igual a la asociación bruta entre las dos variables. Veamos su tamaño en la tabla siguiente:

Estudios	Proporción «Permiso (no)»	
Inferiores	.714	(210)
Superiores	.560	(186)
	.154	

Vemos que hay una pequeña diferencia entre la suma de los efectos directo más indirecto y la asociación bruta bivariada: .150, en el primer caso, *vs.* .154, en el segundo. Ello es debido a la existencia de interacción entre las variables. Efectivamente, analizando la tabla 7 se ve que, por ejemplo, la asociación entre Estudios y Permiso no es la misma para todas las categorías de Antigüedad: para un individuo llegado en el perío-

[6] Una *transmisión* es la suma de los valores de los caminos que unen dos variables.

do 1960-70, el hecho de aumentar sus Estudios (pasar de estudios inferiores a estudios superiores) implica que se dupliquen sus posibilidades de conseguir Permiso de Trabajo (.230 a .520); sin embargo, si ese mismo individuo hubiera llegado entre 1978 y 1980, el hecho de mejorar su educación sólo le supondría un aumento pequeño en sus posibilidades de conseguir ese Permiso de Trabajo (.235 a .330).

Excepto en el caso de que no haya interacción, las *d* y las transmisiones *(transmittances)* no coinciden con los datos brutos. Tal como señala Davis (1967: 130), el hecho de que la discrepancia se pueda considerar tolerable es un problema de opinión del investigador. Cuando las muestras son pequeñas, aun con interacciones no significativas las discrepancias pueden ser grandes. Por el contrario, si las muestras son grandes puede que interacciones significativas no creen problemas.

Continuando con la interpretación de los resultados obtenidos, vemos que la Antigüedad influye positivamente en el Permiso. Tanto los individuos llegados de 1960 a 1970 como los que lo hicieron entre 1978 y 1980 tienen más posibilidades de no tener Permiso de Trabajo que los del período 1971-77 —especialmente los recién llegados (P = 157). Por ejemplo, si comparamos el efecto directo de Antigüedad (1978-80), .157, con el bruto, .176, vemos que .019 del total es espúreo, debido a la influencia de Estudios.

Un próximo paso en el análisis de los datos consiste en el estudio de las nuevas relaciones que se establecen cuando se añade la cuarta variable, los Ingresos. En función de la especificación que hacemos en el gráfico 7 es necesario calcular los coeficientes *f*, *g*, *h* e *i*, además de la constante K_4 y la media de Ingresos.

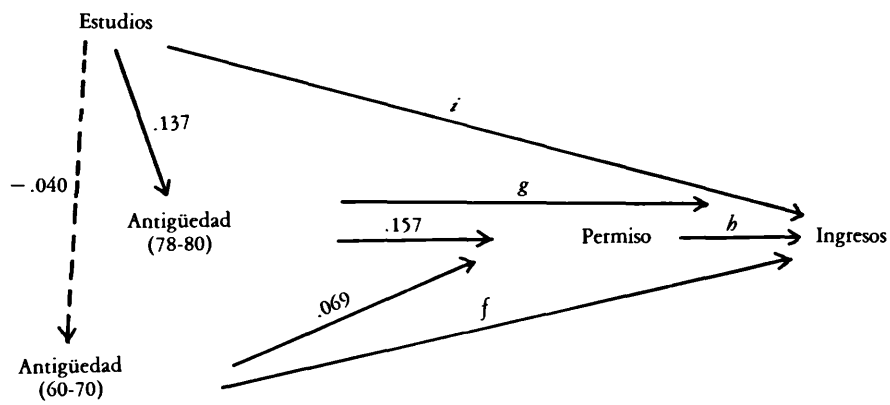


FIGURA 7. Grafo de las relaciones entre Estudios, Antigüedad, Permiso e Ingresos.

Con objeto de calcular estos parámetros no hacemos sino extender la explicación que ofrecíamos al tratar del caso de 3 variables. En primer lugar es necesario calcular el cruce de las 4 variables (en proporciones) (tabla 14).

Si ahora queremos ver la influencia de Estudios sobre Ingresos, hay que distinguir la parte directa de la indirecta, a través de Permiso y Antigüedad. Para no confundir

las influencias ya hemos visto que se hace necesario estandarizar los individuos en aquellas variables que puedan crear confusión. En este caso es necesario hacer que tanto los emigrantes con estudios inferiores como aquellos con estudios superiores tengan la misma Antigüedad en el país y las mismas condiciones laborales (Permiso de Trabajo). En estas circunstancias, la influencia que tengan los Estudios sobre los Ingresos no cabe duda de que se debe sólo y exclusivamente a esta variable. Esta anulación de efectos se consigue haciendo que los Estudios no aparezcan asociados con Antigüedad y con Permiso. Para ello basta con sustituir en la tabla 6 las proporciones originales por las marginales de Permiso, por ejemplo[7]. En la tabla 13 damos estos resultados, incluyendo entre paréntesis el número absoluto de emigrantes que habría en cada combinación de categorías, en el supuesto de que no hubiera asociación.

TABLA 13. Cruce de Estudios con Antigüedad y con Permiso, en el supuesto de independencia múltiple (proporciones y números absolutos).

Estudios	Antigüedad	Permiso		Total
		No	Sí	
Inferiores	1960-70	.64 (17)	.36 (9)	26
Inferiores	1971-77	.64 (56)	.36 (31)	87
Inferiores	1978-80	.64 (62)	.36 (35)	97
Superiores	1960-70	.64 (15)	.36 (8)	23
Superiores	1971-77	.64 (49)	.36 (28)	77
Superiores	1978-80	.64 (55)	.36 (31)	86

Ahora lo que hacemos es utilizar las frecuencias esperadas en el supuesto de no asociación entre las tres variables para ver cuál es la relación entre las 4 variables del modelo. A tal fin construimos la tabla de 4 dimensiones, donde los marginales se corresponden con las frecuencias de las casillas de la tabla 13, siendo las proporciones de las casillas las originales. La tabla 14 da los resultados.

Si ahora queremos ver el efecto de Estudios sobre Ingresos (en la figura 7, i), tan sólo tenemos que plegar Antigüedad y Permiso, cruzando aquellas variables y procediendo como en las tablas 11 y 12. Así $i = .120$. De igual manera se calculan los efectos de Antigüedad y Permiso sobre Ingresos: $f = -.157$; $g = .151$; $h = .232$. Si deseamos construir el sistema de ecuaciones, además de estos coeficientes tendremos que calcular K_4 y la media de Ingresos. La constante es igual a .500 y la media

[7] Ya hemos explicado que serviría cualquier marginal, puesto que tan sólo se trata de que las proporciones de la tabla indiquen que hay independencia múltiple entre las tres variables. (Sobre el concepto de independencia múltiple, así como de los otros tipos de relaciones que se establecen entre las variables en tablas multidimensionales véase el capítulo 11 de este mismo libro).

TABLA 14. Cruce de Estudios con Antigüedad, con Permiso y con Ingresos, en el supuesto de independencia múltiple entre las tres primeras variables (proporciones y números absolutos).

Estudios	Antigüedad	Permiso	Ingresos		Total
			Bajos	Altos	
Inferiores	1960-70	No	.588 (10)	.412 (7)	17
Inferiores	1960-70	Sí	.400 (4)	.600 (5)	9
Inferiores	1971-77	No	.812 (45)	.188 (11)	56
Inferiores	1971-77	Sí	.517 (16)	.483 (15)	31
Inferiores	1978-80	No	.882 (55)	.118 (7)	62
Inferiores	1978-80	Sí	.692 (24)	.308 (11)	35
Superiores	1960-70	No	.538 (8)	.462 (7)	15
Superiores	1960-70	Sí	.143 (1)	.857 (7)	8
Superiores	1971-77	No	.548 (27)	.452 (22)	49
Superiores	1971-77	Sí	.500 (14)	.500 (14)	28
Superiores	1978-80	No	.878 (48)	.122 (7)	55
Superiores	1978-80	Sí	.500 (16)	.500 (15)	31

(proporción) de Ingresos es de .676. El sistema de ecuaciones queda compuesto por las ecuaciones [6], [7] y [8] más la siguiente:

$$\text{Ingresos (bajos)} = .500 + (.530 * .120) + (.464 * .151) + (.124 * -.157) + (.641 * .232) = .763 \quad [9]$$

Comparando el resultado estimado por la ecuación anterior y el valor real de los Ingresos se observa una diferencia del 8.8%. Ello es debido, tal como ya tuvimos ocasión de señalar, al hecho de que, excepto cuando no existe interacción, siempre se observa una discrepancia entre los valores observados y los esperados según el modelo. Finalmente podemos ver los resultados completos del análisis en el gráfico 8.

Mirando los efectos directos de las variables independientes sobre los Ingresos vemos que, salvo Antigüedad (60-70), todos son positivos. Controlando por Antigüedad y Permiso, es decir, suponiendo igualdad de Permiso y Antigüedad para los emigrantes con estudios inferiores y superiores, el hecho de tener estudios inferiores influye positivamente de cara a tener ingresos bajos, aun cuando la relación no sea muy fuerte: .120. La influencia aumenta cuando consideramos el efecto indirecto de esta variable: tener estudios inferiores favorece el hecho de no tener permiso de trabajo, y esto influye para que se tengan ingresos bajos. Junto a esta influencia vía Permiso, la baja educación está asociada con el hecho de llegar tarde a España (1978-80) y esta circunstancia se relaciona positivamente con tener ingresos bajos. En conjunto este efecto indirecto es igual a (.130 * .232) + (.137 * .160 * .232) + (.137 * .151) =

= .056[8]. Es decir, de la asociación total entre Estudios e Ingresos (.171), parte se debe a la influencia directa de los estudios (.120) y parte a su influencia indirecta a través de Antigüedad y Permiso.

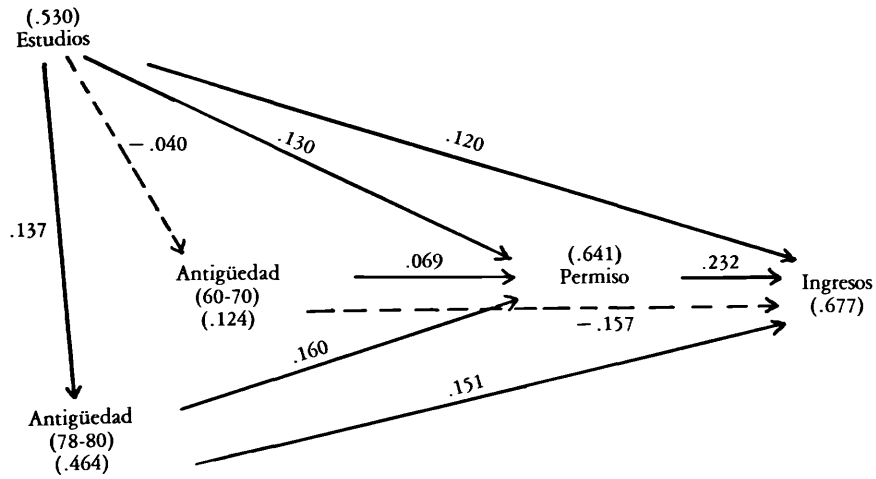


FIGURA 8. Grafo de las relaciones entre Estudios, Antigüedad, Permiso e Ingresos, con los coeficientes de las relaciones entre las 4 variables.

No tener permiso de trabajo también se asocia positivamente con el hecho de tener ingresos bajos. Para dos emigrantes que tengan iguales Estudios y Antigüedad, si uno no tiene Permiso de Trabajo su posibilidad de tener ingresos bajos es .232 superior a la del otro con Permiso de Trabajo. Si calculamos la asociación bruta entre Permiso e Ingresos vemos que su valor es .276; la diferencia entre este número y el efecto directo (.232) es la relación espúrea entre ambas variables, debida a su relación con Estudios y Antigüedad.

Con respecto a la influencia de Antigüedad, ésta es de doble sentido. Los que llegaron en los 60 tienen una mayor probabilidad de tener ingresos altos que aquellos que lo hicieron en el período 1971-77: la diferencia es de .157. En comparación también con estos últimos emigrantes y suponiendo que todos (los del período 78-80 y los del 71-77) tuvieran iguales Estudios y situación legal de trabajo, la probabilidad de tener ingresos bajos para aquellos que llegaron al final de los 70 es .151 superior. En este caso, junto al efecto directo hay otro indirecto vía Permiso que eleva la diferencia de la probabilidad de ser pobre hasta (.151 + (.160 * .232)) = .188. Debido a los Estudios hay una asociación espúrea entre Antigüedad (78-80) e Ingresos de .009 —la

[8] Al hablar de la estimación de los parámetros poblacionales y de la significatividad de los efectos ya veremos cómo la relación entre Estudios y Antigüedad (60-70) no es significativa. Es por ello que la excluimos del cálculo de los efectos indirectos.

diferencia entre la asociación total de Antigüedad (78-80) e Ingresos (.197) y la parte causal de la relación (.188).

A la vista de los resultados del modelo vemos que aquello que más influye de forma directa para tener ingresos altos es obtener Permiso de Trabajo, siguiendo en importancia el tiempo de residencia en España y, por último, los Estudios. Puesto que previamente vimos que, a su vez, conseguir Permiso de Trabajo depende en mayor medida de la Antigüedad que de los Estudios, parece que el problema económico de los emigrantes es un problema de paciencia, en el que poco influye su nivel de formación académica.

12.6. Inferencia Estadística en Tablas Multidimensionales

En la medida en que estamos tratando con datos muestrales se hace necesario proceder al estudio de la significatividad de los coeficientes que hemos estimado[9].

Siguiendo la explicación que hemos ofrecido al tratar de la situación bivariable veremos si cada una de las d obtenidas es estadísticamente significativa calculando su intervalo de confianza. Tomemos la d de la relación entre Estudios e Ingresos. En la tabla 15 se disponen los cálculos.

TABLA 15. Desviación típica muestral de la d de Ingresos y Estudio (datos estandarizados).

Estudios	Proporción «ingresos bajos»	$(1 - P)$	$P(1 - P)$	n	$P(1 - P)/n = \delta_d$
Inferiores	154/210 = .733	.267	.1957	210	.00093
Superiores	114/186 = .613	.387	.2372	186	.00127
				396	.00220

La desviación típica muestral será igual a $\sqrt{.0022}$ y la D poblacional estará comprendida en el intervalo (con un nivel de confianza del 95%) $.120 \pm (1.96\sqrt{.002})$; es decir, entre .212 y .028. Al no estar incluido el cero en el intervalo decimos que la d es estadísticamente significativa. De manera semejante calculamos los intervalos de confianza para el resto de las d. En la tabla 16 ofrecemos las d, junto con su varianza y el resultado de la prueba de significatividad utilizando intervalos de confianza.

De acuerdo con estos resultados el grafo definitivo sería el de la figura 9, siendo la interpretación igual a la ya expresada previamente.

[9] En Sánchez Carrión (1983) se explica la forma de contrastar la significatividad de las interacciones.

TABLA 16. Varianza y significatividad de las *d* estimadas.

Efectos	*d*	Varianza	¿Significativo?
Estudios-Antigüedad (60-70)	−.040	.00112	No
Estudios-Antigüedad (78-80)	.137	.00246	Sí
Estudios-Permiso	.130	.00230	Sí
Estudios-Ingresos	.120	.00220	Sí
Antigüedad (60-70)-Permiso	.069	.00626	No
Antigüedad (78-80)-Permiso	.157	.00260	Sí
Antigüedad (60-70)-Ingresos	−.157	.00650	Sí
Antigüedad (78-80)-Ingresos	.151	.00237	Sí
Permiso-Ingresos	.232	.00247	Sí

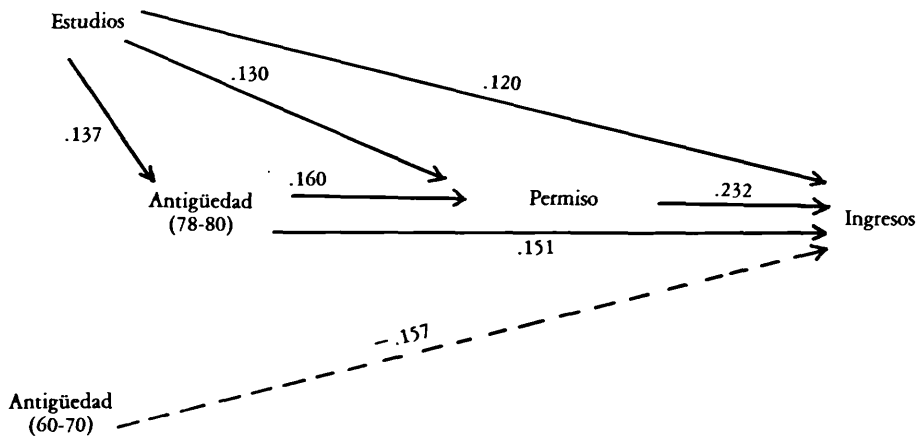


FIGURA 9. Grafo de la relación entre las 4 variables. Incluye solamente los coeficientes significativos.

12.7. La medida del impacto causal

La idea de causalidad se basa en el hecho de que si una variable es la causa de otra, el cambio en la primera produce un cambio en la segunda. Esta definición se puede operativizar mediante la tasa de cambio, que mide la cantidad de cambio que sufre la variable dependiente ante un cambio determinado de la variable independiente. En los sistemas de la *D*, la tasa de cambio viene dada por la *d*. Así, en el modelo que acabamos de estudiar, tomando como ejemplo la relación entre Permiso e Ingresos veíamos en el apartado 4 que la *d* bruta era igual a .276. Si suponemos una relación lineal entre ambas variables este coeficiente se puede interpretar diciendo que «un aumento de 10 puntos en la proporción de emigrantes sin permiso de trabajo

tendrá como resultado un aumento de 2.76 puntos en la proporción con ingresos bajos».

En el ejemplo anterior hemos visto que resulta sencillo de medir el impacto causal producido por un cambio en la variable independiente. Cuando las variables dependiente e independiente no están sólo relacionadas directamente —ejemplo de Estudios e ingresos— o cuando la variación se produce simultáneamente en más de una variable independiente que afecta a más de una variable dependiente —todo un sistema—, el cálculo de los efectos causales resulta algo más complejo. Utilizando los datos de una investigación realizada por García Ferrando (1982) sobre la práctica deportiva de los españoles vamos a mostrar cómo se resuelve el problema planteado, utilizando las proporciones recurrentes.

Los datos de partida, dispuestos en forma de proporciones recurrentes, están contenidos en la tabla 17. La tabla recoge el cruce de Edad con Estudios y con Práctica Deportiva; siendo las relaciones causales entre las variables las de la figura 10.

TABLA 17. Proporciones recurrentes para el sistema de la figura 10.

	A = Edad			
	15-25	26-40	40 +	
1	.237 (932)	.286 (1127)	.477 (1876)	(3935)

	B = Estudios				
	Edad	Primarios	Bachiller	Grado Medio	Universitarios	
2	15-25	.328 (306)	.544 (507)	.063 (59)	.064 (60)	(932)
3	26-40	.551 (621)	.267 (301)	.105 (118)	.077 (87)	(1127)
4	40 +	.558 (1532)	.113 (212)	.048 (90)	.022 (42)	(1876)
		.625	.259	.068	.048	(3935)

	C = Práctica Deportiva				
	Edad	Estudios	Sí	No	
5	15-25	Primarios	.408 (79)	.592 (115)	(194)
6	15-25	Bachiller	.671 (215)	.229 (106)	(321)
7	15-25	Grado Medio	.712 (26)	.288 (11)	(37)
8	15-25	Universitarios	.633 (24)	.367 (14)	(38)
9	26-40	Primarios	.200 (139)	.800 (555)	(694)
10	26-40	Bachiller	.419 (141)	.581 (195)	(336)
11	26-40	Grado Medio	.551 (73)	.449 (59)	(132)
12	26-40	Universitarios	.621 (60)	.379 (37)	(97)
13	40 +	Primarios	.060 (102)	.940 (1602)	(1704)
14	40 +	Bachiller	.240 (57)	.760 (179)	(236)
15	40 +	Grado Medio	.333 (33)	.667 (67)	(100)
16	40 +	Universitarios	.452 (21)	.548 (25)	(46)
			.281	.719	(3935)

Según los datos, en el momento de hacer la encuesta practicaban deporte una proporción del .281 de los entrevistados; y tenían estudios primarios, de bachiller, medios y universitarios el .625, .259, .068 y .048, respectivamente. Suponiendo que en la próxima generación de españoles la población se habrá envejecido y que el nivel de estudios habrá aumentado, se trata de ver la repercusión del envejecimiento sobre el nivel educacional de los españoles y de ambas variables sobre la práctica deportiva.

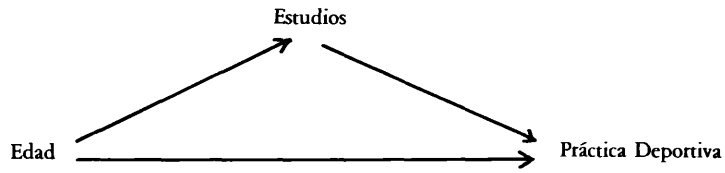


FIGURA 10. Grafo de la relación entre Edad, Estudios y Práctica deportiva.

Cuando tenemos que predecir los valores de más de una variable hay que proceder según el orden de causalidades. Por lo tanto, veamos primero la influencia de la Edad. Supongamos que en la próxima generación las propoprciones de individuos para cada grupo de edad sean .150, .320 y .530 —recordemos que los grupos de edad son 15-25, 26-40 y 40 y más años. ¿Qué consecuencias tendrá este envejecimiento? Para saberlo basta con cambiar la línea 1 de la tabla 17, sustituyendo los valores originales por los que acabamos de ofrecer, a la vez que calculamos los marginales de las variables consecuentes para los datos estandarizados. En la tabla 18 se ofrecen los resultados, siguiendo la presentación que hace James A. Davis (1982).

TABLA 18. Niveles ajustados para el sistema de la tabla 17.

		Edad			Estudios				Práctica deporte
Línea	Cambio	15-25	26-40	40 +	Primar.	Bachil.	G. Medio	Univer.	Sí
0	Datos brutos	.237	.286	.477	.625	.259	.068	.048	.281
	Después ajuste de:								
1	a) edad	.150	.320	.530	.659	.227	.068	.046	.246
2	b) estudios	.150	.320	.530	.400	.400	.150	.050	.308
	Cambio después de:								
3	c) etapa 1	−.087	.034	.053	.034	−.032	0	−.002	−.035
4	d) etapa 2	0	0	0	−.259	.173	.082	.004	.062
	Total	−.087	.034	.053	−.225	.141	.082	.002	.027

La línea 0 muestra los marginales originales para cada una de las tres variables. La línea 1 da los marginales después de ajustar la edad. En la línea 3 se ofrecen las diferencias entre las líneas 0 y 1. Según estos datos, el cambio de edad de la hipotética futura generación repercutirá en un aumento de 3.4 puntos en el número relativo de individuos con estudios primarios y una disminución de 3.2 y 0.2 puntos, respectivamente, en el de los bachilleres y universitarios; no sufrirá modificación la proporción de graduados medios. Como consecuencia de ese mismo envejecimiento cabe predecir una disminución de 3.5 puntos en el porcentaje de practicantes deportivos (línea 3).

La próxima variable que se modifica son los Estudios. Con segundas y posteriores variables el ajuste se complica algo, debido a que, además de tener que ajustar más de una línea es necesario tener en cuenta la relación original entre las variables. Supongamos que en la generación siguiente las proporciones de individuos con estudios primarios, bachilleres, grados medios y universitarios sean .400, .400, .150 y .050, respectivamente, y que la asociación entre Edad y Estudios se mantenga constante. Para ello, primero calculamos el marginal de la variable a ajustar (modificar), teniendo en cuenta el cambio experimentado por el ajuste previo de la Edad. Tomando el grupo de estudios primarios, en la línea 1 de la tabla 18 vemos que la proporción en cuestión es de .659; como queremos hacer que sea de .400, hemos de cambiar la proporción de estudios primarios en −.259 —igual haríamos con los otros niveles de estudios. Puesto que hemos determinado mantener constante el nivel de asociación entre Edad y Estudios, esta modificación de −.259 habrá que efectuarla sobre todos los niveles de Edad (líneas 2, 3 y 4 de la tabla 17). Después de efectuar la misma operación con todos los niveles de Estudios, estas líneas 2, 3 y 4 de la tabla 17 quedan como sigue (tabla 19).

TABLA 19. Cruce de edad con Estudios (proporciones), tras ajustar Estudios.

Edad	Primarios	Bachiller	Grado Medio	Universitarios	Total
15-25	(.328 − .259) = .069	.717	.145	.068	590
26-40	.292	.440	.187	.081	1259
40 +	.558	.286	.130	.026	2086
	.400	.400	.150	.050	3935

De esta forma, los marginales de Estudios toman el valor de .400, .400, .150 y .050, para los 4 niveles respectivos, sin que se altere la asociación entre las dos variables[10]. Ahora podemos proceder como en el caso previo, calculando los margina-

[10] Tomando la fila de estudios primarios se observa que el total de individuos en este nivel es de 1.573 —esto es, ((.069)(590) + (.292)(1.259) + (.558)(2.086))—, que dividido por 3.935 (el total de individuos) da .400. Igual se haría con las restantes filas.

Por otra parte, vemos en la tabla 17 que la diferencia entre la proporción de 15-25 y estudios primarios menos 26-40 y estudios primarios es de (.328-.551) = −.223; esta diferencia es la misma que la observada en la tabla ajustada (tabla 19): .069 − .292 = −.223. Por lo tanto, la intensidad de la asociación se mantiene.

les de Práctica Deportiva para los datos estandarizados. En las filas 2 y 4 de la tabla 18 se muestran los resultados obtenidos: la primera variable, Edad, por ser anterior al ajuste efectuado, no cambia; la proporción esperada de practicantes es de .308, lo que supone un aumento de 6.2 puntos. Analizando las filas 3 y 4 de la tabla se pueden resumir los resultados que hemos obtenido de la siguiente manera:

1. Cabe esperar una disminución de la población entre 15 y 25 años de casi 9 puntos (−.087). Por el contrario, aumentarán los individuos de 26 a 40 años (3.4 puntos) y de más de 40 años (5.3 puntos).
2. Las personas con estudios primarios disminuirán en 22.5 puntos. En este resultado hay una influencia positiva del envejecimiento de la población, que hace subir el porcentaje de personas con estudios primarios en 3.4 puntos, y una disminución del 25.9% no explicada por ninguna variable del modelo, que es atribuible a nuestros supuestos.

En el caso de los estudios universitarios —tomamos sólo este nivel y el anterior como ejemplos de la forma en que cabe interpretar los resultados— el envejecimiento de la población hace bajar en 0.2 puntos el porcentaje de individuos con este nivel, mientras que hay un aumento de 0.4 puntos en este colectivo que no es explicado. El resultado final es que en esta hipotética generación con la que estamos trabajando cabe esperar un 0.2% más de titulados superiores.

3. Como resultado de los diversos efectos, los deportistas aumentarán en 2.7 puntos. Desglosando este resultado final vemos en las líneas 3 y 4 que el envejecimiento hará disminuir en 3.5 puntos el número de practicantes, mientras que la mejora del nivel educativo de la población aumentará este colectivo en 6.2 puntos.

Vemos así cómo se puede proceder a resolver problemas de simulación con variables cualitativas, utilizando proporciones recurrentes y calculando una serie de tablas estandarizadas.

12.8. Resumen

A lo largo de las páginas de este artículo hemos mostrado cómo se puede ampliar el estudio de las variables nominales y ordinales, tratando de seguir los procedimientos utilizados cuando se dispone de información interval. Para ello hemos explicado la descomposición de una tabla multidimensional en una serie de ecuaciones y su representación en un grafo. El procedimiento seguido para estimar los coeficientes ha sido el cálculo de las proporciones recurrentes y su estandarización. Dichos coeficientes son susceptibles de ser sometidos a prueba estadística. Por último hemos mostrado cómo se pueden resolver ejercicios de simulación cuando se dispone de información cualitativa.

Como nota final, digamos que este trabajo de divulgación de la obra de James A. Davis queda incompleto al no poder abordar aquí la aplicación que el autor hace del sistema de la *D* al análisis del cambio —la «ola del futuro en la investigación social», según sus propias palabras (Davis 1975, 1976, 1978). En una futura publicación esperamos poder proceder a la explicación del estudio del cambio.

BIBLIOGRAFIA

ALWIN, D. F. y HAUSER, R. M.: «The decomposition of effects in path analysis». *American Sociological Review*, 40, 1975, pp. 37-47.

ALWIN, D. F. y JACKSON, D. J.: «Measurement models for response errors in surveys: issues and applications», en K. F. SCHUESSLER (ed.): *Sociological Methodology*, San Francisco, Jossey-Bass, 1980.

ARNOLD, J. B.: «A multidimensional scaling study of semantic distance». *Journal Exptl. of Psychology*, n.° 90, 1971, pp. 349-372.

ASHER, B. H.: *Causal modeling series: Quantitative Applications in the Social Sciences.* Sage Pub. Inc., 1976.

BAGOZZI, R. P.: *Causal models in Marketing.* Nueva York, John Wiley, 1977.

BAILEY, K. D.: «Cluster analysis», en D. R. HEISE (ed.): *Sociological Methodology.* San Francisco, 1975, pp. 59-127.

BALA KRISH NAN, V. y SANGHVI, L. D.: «Distance between populations on the basis of attribute data». *Biometrics*, n.° 24. 1968.

BARLOW, R. E.: *Statistical inference under order restrictions.* Londres, Wiley, 1972.

BATISTA, J. M.: «Modelo general para el análisis de ecuaciones de estructura lineal, LISREL. Relación con el modelo para el análisis de la estructura de la covarianza. Aplicación de un programa preventivo frente al consumo de drogas en España». Universidad Politécnica de Barcelona, Tesis doctoral, 1982.

— «Introducción al análisis factorial confirmatorio», en *Publicaciones de Bioestadística y Biomatemática*», n.° 10. Publicacions Edicions Universitat de Barcelona, 1983.

BATISTA, J. M. y CUADRAS, C.: «El análisis de la causalidad y el planteamiento LISREAL. Una presentación desde los modelos de medida». *QUESTIO*, universidad Politécnica de Barcelona.

BATISTA, J. M. y ESTIVILL, X.: «Definición de zonas homogéneas para la elaboración del plan territorial de Catalunya mediante técnicas de análisis multivariable». Dirección General de Política Territorial de la Generalitat de Catalunya, 1983.

BEALE, E. M. L.: *Cluster Analysis.* Londres. Scientific Control Systems. 1969.

BENJAMIN, B.: «Inter-generation differences in occupation», *Population Studies, II*, 1958, pp. 262-268.

BENTLER, P. M.: «Multivariate analysis with latent variables: causal modeling», en *Ann. Rev. Psychol.* n.° 31, 1980, pp. 419-456.

BENZECRI, J. P.: «L'analyse des données», tomo II, *L'analyse des correspondences*, París, Dunod, 1973.

— «La taxonomie», en *L'analyse des données*, tomo I. Duriod, 1979.

BERTIER, P. y BOUROCHE, J. M.: *Analyse des données multidimensionelles.* París, P.U.F., 1975.

BIRNBAUM, A. y MAXWELL, E. A.: «Classification procedures based on Baye's Formula Applied Statistics». 1966.

BISHOP, Y. M., FIENBERG y HOLLAND, P. W.: *Bisaete Multivariate Analysis: Theory and Practice,* Cambridge Univ. Press, 1975.

BLALOCK, H. M.: «Theory building and causal inferences», en H. M. BLALOCK y A. B. BLALOCK (eds.): *Methodology in Social Research,* Nueva York, McGraw Hill, 1968.

— «The measurement problem: A gap between the languages of theory and research», en H. M. BLALOCK y A . B. BLALOCK (eds.): *Methodology in Social Research.* Nueva York, McGraw-Hill, 1968.

— «Multiple indicators and the causal approach to measurement error». *American Journal of Sociology,* n.° 75, 1969, pp. 264-272.

— *Causal Inferences in Nonexperimental Research.* The Norton Library. 1961.

— *Estadística Social.* Madrid, F.C.E., 1980.

— «Estimating measurement error using multiple indicators and several points in time». *American Sociological Review,* n.° 35, 1970, pp. 101-111.

BOCK, R. D. y BARGMAN, R. E.: «Component of variance analysis as a structural and discriminal analysis for psychological tests». *British Journal of Statistical Psychologica,* n.° 8, 1960, pp. 151-163.

— «Analysis of covariance structures». *Psychometrika,* 31, 1966, pp. 507-534.

BOOSMA, A.: «The robutness of LISREL against small sample sizes in factor analysis models», en JÖRESKOG and WOLD, *Systems under indirect observation,* 1981.

BROWNE, M. W.: «Generalized least-squares estimators in the analysis of covariance structures», en D. J. AIGNER y A. S. GOLDBERGER (eds.): *Latent variables in Socio-Economic Models.* Amsterdam, North-Holland, 1977.

— «Covariance structures« en D. M. HAWKINS (ed.): *Topics in Applied Multivariate Analysis,* Cambridge University Press, 1982.

BROWN, M. B.: «Screening effects in multidimensional contingency tables». *Applications of Statistics,* n.° 25, 1976, pp. 37-46.

BYNNER, J. y ROMNEY, D.: «A method for overcoming the problem of concept-scale interaction in semantic differential research». *British Journal of Psichology,* n.° 63, 1972, pp. 229-234.

CAMPBELL, D. T. y FISKE, D. W.: «Convergent and discriminant validation by the multitrait-multimethod matrix», *Psychological Bulletin,* n.° 56, 1959, pp. 81-105.

CARROL, J. D.: «Individual differences and multidimensional scaling» en R. A. SHEPARD y otros: *Multidimensional scaling: Theory and applications in the behavioral sciences.* Vol. I. Nueva York, Seminar Press, 1972.

CARROL, J. D. y CHANG, J. J.: *Nonmetric multidimensional analysis of paired comparisons data,* Murray Hill, Mimeo, 1964.

— «Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-young" decomposition». *Psychometrika,* 35, 1970, pp. 283-319.

CARROLL, J. D. y WISH, M.: «Multidimensional perceptual models and measurement methods», en E. C. CARTERETTE y M. P. FRIEDMAN (eds.): *Handbook of perception.* Nueva York, Academic Press, 1975.

CATTELL, R. B.: «A note on correlation clusters and clusters search methods». *Psicometrika,* 1944.

COOMBS, C. H.: *A theory of data.* Nueva York, Wiley, 1964.

CORMACK, R. M.: «A review of classification». *J. Royal Stat. Soc.* (A) n.° 134, 1971, pp. 321-367.134.

COSTNER, H. L.: «Theory, deduction and rules of correspondence». *American Journal of Sociology*, n.° 75, 1969, pp. 245-263.

COSTNER, H. L. y SCHOENBERG, R.: «Diagnozing indicator ills in multiple indicator models» en A. S. GOLDBERGER y O. D. DUNCAN: *Structural Equation Models in the Social Sciences*. Nueva York, Seminar Press, 1973.

COXON, A. P. M.: «Reply to Jones». *Sociology*, n.° 6, 1972, pp. 453-455.

COXON, A. P. M. y JONES, C. L.: «Occupational similarities: some subjective aspects of social stratification». *Quality & Quantity* n.° 8, 1974, pp. 139-157.

CUADRAS, C. M.: *Métodos de análisis multivariante*. Barcelona, Eunibar, 1981.

DALING, J. R. y TAMURA, H. T.: «Use of ortogonal factors for selection of variables in a regression equation» *J.R.S.S.* (C), 1970, pp. 260-268.

DAVIS, J. A.: *Elementary Survey Analysis*. Nueva Jersey, Prentice Hall, 1971.

— «Hierarchical models for significance tests in multivariate contingency tables». *Sociological Methodology* 1973-1974. San Francisco. Jossey Bass, 1975.

— «Background characteristics in the US adult population 1952-1973: a survey metric model». *Social Science Research*, 5, 1976, pp. 349-383.

— «Studing categorical data overtime». *Social Science Research*, 7, 1978, pp. 151-179.

— «Communism, conformity, cohorts and categories». *American Journal of Sociology*, 81, 1975, pp. 491-513.

— «Contingency tables analysis: proportions and flow graphs». En JAMES ALT (ed.): «Advances in Quantitative Analysis». *Quality and Quantity*, 1. 1980, pp. 117-155.

— «Analyzing contingency tables with linear flow graphs: D systems», en D. R. HEISE (ed.): *Sociological Methodology*, San Francisco. Jossey Bass, 1976.

— «Extending Rosenberg's Technique for Standardizing Percentage Tables». Artículo sin publicar, 1982.

DIXON, W. J.: *BMDP: Biomedical Computer Programs*. Los Angeles. UCLA, 1975.

DRENTH, P. J. D.; PETRIE, J. P. y BLEICHRODT, N.: *AKIT: Amsterdamse Kinder Intelligentie Test*. Amsterdam. Swets & Zeitlinger, 1968.

DUNCAN, O. D.: «Path analysis: sociological examples». *American Journal of Sociology*, n.° 72, 1966, pp. 1-16.

— *Introduction to structural equation Models*. Nueva York, Academic Press, 1977.

ECKART, C. y YOUNG, G.: «Approximation of one matrix by another of lower rank», *Psicometrika*, n.° 1, 1936, pp. 211-218.

EISLER, H. y ROSKAM, E. E.: «Multidimensional similarity: an experimental study of vector, distance and set theoretical models». Nijmegen, Dept. of Psychology. *Mimeo*, 1973.

EVERITT, B.: *Cluster Analysis*. Londres, H.E.B., 1974.

— «Cluster Analysis» en varios autores *The analysis of surveydata*. Chichester, John Wiley & Sons, 1977.

— «Cluster Analysis» en J. ALT: *Advances in quantitative analysis*. N..° especial Quality & Quantity, vol. 14, n.° 1, 1980.

— «Notas del curso de Análisis de Cluster». Universidad de Essex, 1982.

EVERITT, B.; GOURLAY, A. J. y KENDELL, R. E.: «An attemp at validation of traditional psychiatric syndromes by Cluster Analysis». *British Jr. of Psychiatry*, n.° 119, 1971, pp. 299-412.

FARRIS, J. S.: «On the cophenetic correlation coefficient». *Systematic Zoology*, n.° 18, 1969.

FIEMBERG, S. E.: *The analysis of crossclassified data*. Cambridge. MIT Press, 1977.

FISHER, F. M.: *The identification problem in econometric*. Nueva York, Mc-Graw Hill, 1966.

— «The use of multiple measurement in taxonomic problems». *Aun. Eugenics*, n.° 7, 1936.

FLETCHER, R.; POWELL, M. J. D.: «A rapidly convergent descent method for minimization». *Comput Journal*, n.° 2, 1963, pp. 163-168.

García Ferrando, Manuel: *Socioestadística: Introducción a la Estadística en Sociología.* Madrid. Centro de Investigaciones Sociológicas, 1984.
Goldberger, A. S.: *Teoría econométrica.* Tecnos. Madrid, 1970.
Goodman, L. A.: «The multivariate analysis of qualitative data: interactions among multiple classifications». *Journal American Statistical,* n.° 65, 1970, pp. 226-256.
— «A modified multiple regression approach to the analysis of dichotomous variables». *American Sociology Review,* n.° 37, 1972. pp. 28-46.
— «A general model for the analysis of surveys». *American Journal of Sociology,* n.° 77, 1972, pp. 1035-1086.
— «Causal analysis of data from panel studies and other kinds of surveys». *American Journal of Sociology,* n.° 78, 1973, pp. 1135-1191.
— «The analysis of multidimensional contingency tables when some variables are posterior to others: a modified path analysis approach». *Biometrika,* n.° 60, 1973, pp. 178-192.
Gower, J. C.: «A general coefficient of similarity and some of its propierties». *Biometrics,* 1971.
Gower, J. C. y Ross: «Minimum spanning trees and single linkage cluster analysis». *Applied Statistics,* 1969.
Green, P. E., Frank, R. E. y Robinson, P. J.: «Cluster analysis in test market selection». *Management Sciences,* n.° 13, 1967.
Guttman, L. A.: «A new approach to factor analysis», en P. F. Lazarsfeld (ed.): *Mathematical Thinking in the Social Sciences.* Columbia Univ. Press, 1954.
— *The quantification of a class of atribute.* Nueva York. Social Science Research Council, 1941.
Haberman, S. J.: «Long linear fit for contingency tables». *Applied Statistics,* n.° 21, 1972, pp. 218-225.
Hanushek, E. A. y Jackson, J. E.: *Statistical Methods for Social Scientists.* Academic Press. Nueva York, 1977.
Harman, H. H.: *Modern Factor Analysis.* Chicago. university of Chicago Press, 1975.
Harrison, I.: «Cluster Analysis», en *Metra 7,* 1968.
Harrison, P. J.: «A method of cluster analysis and some applications». *Applied Statistics.* 1968.
Hartigan, J. A.: *Clustering algorithms.* Nueva York, Wiley, 1975.
Hausman, J. A.: «Specification and estimation of simultaneous Equation Models» en Griliches, Z. e Intriligator, M. D. (eds.): *Handbook of Econometrics,* Amsterdam, North-Holland, 1983.
Heise, D. R.: *Causal Analysis.* Nueva York, John Wiley & Sons, 1969.
Hildebrand, D. K. y otros: *Analysis of ordinal data.* Londres, Sage, 1977.
Hodson, R. R.; Kendall, D. G. y Tautu, P.: *Mathematics in Archeological and Historical Sciences.* Edimburgo University Press, 1971.
Horan, C. B.: «Multidimensional scaling: Combinig observations when individuals have different perceptual structures». *Psychometrika* 34, 1969, pp. 139-165.
Horst, P.: *Factor analysis of data matrices.* Holt, Rinehart & Winston, 1965.
Hotelling, H.: «Analysis of complex statistical variables into principal components». *Journal. Educ. Psychol.* n.° 24, 1933, pp. 417-441.
Jambu, M.: Classification automatique pour l'analyse des données. Dunod, 1978.
Jardine, N. y Sibson, R.: *Mathematical taxonomy.* Nueva York, Wiley, 1971.
Johnston, J.: *Métodos econométricos.* Vicens Vives, 1963.
Jollife, I. T.: «Discarding variables in a principal components Analysis». *J.R.S.S.* (c), n.° 21, 1972, pp. 160-173 y n.° 22, 1973, pp. 21-31.
Jones, C. L.: «On occupational attributes». *Sociology,* 6, 1972, pp. 451-452.

JONES, C. L. y MCPHERSON, A. F.: «Implications of non-response to postal surveys». *Scot. Ed. Studies*, n.º 4, 1964, pp. 143-151.

JÖRESKOG, K. G.: «Estimation and testing of simplex models». *British Journal of Mathematical and Statistical Psichology*, n.º 23, 1970, pp. 121-145.

— «A general approach to confirmatory factor analysis». *Psychometrika*, n.º 34, 1969, pp. 183-202.

— «Analysis of covariance structures». *Scand. J. Statist.* n.º 8, 1981, pp. 65-92.

— «Factor analysis by least-squares and maximum-likelihood methods» en varios autores; *Statistical Methods for Digital Computers.* J. Wiley, 1977.

— «A general method for estimating a linear structural equation System», en A. S. GOLDBERGER y O. D. DUNCAN (eds.): *Structural Equation Models in the Social Sciences*, Seminar Press, 1973.

— «A general method for analysis of covariance structures». *Biometrika*, 57, 1970, pp. 239-251.

— «Structural equation models in the Social Sciences; specification, estimation and testing», en P. R. KRISHNAIAH (ed.): *Applications of statistics*, Amsterdam, North Holland, 1977.

— «Structural analysis of covariance and correlation matrices». *Psychometrika*, n.º 43, 1978, pp. 443-477.

— «Analyzing psychological data by structural analysis of covariance matrices», en C. KRAUTZ, et. al.: *Contemporary Developments in Mathematical Psychology.* Freeman, 1974. Vol. II.

— «Statistical analysis of sets of congeneric tests». *Psychometrika*, n.º 36, 1971, pp. 109-133.

— «A general computer program for analysis of covariance structures». *Biometrika*, n.º 57, 1970, pp. 239-251.

JÖRESKOG, K. G.; GRUVAENS, G. y VAM THILLO, M.: «A general computer program for analysis of covariance structures». *Research Bulletin*, Princeton, 1970.

JÖRESKOG, K. G. y LAWLEY, D. N.: «New methods in maximum likelihood Factor Analysis». *British Journal Math. Statistical Psychology*, n.º 21, 1968, pp. 85-96.

JÖRESKOG, K. G. y SORBOM, D.: *LISREL IV: A general computer program for estimation of a linear structural equation by maximum-likelihood methods.* Chicago International Educational Services, 1978.

— *LISREL V: Analysis of linear structural relationships by Maximum Likelihood and Least Squares Methods.* Chicago, National Educational Services. 1983.

— «Statistical models and methods for analysis of longitudinal data». En AIGUER, D. J. y GOLDBERGER, A. S. (eds.): *Latent Variables in Socio-Economic Models.* Amsterdam, North-Holland, 1977.

JÖRESKOG, K. G. y VAN THILLO, M.: «LISREL: A general computer program for estimating Linear Structural Equation System involving multiple indicators of unmeasured variables», *Research Report 73-5.* Statistics Department, Uppsala Univ. 1973.

KAISER, H. F.: «The Varimax criterion for analytic rotation in factor analysis». *Psychometrika*, n.º 23, 1958, pp. 187-200.

KAMP, L. J. y MELLENBERGH, G. J.: «Agreement between raters». *Educational and Psychological Measurement*, n.º 36, 1976, pp. 311-317.

KELLEY, E. L. y FISKE, D. M.: *The prediction of performance in clinical psicology.* Ann Arbor. University of Michigan, 1959.

KLECKA, W. R.: «Discriminant analysis», en NIE, N. y otros: *SPSS: Statistical Package for the Social Sciences.* Nueva York, Mc-Graw Hill, 1975.

— *Discriminant analysis.* Beverly Hills y Londres. Sage, 1980.

KNOKE, D. y BURKE, P. J.: *Log-Linear Models.* Beverly Hills y Londres. Sage, 1980.

KRISHNAIAH, P. R.: *Multivariate analysis*, Nueva York, Academic Press, 1966.

KRITZER, H. M.: «Approaches to the analysis of complex contingency tables: a guide for the perplexex». *Sociological Methods and Research*, n.º 3 (febrero), 1979, pp. 305-329.
KRUSKAL, J. B.: «Multidimensional Scaling by optimizing goodness of fit to a nonmetric hypothesis». *Psychometrika*, n.º 29, 1964, pp. 1-27.
— «Nonmetric multidimensional scaling: a numerical method». *Psychometrika*, n.º 29, 1964, pp. 115-129.
— «Geometric interpretation of diagnostic data for a digital machine». *Bell. Syst. Tech. J.*, n.º 45, 1966, pp. 1299-1338.
KRUSKAL, J. B. y WISH, M.: *Multidimensional Scaling*. Beverly Hills y Londres. Sage, 1978.
KRUSKAL, J. B. y SHEPARD, R. N.: «A nonmetric variety of linear factor analysis». *Psychometrika*, n.º 39, 1974, pp. 123-157.
LAND, K. C.: «Significant others, the self-reflexive act and the attitude formation process: a reinterpretation». *American Sociological Review*, vol. 36, 1971, 1085-1098.
LAWLEY, D. N.: «The estimation of factor loadings by the maximum-Likelihood method», *Proc. R. Soc. Edind.* A 60, 1940, pp. 64-82.
LAWLEY, D. N.; MAXWELL, A. E.: *Factor Analysis as a statistical Method*. Butterworth and Co., 1971.
LAZARSFELD, P. F. y ROSENBERG, M.: *The Languaje of Social Research: a Reader in the Methodology of Social Research*. Nueva York, Free Press, 1955.
LEBART, L.; MORINEAU, A. y TABARD, N.: *Techniques de la description statistique, methodes et logiciels pour l'analyse des grands tableaux*. París, Dunod, 1977.
LEBART, L. y MORINEAU, A.: *SPAD, Système portable pour l'analyse des données*. París, Cesia, 1982.
LEBART, L.; MORINEAU, A. y FENELON, J. P.: *Traitement des données statistiques*. París, Dunod, 1979.
LEBART, L. y FENELON, J. P.: *Statistique et informatique appliquées*. París, Dunod, 1975.
LERMAN, I. C.: *On two criteria of classification*. Nueva York, Academic Press, 1969.
— *Les bases de la classification automatique*. París, Gauthier-Villers, 1970.
LEVINE, J. H.: «The sphere of influence», *American Sociological Review*, n.º 37, 1972, pp. 14-27.
LINGOES, J. C.: «A general survey of the Guttman-Lingoes nonmetric program series» en SHEPARD et al., vol. I, 1972.
LORD, F. M. y NOVICK, M. E.: *Statistical Theories of mental test scores*. Addison Wesley, 1968.
MADDALA, G. S.: *Econometrics*. McGraw-Hill. Nueva York, 1970.
MCDONALD, K. I.: «MDSCAL and distances between socioeconomic groups», en K. HOPE (ed.): *The analysis of Social mobility. Methods and Approaches*. Oxford. Clarendon Press, 1972, pp. 211-234.
MCRAE, D. JR.: *Issues and Parties in Legislative Voting*. Nueva York. Harper & Row, 1970.
MAHALANOBIS, P. C.: *On the generalized distance in statistics*, 1936.
MARSDEN, P. V. (ed.): *Linear Models in Social Research*. Londres y Beverly Hills, Sage, 1971.
MAXWELL, A. E.: *Multivariate analysis in behavioural research*. Chapman & Hall, 1977.
MICHENER, C. C. y SOKAL, R. R.: *A quantitative approach to a problem in classification evolution*, 1957.
MORRISON, D. F.: *Multivariate statistical methods*. Nueva York, McGraw-Hill, 1976.
MOSER, C. A. y KALTON, G.: *Survey methods in Social Investigation*. Londres, Heinemann Ed. Books, 1977.
MOSER, C. A. y WOLF SCOTT: *British Towns, a statistical study of their social and economic differences*. Oliver & Boy Ltd., 1961.
MULAIK, S. A.: *The foundations of factor analysis*. Nueva York, McGraw-Hill, 1972.

MUTHEN, B.: «Contributions to factor analysis of dichotomous variables». *Psychometrika*, n.º 43, 1978, pp. 551-560.

NAPIOR, D.: «Nonmetric multidimensional techniques for summated ratings», en SHEPARD et al., vol. I, 1972.

NEEDHAM, R. M.: *Computer methods for classification and grouping.* Londres. Mouton & Co. 1965.

NELDER, J. A. y WEDDERBURN, M.: «Generalized Linear Models», *Journal Royal Statistic of Sociology*, n.º 135, 1972, pp. 370-384.

NETHERLANDS INTERUNIVERSITY DEMOGRAFIC INSTITUTE: «Netherlands survey on fertility and parenthood motivation», 1975.

NIE, N. H., HULL, C. H., JENKINS, J. G., STEINBRENNER, K. y BENT, D. H.: *SPSS, Statistical Package for the Social Sciences.* Nueva York, McGraw-Hill, 1975.

NIGEL GILBERT, G.: *Modelling Society: an Introduction to Log-Linear Analysis for Social Researchers.* Londres, George Allen & Unwin, 1981.

PAYNE, C.: «The log-linear model for contingency tables» en O'MUIRCHEARTAIGH y PAYNE (EDS.): *The Analysis of Survey Data.* Londres, Wiley, 1977, vol. II, pp. 105-145.

PIPJER, W. M. y SANIS, W. E.: «The effect of identification restrictions on the test statistics in latent variables models». Ponencia presentada en la Conferencia sobre *Sistemas bajo observaciones indirectas.* Ginebra, 1979.

PULIDO: *Modelos econométricos.* Madrid, Pirámide, 1983.

RABINOWITZ, G. B.: «An introduction to non-metric multidimensional scaling», *American Journal of Political Science*, vol. 19, pp. 343-390, 1975.

RAO, C. R.: *Advanced statistical methods in biometric research.* Nueva York, Wiley, 1952.

REUCHLIN, M.: *Méthodes d'analyse factoriel a l'usage des psychologes.* Presses Universitaires de France, 1964.

REYNOLDS, H. T.: *Analysis of nominal data.* Londres y Beverly Hills, Sage, 1977.

ROSENBERG, M.: *The logic of survey analysis.* Nueva York, Basic Books, 1968.

— «Test factor standarization as a method of interpretation», *Social Forces*, n.º 41, 1962, pp. 53-61.

SÁNCHEZ CARRIÓN, J. J.: «Introducción al análisis multidimensional no métrico». *Revista Española de Investigaciones Sociológicas* (en prensa), 1985.

— *The Analysis of Tabular Data Comparing log-linear models and D-systems.* Research Project Diploma. Universidad de Essex, 1983.

SANDERS, D.: «Path analysis and causal modelling», en J. ALT (ed.): *Advances in Quantitative Analysis. Quality & Quantity*, n.º 1, enero 1980.

SARIS, W. E.: «Different questions, different variables», en FORNELL: *A Second Generation of Multivariate Analysis.* Vol. II. Nueva York, Praeger Publishers, 1983.

— *The use of linear structural equation models in non-experimental research*, vol. I y II. Report of the Department of Methodology, Free University of Amsterdam, 1978.

— «Linear structural relationships», en J. ALT (ed.): *Advances in Quantitative Analysis. Quality and Quantity*, n.º 14, enero 1980.

SARIS, W. E.; VAN LOORN, L. y LODGE, M.: «How certain are the resultats of voting studies», en SARIS, W. E. (ed.): *Linear structural relationships. Measurement models* (vol. II). Free University of Amsterdam, 1982.

SARIS, W. E. y STRONKHORST, L. H.: *Introduction to causal Models in non-experimental Research.* Amsterdam, S.S.O., 1984.

SARIS, W. E.; DEN RONDEN, J. y SATORRA, A.: «Testing structural equation models», en CUTTANCE, P. (ed.): *Structural Equation Models*, 1984.

SATORRA, A.: *Potència del contrast de la Ráo de Versemblança en Models d'equacions estructurals.* Barcelona, Tesis Doctoral. Facultad de Matemáticas, 1983.

SCHÖNEMANN, P. H. y WANG, N. M.: «An individual difference model for the multidimensional analysis of preference data», *Psychometrika*, n.° 37, 1972, pp. 275-309.
SHEPARD, R. N.: «The analysis of proximities: multidimensional scaling with an unknown distance function», *Psychometrika*, n.° 27, 1962, pp. 125-139 y 219-246.
— «On subjectively optimum selections among multiatribute alternatives», en M. W. SHELLEY y G. L. BRYAN: *Human Judgements and Optimality*. Nueva York, Wiley, 1964.
— «Metric structures in ordinal data», *Journal of Mathematical Psychology*, n.° 3, 1966, pp. 287-315.
SHEPARD, R. N.; ROMNEY, A. K y NERLOVE, S. B.: *Multidimensional scaling: theory and applications in the Behavioral Sciences*. Vol. I. Nueva York, Academic Press, 1972.
SHERMAN, C. R.: «Nonmetric multidimensional scaling: a Monte Carlo study of the basic parameters», *Psychometrika*, n.° 37, 1972, pp. 323-355.
SNEATH, P. H. A.: «Recent developments in theoretical and quantitative taxonomy». *Sistematic Zoology*, 1961.
SIBSON, R.: «Order-invariant methods for data analysis», *Journal Royal Statistica Sociologica* (B), n.° 34, 1972, pp. 311-349.
SIMON, H.: «Spurious correlation: a causal interpretation». *Models of man*. Nueva York, Wiley, 1957.
SIMPSON, E. H.: «The interpretation of interaction in contingency tables». *Journal Royal Statistica Sociologica* (A), 1951, pp. 238-241.
SOKAL, R. R. ŷ ROHLF, F. J.: *The comparison of dendograms by objective methods taxon*, 1962.
SOKAL, R. R. y SNEATH, P. H. A.: *Principles of numerical taxonomy*. Freeman, Londres, 1963.
SPARKS, D. N.: «Euclidean cluster analysis». *Applied Statistics*, 1970.
SPEARMAN, C. H.: «General intelligence objectively determined and measured». *American Journal of Psichology*, n.° 15, 1904, pp. 201-293.
STEVENS, S. S.: *Psychophisics: Introduction into its perceptual, neural and social prospects*. J. Wiley, 1975.
— «A metric for the Social Consensus», *Science*, n.° 151, 1966, pp. 530-541.
— «On the theory of scales of measurements», *Science*, vol. 103, n..° 2.684, 1946.
STINCHCOMBE, A. L.: *Constructing Social Theories*. Nueva York, Harcourt Brace Jovanovich, 1958.
STRONKHORST, L. H. y SATORRA, A.: *A fertility study: causal modeling approach*. Próxima publicación.
THOMAS, J. J.: *Introducción al análisis estadístico para economistas*. Barcelona. Boixareu Editores, 1980.
THURSTONE, L. L.: «Multiple factor analysis», *Psychological Review*, n.° 38, 1931, pp. 406-407.
— *The vector of mind*. Chicago. University of Chicago Press, 1935.
TORGERSON, W. S.: *Theory and methods of scaling*. Londres, Wiley, 1958.
TRYON, R. C. y BAILEY: *Cluster analysis*. Nueva York, McGraw-Hill, 1970.
UPTON, G.: *The analysis of cross-tabulated data*. Chichester, Wiley, 1978.
— «Contingency table analysis: log-linear models», en J. ALT (ed.): *Advances in Quantitative Analysis. Quality and Quantity*, 1. Enero 1980, pp. 155-180.
— «Log-Linear models, screening and regional industrial surveys». *Reg. Studies*, n.° 15, 1981, pp. 33-45.
VAN DE GEER: *Introduction to multivariate analysis for the Social Sciences*. Londres, Freeman, 1971.
WERTS, C. E.; JÖRESKOG, K. G. y LINN, R. L.: «A multitrait-multimethod model for studying growth». *Educational & Psychological Measurement*, n.° 33, 1973, pp. 655-678.

WERTS, C. E.; LINN, R. L. y JÖRESKOG, K. G.: «Quantifying unmeasured variables», en BLALOCK, H. M. (ED.): *Measurement in the Social Sciences.* McMillan, 1974, pp. 270-293.

WILEY, D. E. y WILEY, J. A.: «The estimation of measurement error in panel data». *American Sociological Review,* n.° 35, 1970, pp. 112-117.

WILLIAMS, W. T. y LAMBERT, J. M.: *Multivariate Statistical Methods, among-groups covariation.* Dowden Hutchinsob Ross, 1975.

WISH, M.; DEUTSCH, M. y BIENER, L.: «Differences in conceptual structures of nations». *J. Pers. and Soc. Psychol.,* n.° 16, 1970, 361-373.

WOLD, H.: «Econometric Model Building», en *Essays on the Causal Chain Approach.* North-Holland, 1964.

WONNACOT, T. H. y WONNACOTT, R. J.: *Introductory Statistics.* Londres, Wiley & Sons, 1977.

WRIGHT, S.: «The methods of Path Coefficients». *Ann. Math. Stat.,* 5, 1934, pp. 161-215.

ZAAL, J.: *Sociaal emotioneel gedrag van kinderen van 5-7 jaar beoordeeld door de leerkrachten.* Tesis doctoral, inédita, 1978.