UNA REVISIÓN DE MODELOS GRÁFICOS ENCADENADOS PARA TABLAS DE CONTINGENCIA MULTIDIMENSIONALES

Claudio Rafael Castro López* y Purificación GalindoVillardón[1]**
* Facultad de Estadística e Informática, Universidad Veracruzana, México
** Departamento de Estadística, Universidad de Salamanca, España

RESUMEN
En el marco de los modelos gráficos encadenados, se presenta una descripción de los conceptos en los que está basada esta metodología de análisis estadístico, para modelizar una tabla de contingencia multidimensional con variables respuesta. Se abordan los modelos gráficos encadenados, sus propiedades, estimación y búsqueda del modelo. Se recurre a datos producidos por un estudio de opinión, para ilustrar una aplicación y mostrar las potencialidades de aplicación, que tiene la modelización gráfica.

ABSTRACT
Within the framework of the graphical chain models of Markov, a description of the concepts that inspire this methodology of statistical analysis, with respect to modeling a table of multidimensional contingency with response variables. The chain graphical models, their properties, estimation and search of the model are approached. One resorts to the data produced in an opinion study in where a set of variables is used to show the application potentialities that the graphical modeling has.

Key words: Contingency table, log-linear model, graphical chain model, properties of Markov, EH algorithm, opinion study.

MSC 62J12

1 INTRODUCCIÓN

En este trabajo presentamos los modelos gráficos encadenados, como una alternativa al análisis de tablas de contingencia multidimensionales.

El potencial de los modelos log-lineales jerárquicos para detectar interrelaciones entre variables de una tabla de contingencia y la relación que estos modelos tienen con los modelos gráficos, permite un esquema de análisis interesante y con consistentes bases matemáticas.

El objetivo de este artículo es mostrar el potencial de los modelos gráficos encadenados en el contexto del análisis de tablas de contingencia multidimensionales. Para ello en la sección 2, incorporamos alguna notación básica acerca de la teoría de grafos y los modelos gráficos, revisamos algunas de las características y propiedades de la modelización gráfica, sus vínculos a los modelos log-lineales y en este contexto definimos la modelización con grafos encadenados en el marco de las propiedades de Markov. En la sección 3 caracterizamos los modelos gráficos encadenados. En la sección 4 presentamos una aplicación de modelización de una tabla *p*-dimensional con modelos gráficos

[1] Email ccastro@uv.mx, pgalindo@usal.es

encadenados. En la sección 5 se presenta una panorámica del software que es posible emplear para modelización gráfica y en la sección 6 se presenta la conclusión de este trabajo.

2 TEORÍA DE GRAFOS Y SU VÍNCULO CON TABLAS DE CONTINGENCIA MULTIDIMENSIONALES

La estructura de asociación subyacente entre un conjunto Γ de variables categóricas, contenidas en una tabla de contingencia *p*-dimensional, puede ser convenientemente descrito en base a la estructura de independencia condicional por los llamados modelos log-lineales jerárquicos, los cuales plantean una relación lineal entre las frecuencias esperadas para las celdas de la tabla de contingencia (CHRISTENSEN, 1990); (AGRESTI, 1990); (WERMUTH y COX, 1998). Los modelos log-lineales jerárquicos permiten definir la clase generadora del modelo y su interpretación en términos de independencia condicional. Así, tratándose del modelo [***ij***][***jk***], nos referimos a la independencia condicional ($i \coprod k / j$). Cuando el número de variables se incrementa, los parámetros a estimar aumentan de manera considerable haciendo que el análisis y la estimación de los modelos log lineales sea poco practica (WHITTAKER 1990). Sin embargo, es posible plantear una fuerte relación entre estos modelos y los modelos gráficos.

Un *grafo* G es un objeto matemático que se forma de un conjunto de vértices **V** y un conjunto de bordes **E** (dirigidos o no), que representa pares de elementos de **V**, denotamos el grafo por: G(**V,E**). Podemos decir que G(**V,E**) es una representación gráfica, en la cual todos los *vértices* están caracterizados por pequeños círculos ó puntos (• para identificar variables categóricas), y líneas (*bordes*), no necesariamente direccionales, que unen los vértices. Si todo par de vértices esta unido por un borde, lo denominamos completo. Un conjunto completo de vértices, que no puede ser ampliado a un conjunto completo mayor por la adición de más vértices se denomina un *clique* maximal. Un modelo log-lineal puede ser descrito por un grafo a través de la identificación de la clase generadora del modelo, conjunto de todos los cliques del grafo (ASSMUSEN y EDWARDS, 1983).

Los modelos que son generados por esta vía son llamados modelos gráficos y pueden ser interpretados en términos de independencia condicional. En general, el grafo de independencia condicional de **V** es el grafo **G**{**V,E**}, donde **V** = {***i, j, k***,……} y el par (***i, j***) no está en el conjunto **E**, si y solo si $i \coprod j / \mathbf{V} \setminus (i, j)$.

Las propiedades de Markov permiten establecer la independencia condicional en un modelo gráfico. En base a la propiedad de global, también denominada de separación, se establece que si para **A**, **B** y **S** subconjuntos de **V**, donde $\mathbf{A} \cup \mathbf{B} \cup \mathbf{S} = \mathbf{V}$, en el grafo **G** de independencia, cada vértice en **A** esta separado de cada vértice de **B** por un vértice de **S**, entonces $\mathbf{A} \coprod \mathbf{B} / \mathbf{S}$. De esta propiedad se deriva la afirmación ***i*** es independiente de ***j***, dado el resto de las variables, esto es $i \coprod j / \mathbf{V} \setminus (i, j)$ (LAURITZEN, 1996).

Para la selección del modelo gráfico, que denotamos por **M**(G), existen diferentes estrategias, las cuales son semejantes a los procedimientos de estimación en los modelos log-lineales (AGRESTI, 1990); una de ellas es el método paso a paso, en donde se inicia a partir de un modelo (generalmente saturado) y se eliminan bordes hasta que algún criterio se cumple. Otra alternativa es seleccionar el modelo que optimiza uno de los llamados criterios de información AIC (*Akaike's Information Criterion*) o BIC (*Bayesian Information Criterion*), (EDWARDS 2000). Las estimaciones máximo verosímiles (MLEs) para seleccionar el modelo que mejor se ajusta a los datos, pueden ser obtenidas de los llamados modelos descomponibles (DARROCH, LAURITZEN y SPEED, 1980). Dos subconjuntos de vértices **A** y **B** de G(**V,E**), forman una descomposición del grafo, si **A** U **B**

= **V**, de tal modo que **A** y **B** están separados por **A** ∩ **B** en el grafo, y **A** ∩ **B** es un subconjunto completo de vértices. En palabras, dos subgrafos forman una descomposición **A** ∩ **B** = **C** de los vértices del grafo, si el grafo es la unión de dos subgrafos y la intersección es un subgrafo completo. Así el conjunto **C** es un *clique* separador.

En resumen, si un grafo y sus subgrafos pueden ser descompuestos recursivamente, entonces decimos que el grafo es reducible o descomponible, hasta que todos los subgrafos sean completos. Cuando el grafo o subgrafo es no separable o irreducible, forma lo que se denomina, una componente irreducible o subgrafo *maximal*.

En frecuentes ocasiones, cuando se tiene una tabla multidimensional, estamos interesados en las relaciones que puedan existir entre variables explicativas y una o mas variables respuesta, de tal manera que las variables explicativas nos interesan en la medida de que éstas sean capaces de explicar efectivamente la respuesta. En este contexto, se forma un tipo de modelos gráficos, donde los bordes están representados por flechas, es decir, que la existencia de un borde, además de indicar la relación entre las dos variables, indica la existencia de una influencia de una de las variables con respecto de la otra. Se forman así los llamados grafos con bordes dirigidos, los cuales permiten ampliar las posibilidades de análisis y conocer con mayor precisión la estructura de asociación entre las variables.

Los grafos con bordes dirigidos (flechas), producen los denominados modelos gráficos acíclicos dirigidos **M**(DAGs), también conocidos como diagramas de influencia probabilística (LAURITZEN, 1996 y 1999); (EDWARDS, 2000). A diferencia de los grafos con bordes no dirigidos, en donde el orden de los vértices no tiene importancia, en los grafos con bordes dirigidos, los vértices consideran un orden, tal que $\boldsymbol{v}_i \rightarrow \boldsymbol{v}_j$ solo cuando i < j. De esta forma, si dos vértices están conectados, damos por supuesto que existe un orden entre ellos del menor al mayor. Si suponemos que dentro del conjunto de vértices **V**, tenemos una secuencia $(\boldsymbol{v}_1,.....\boldsymbol{v}_k)$, un camino dirigido del vértice $\boldsymbol{v}_1$ al vértice $\boldsymbol{v}_k$ requiere que $\boldsymbol{v}_i$ y $\boldsymbol{v}_{i+1}$ sean adyacentes, es decir los vértices estén conectados por un borde dirigido, para cada i = 1,……k-1. Si hay un camino dirigido de ***i*** a ***j***, entonces ***i*** es llamado un *ancestro* de ***j*** y ***j*** es llamado un *descendiente* de ***i***. El conjunto de *ancestros* es denotado por *an*(***j***) y el conjunto de *descendientes* por *de*(***i***). El conjunto ancestral an^+ (**C**) de **C** **V**, se obtiene de considerar: Si hay un camino de ***i*** a ***j***, entonces ***i*** es un *ancestro* de ***j***. Denotamos el conjunto de vértices ancestros de **C** por *an*(**C**) y por tanto an^+(**C**) = **C** □ *an*(**C**), se denomina el *conjunto ancestral* de **C** (EDWARDS 2000).

Cuando el primero y último vértices coinciden, es decir $\boldsymbol{v}_1 = \boldsymbol{v}_k$, entonces el camino es llamado un ciclo dirigido. Los modelos gráficos con bordes dirigidos, denotados por **M**(DAG), se desarrollan con grafos que no tienen ciclos, ya que no podemos pensar en que una variable respuesta sea explicativa de si misma y además no existe una distribución de probabilidad para modelizar esta situación. Por tanto, un grafo que contiene ciclos dirigidos no representa una versión dirigida de un grafo de independencia. Según EDWARDS (2000), la distribución de probabilidad *p* para un **M**(DAG) se explica como:

$$p(\boldsymbol{v}_1,.....\boldsymbol{v}_n) = p(\boldsymbol{v}_1)p(\boldsymbol{v}_2/\boldsymbol{v}_1).....p(\boldsymbol{v}_n/\boldsymbol{v}_{n-1}\ \boldsymbol{v}_{n-2}..... \boldsymbol{v}_1)$$

En base a esto, como un DAG no tiene ciclos completos y un borde dirigido de $\boldsymbol{v}_i$ a $\boldsymbol{v}_j$ existe a menos que $p(\boldsymbol{v}_j|\boldsymbol{v}_{j-1} ..... \boldsymbol{v}_1)$, no dependa de $\boldsymbol{v}_i$, Se sigue que:

$$\boldsymbol{v}_i \coprod \boldsymbol{v}_j / \{\boldsymbol{v}_1,.....\boldsymbol{v}_j\} \setminus \{\boldsymbol{v}_i , \boldsymbol{v}_j\}$$

Por tanto una relación de independencia condicional entre $\boldsymbol{v}_i$ y $\boldsymbol{v}_j$, se establece con las variables anteriores al par $(\boldsymbol{v}_i,\boldsymbol{v}_j)$ de variables. En otras palabras, mientras en un grafo no dirigido, la independencia condicional se establece en base a una única distribución, en los DAGs se establece en base a una secuencia de distribuciones marginales (WHITTAKER, 1990).

El grafo moral (*moral graph*) para DAGs (LAURITZEN y WERMUTH, 1989), permite identificar la propiedad de *d*-separación de Markov, que establece si **A**, **B** y **S** son tres conjuntos disjuntos de **V** vértices en un DAG determinado, decimos que $\mathbf{A} \coprod \mathbf{B} / \mathbf{S}$, siempre que **A** y **B** estén separados por el conjunto de vértices **S** en el grafo moral G^m, (grafo con bordes no dirigidos formado por el criterio de moralización), del menor subconjunto ancestral que contiene a los vértices **A**, **B** y **S** El proceso de estimación y búsqueda del modelo para los datos, se plantea asumiendo que DAG_0 es un subgrafo del DAG, obtenido de eliminar uno o más bordes de éste. El modelo $\mathbf{M}(DAG_0)$ es un submodelo del modelo **M**(DAG) y con esto podemos considerar a la razón de verosimilitud, como un estadístico de prueba para la hipótesis de la distribución extendida de probabilidad de Markov *p*, factorizada de acuerdo a un grafo con bordes dirigidos tales que $p \in \mathbf{M}(DAG_0)$ asumiendo que $p \in \mathbf{M}(DAG)$ (LAURITZEN, 2002).

3 MODELOS GRÁFICOS ENCADENADOS

En base a lo anterior, podemos definir un grafo encadenado (CG), como una combinación ordenada de grafos en bloques. Esto es, grafos que se forman de bordes dirigidos y no dirigidos, también llamados grafos estructurados en bloques recursivos. La ordenación de las variables en bloques en un CG, permite considerar a las variables del primer bloque como variables estrictamente explicativas, de las variables del último bloque, y las variables en bloques intermedios son respuesta de las que le preceden y explicativas de las posteriores. Algunos pares de variables unidos por bordes no dirigidos indican la asociación entre ellas y si el borde es dirigido se indica la dependencia de una variable de la otra. Todos los bordes en el mismo bloque son no dirigidos, y todos los bordes entre diferentes bloques son dirigidos.

En un grafo encadenado, un camino es una secuencia de vértices ($\boldsymbol{v}_1,\boldsymbol{v}_2\ldots..,\boldsymbol{v}_k$), tal que $\boldsymbol{v}_i \rightarrow \boldsymbol{v}_{i+1}$ ó ($\boldsymbol{v}_i$, $\boldsymbol{v}_{i+1}$) para cada i = 1,….,k-1. En otras palabras, un camino puede tener bordes dirigidos y no dirigidos, pero siempre en la dirección que marca el borde dirigido. A partir de la estructura de un grafo encadenado pueden ser obtenidas las *componentes conectadas* de un grafo encadenado después de eliminar todos los bordes dirigidos. Escribiremos $C_1,\ldots..,C_r$, para identificar *las componentes conectadas* de un grafo encadenado, una componente conectada de un grafo, es un subconjunto de vértices $\mathbf{A} \subseteq \mathbf{V}$, tal que cualquier vértice en el subgrafo $CG_{\mathbf{A}}$ esta conectado con cualquier otro vértice a través de un camino. Ellas, contienen solo bordes no dirigidos, y si dos componentes están conectadas, se conectan por bordes dirigidos. Por otra parte, todos los bordes dirigidos entre cualesquiera dos componentes deben tener la misma dirección. Así podemos construir un grafo dirigido cuyo "nuevo conjunto de vértices" son las componentes del CG**.** Obvio es que este nuevo grafo es un DAG, el cual se denomina DAG *componente*. Si consideramos el hecho de que las componentes obedecen una α ordenación, esto permitirá plantear varias ordenaciones distintas.

Haciendo una comparación entre los bloques y las componentes, estas últimas aumentan las posibilidades de análisis, puesto que es posible hacer una buena partición de las variables que componen los bloques. Esto es, cada bloque es la unión de dos o más componentes. Dos variables de un bloque estarán en la misma componente siempre que ellas estén conectadas, es decir, hay un camino entre ellas en el grafo. En FRYDENBERG (1990) se demuestra que todas las representaciones de bloques son probabilísticamente equivalentes y define cuando dos CGs son equivalentes.

.

Al igual que en los casos anteriores, la estructura de dependencias en los CGs se establece en base a las propiedades de Markov. Las propiedades de Markov en los CG han sido estudiadas por FRYDENBERG (1990), LAURITZEN (1996), PEARL (1986); PEARL (1990); y STUDENY (1997), entre otros autores.

La falta de un borde no dirigido entre dos vértices ***i, j*** en el mismo bloque $\mathbf{BL_t}$, o la falta de un borde dirigido de $i \in \mathbf{BL_t}$ a $j \in \mathbf{BL_{t+1}}$, verifica que $i \coprod j$ / $\mathbf{BL}_1 \cup \mathbf{BL}_2 \cup \ldots\ldots \cup \mathbf{BL}_{t+1}$ Esta es una versión de la propiedad de pares de Markov para CGs, pero la interpretación de un borde faltante entre un par de variables (***i, j***), es que ellas son condicionalmente independientes de todas las variables "antes de", y "coexistentes con" respecto del par que se trate. Por tanto, "antes de" y "coexistentes con", se refiere a todas las variables en los bloques previos y variables que pertenecen al mismo bloque respecto de ***i, j***. Al igual que en los grafos anteriores, en los CGs las independencias se verifican en base a un criterio de *c*-separación en el grafo moral $\mathbf{G}^m$ correspondiente al CG. El criterio de *c*-separación de Markov para grafos encadenados, menciona que para tres conjuntos disjuntos de vértices **A**, **B**, y **S** de un CG, $\mathbf{A} \coprod \mathbf{B} / \mathbf{S}$ siempre que **S** separa a **A** de **B** en $\mathbf{G}_c^m$, donde **C** = an^+(**A**□ **B**□ **S**), **C** es el conjunto ancestral obtenido de $\mathbf{G}^m$. Una propiedad interesante es que **C** tiene la característica de ser el conjunto condicionante mínimo, entre el par de variables que se encuentran bajo estudio. Asumiendo que CG_c es un subgrafo del grafo encadenado CG obtenido por eliminar uno o mas bordes de CG. Entonces un modelo **M**(CG_c) es un submodelo del modelo gráfico encadenado **M**(CG) y por tanto, consideramos el estadístico de razón de verosimilitud para verificar la hipótesis de que distribución extendida de probabilidad de Markov *p*, factoriza de acuerdo a un grafo encadenado, si $p \in$ **M**(CG_c), asumiendo que $p \in$ **M**(CG) (LAURITZEN 1996). El proceso de estimación se inicia, mediante métodos iterativos, con la eliminación de un borde de CG para obtener CG_c. Este puede ser o bien un borde dirigido (***i*→ *j***) o un borde no dirigido ((***i,j***),(***j,i***).

De todo lo anterior, podemos decir que la estructura de dependencias subyacente en una tabla de contingencia *p*-dimensional puede ser conocida en base a modelos gráficos.

La selección del modelo en métodos paso a paso es apropiada para identificar un modelo consistente con los datos. Sin embargo se pueden pasar por alto otros modelos que estiman los datos aún mejor. EDWARDS y HAVRANEK (1987), proponen un algoritmo de búsqueda de modelos parsimoniosos (algoritmo EH), consistentes con los datos. Durante la búsqueda, una secuencia de modelos son estimados; estos se clasifican como aceptados (se consideran consistentes con los datos) o rechazados (inconsistentes). La decisión esta basada en la prueba χ^2 de la *deviance*; esto es, si el modelo tiene un $p \leq \alpha$ entonces es rechazado, por tanto todos sus submodelos son también rechazados. En otro caso, el modelo es aceptado. Si un modelo es aceptado, entonces todos los modelos que lo contienen son aceptados. En cualquier etapa del proceso de búsqueda, los modelos rechazados y aceptados se dividen en tres subconjuntos, un conjunto lo forman aquellos modelos que contienen tanto un modelo aceptado como uno o varios submodelos de éste, y que también este puede ser considerado consistente con los datos. Estos modelos son denominados d-aceptados (débilmente aceptados). Por otro lado, otro conjunto se forma de aquellos modelos que contienen modelos rechazados, y submodelos de éstos que se consideran no consistentes con los datos, se denominan modelos d-rechazados. Entre estos dos conjuntos se forma un tercero, aquellos modelos cuya consistencia con los datos no puede ser determinada. La Figura 1, describe la idea comentada.

La Figura anterior muestra 8 modelos (representados por círculos negros y blancos), que se encuentran en un proceso de búsqueda del algoritmo EH, donde dos han sido d-aceptados, dos d-rechazados y cuatro son indeterminados. La última etapa del proceso consiste en probar modelos *minimales* o *maximales* en el conjunto de los indeterminados,

para con esto minimizar la lista de modelos indeterminados. La Figura 2 presenta un ejemplo de un proceso final de búsqueda por el algoritmo EH.

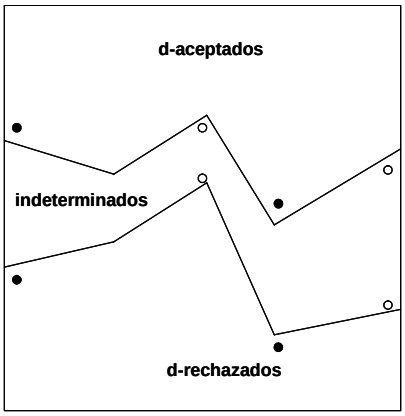

Figura 1: Esquema de las tres clases de modelos según el algoritmo EH

4 SOFTWARE PARA MODELIZACIÓN GRÁFICA

En general, es poco el software estadístico desarrollado en el área de los modelos gráficos. En lo que se refiere a la modelización con grafos encadenados, el software estadístico de propósito general como SPSS, SAS, S plus, etc., no considera módulos específicos para modelizar con grafos. Probablemente, el programa más importante para trabajar modelos gráficos sea MIM (EDWARDS, 1987 y 2000). MIM es un programa frecuentemente mencionado en las publicaciones sobre el tema

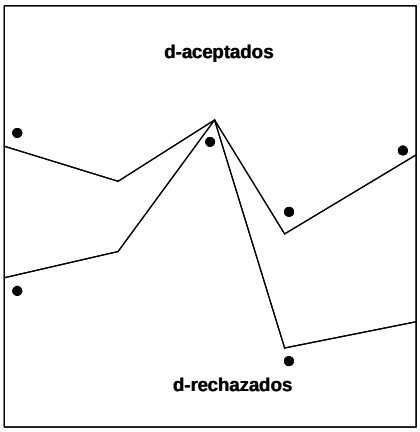

Figura 2: Esquema final de búsqueda del algoritmo EH

MIM trabaja bajo Windows, la manipulación de datos es convencional, detecta propiedades de los modelos, tales como *descomponibilidad* y *colapsabilidad* gráfica / No gáfica (ASMUSSEN y EDWARDS, 1983). El algoritmo general de estimación, máximo verosímil MIPS (*modified iterative proprtional fitting*), presentado en FRYDENBERG y EDWARDS (1989). Incorpora pruebas estadísticas como: razón de verosimilitud, pruebas F y diversas pruebas condicionales exactas, las cuales son evaluadas por enumeración exhaustiva o muestreo Montecarlo.

El procedimiento de la selección del modelo es implementado en una selección paso a paso y en el procedimiento algorítmico EH (EDWARDS y HAVRANEK, 1985 y 1987). La

selección paso a paso cubre los conocidos procedimientos backward y forward, en donde un borde es eliminado o agregado, basado en los p-valores obtenidos de las pruebas de razón de verosimilitud χ^2. Del algoritmo EH ya hemos comentado acerca de sus características para trabajar. El algoritmo EM en DEMPSTER, LAIRD y RUBIN (1977), está implementado en MIM, para el caso de datos faltantes.

MIM, se encuentra disponible en la plataforma de computación estadística y gráfica **R** (HOJSGAARD, 2004). El programa *mim***R**, presenta algunas facilidades para modelización gráfica y es parte del proyecto *g***R** –*project* (http://www.R-project.org/gR). En general en *mim***R** se encuentran muchas similitudes con respecto de MIM, sin embargo el núcleo de *mim***R** esta en *gm***R** y en *mimObjects*, es decir, un área de trabajo de datos y otra para modelar objetos en **R**. El nombre de *gm***R** se refiere a "*graphical meta Data*" y en esta área de trabajo, además de especificar variables y categorías, es posible investigar algunas propiedades de un modelo sin referencia específica a algún grupo de datos, por ejemplo descomponibilidad y colapsabilidad. Un *mim object* vincula la formula de un modelo a un *gmData*. Actualmente, solo es posible trabajar en mim**R** con modelos simétricos, es decir, modelos sin variable respuesta.

Por otra parte, un programa muy utilizado, para modelos gráficos log-lineales es CoCo (COmplete COntyngency Tables), de BADSBERG (1995). El programa está diseñado para analizar tablas de contingencia a través de modelos gráficos log-lineales, además está integrado dentro de XLISP-STAT (TIERNY 1990), esto hace posible utilizar comandos de este ambiente. La estimación y pruebas de hipótesis son realizadas bajo los mismos algoritmos que se mencionan en MIM. Los valores faltantes y ceros estructurales pueden ser tratados con el Algoritmo EM.

5 APLICACIÓN: MODELIZANDO UNA TABLA DE CONTINGENCIA *P*-DIMENSIONAL CON GRAFOS ENCADENADOS

La necesidad de incorporar a las universidades, una serie de dinámicas que los procesos de globalización plantean a los países en desarrollo, requiere de considerar aspectos de muy diversa índole, tales como actualizar planes y programas de estudio, consolidar una plantilla de profesores mejor formada, tener una infraestructura adecuada y actual (bibliotecas, plataforma de computo, etc.), desarrollo en investigación, entre otros muchos aspectos.

En la provincia de Veracruz, en México, se ubica la Universidad Veracruzana, esta universidad es una universidad pública, y por tanto debe responder a las necesidades que en los esquemas de desarrollo actuales se tienen, tanto a nivel nacional, como a nivel internacional. En este contexto, la Universidad Veracruzana, cuenta con 5 *campus* a lo largo de un estado de aproximadamente 6 millones de habitantes.

En este contexto, en noviembre de 2000, se realiza una encuesta de opinión a 1313 entrevistados, todos ellos miembros del colectivo universitario (estudiantes, profesores, investigadores, empleados y directivos). El objetivo del estudio es conocer la opinión de este colectivo universitario, acerca de diferentes proyectos académicos llevados a cabo por los funcionarios de la universidad. La población bajo estudio es un núcleo de más de 50 mil personas distribuidas en los cinco *campus*, en los cuales se presentan entornos con diversos problemas de desarrollo económico y cultural, y necesidades de carácter académico, por ello, se tiene la expectativa de que el estudio provea información relevante respecto de lo que opinan los colectivos bajo estudio.

De cada entrevistado se tiene un conjunto de variables demográficas, tales como el grupo de edad al que pertenece (variable ***d***), actividad que desarrolla para la universidad, (variable ***b***) y *campus* a donde desarrolla su actividad (variable ***a***). Las variables, son

incluidas en el primer bloque del modelo. Recordemos que la ordenación de las variables es un aspecto clave en este tipo de análisis estadístico, sin embargo las variables demográficas o indicadoras de estatus socio-económico se ubican siempre en el primer bloque del modelo, de tal manera que el bloque final, contendrá una (o varias) variable(s) respuesta(s), que sean de interés en el análisis (siempre de acuerdo a los objetivos del estudio). Y un conjunto de variables intermedias, que son respuesta al (los) primer(os) bloques y potenciales explicativas en los bloques subsecuentes (STANGHELLINI, 2003). Las variables intermedias incluidas en el análisis son variables que contienen las respuestas a cuestiones acerca de: ¿Los proyectos realizados han contribuido a mejorar la plantilla de profesores? (variable ***s***), Y ¿Cree que las acciones realizadas han contribuido al mejoramiento de nuestra Universidad? (variable ***x***). Como variable respuesta de interés principal en el análisis, se incluyo la variable (***m***) que contiene la respuesta del entrevistado a la cuestión: ¿Considera que la universidad responde a los retos que su entorno plantea?

Las categorías de cada variable son definidas en la tabla 1.

Si consideramos las variables y sus categorías respectivas, la tabla de contingencia, contiene 160 celdas.

La Figura 1 de la siguiente página, muestra la ordenación de las seis variables, se presentan también algunos resultados porcentuales de alguna categoría de la variable.

Un modelo gráfico encadenado para variables categóricas es estimado como una secuencia de modelos log-lineales. Esta forma de estimación hace a este tipo de modelización particularmente atractiva. El primer modelo log-lineal para las variables en el primer bloque es obtenido, seguido de ello se estima el modelo log-lineal para la distribución de probabilidad condicional de las variables del segundo bloque, dadas las variables del primer bloque, análogamente se realiza la estimación del modelo log-lineal para la distribución de probabilidad condicional de las variables del tercer bloque, dados los dos primeros bloques, y así de manera sucesiva. Si en cada etapa las estimaciones del modelo log-lineal son máximo verosímiles (ML), entonces el modelo resultante es ML.

Niveles de ***a***	Niveles de ***b***	Niveles de ***d***	Niveles de ***s***	Niveles de ***x***	Niveles de ***m***
a_1 = Campus 1 a_2= Campus 2 a_3 = Campus 3 a_4 = Campus 4 a_5 = Campus 5	b_1 = Estudiante b_2 = Prof ./ Investig. Empleado, o Funcionario	d_1 = 18-24 d_2 = ≥ 25	s_1 = Demasiado ó lo suficiente s_2 = Poco ó nada	x_1 = Demasiado ó lo suficiente x_2 = Poco ó nada	m_1 = Sí m_2 = No

Tabla 1: Variables consideradas para el modelo, y sus correspondientes categorías

Después de verificar las pruebas de independencia condicionada y usando una estrategia de selección paso a paso en MIM (EDWARDS, 2000), el **M**(CG) seleccionado se muestra en la Figura 2, de la página siguiente, de donde es posible identificar la factorización de la distribución de probabilidad conjunta (*p*):

$$p = p(\boldsymbol{m}/\boldsymbol{x},\boldsymbol{s},\boldsymbol{a},\boldsymbol{b})\; p(\boldsymbol{x}/\boldsymbol{s},\boldsymbol{a},\boldsymbol{b})\; p(\boldsymbol{s}/\boldsymbol{b})\, p(\boldsymbol{b},\boldsymbol{d})\, p(\boldsymbol{a},\boldsymbol{b})$$

Las Tablas 2, 3 y 4, de la página siguiente, muestran algunos patrones de categorías de las variables explicativas, los porcentajes más altos y más bajos observados para la categoría 1 de la variable respuesta.

Una ventaja de modelizar con grafos, es la posibilidad de trazar en el grafo un camino que permite un análisis descriptivo de las variables consideradas. Por ejemplo, el camino ***d*** → ***s***, ***s***→ ***x***, ***x*** → ***m***, describe las personas jóvenes que opinan que la plantilla del profesorado se ha actualizado, sin embargo no se ha hecho mucho por actualizar los contenidos de programas de estudio, sin embargo se considera que la Universidad responde a los retos que el entorno le plantea.

En la Tabla 2 podemos observar que un 23% de entrevistados menores de 25, opina que el profesorado se ha actualizado.

De la tabla 3 podemos decir que de los entrevistados en el *campus* 1, siendo éstos cualquier tipo de universitario y que considera que el profesorado se ha actualizado, un 23 % de ellos, opina que también los contenidos de los programas se han actualizado. La opinión mas desfavorable respecto de estos cuestionamientos se ubica en los *campus* 5 y 4.

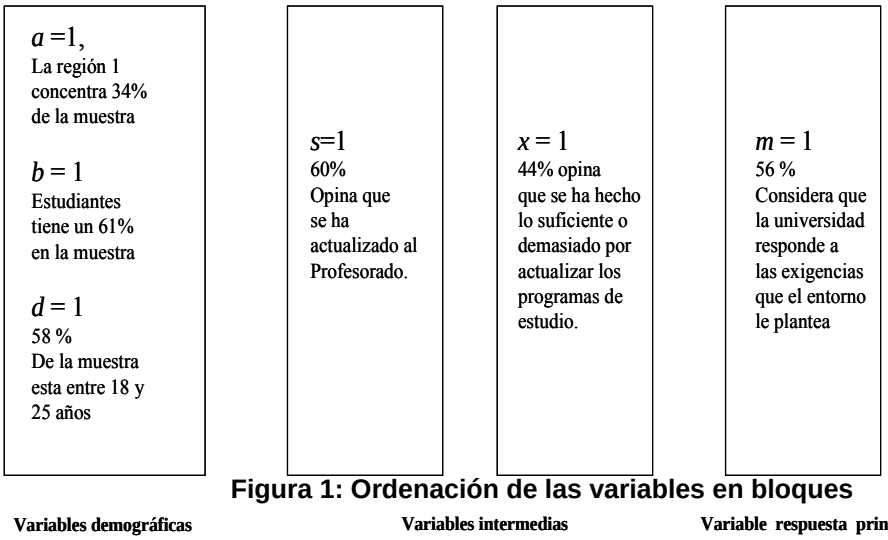

Figura 1: Ordenación de las variables en bloques

Deviance: 0.0000 *p*: 1.0000

Bloque	1	2	3	4
	[***abd***]	[***abd***][***ds***]	[***absx***][***abds***]	[***absxm***][***abdsx***]

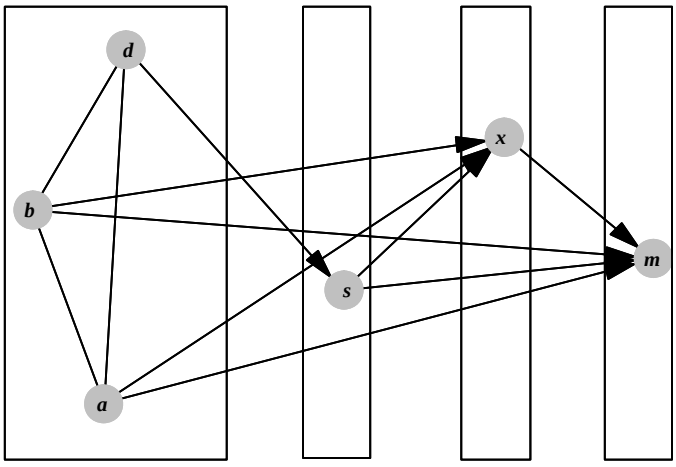

Figura 2: Modelo en bloques encadenados obtenido

Variable explicativa	*Variable respuesta*	Porcentaje
d	***s***	23.0 %
1	1	

Tabla 2: Porcentaje para categoría 1 variable *d* y categoría 1 variable *s*

En la Tabla 4 podemos observar que de los entrevistados en los *campus* 1 y 2 cerca de un veintiún por ciento de ellos opina que la Universidad responde.

Una ventaja de modelizar con grafos, es la posibilidad de trazar en el grafo un camino que permite un análisis descriptivo de las variables consideradas. Por ejemplo, el camino ***d*** → ***s***, ***s***→ ***x***, ***x*** → ***m***, describe las personas jóvenes que opinan que la plantilla del profesorado se ha actualizado, sin embargo no se ha hecho mucho por actualizar los contenidos de programas de estudio, sin embargo se considera que la Universidad responde a los retos que el entorno le plantea.

Variables explicativas			*Variable respuesta*	Porcentaje
a	***b***	***s***	***x***	
1	1	1	1	Alto:23.0 %
1	2	1	1	
5	2	2	1	Bajo: 1.6 %
4	1	2	1	

Tabla 3: Porcentajes obtenidos para tres patrones de categorías de variables explicativas y la categoría 1 de la variable respuesta

Variables explicativas				*Variable respuesta*	Porcentaje
A	***b***	***s***	***x***	***m***	
1	2	1	1	1	Alto:21.0 %
2	1	1	2	1	
2	2	2	2	1	Bajo: 1.5 %
3	2	2	1	1	
4	1	2	1	1	

Tabla 4: Porcentajes para cuatro patrones de categorías de variables explicativas y la categoría 1 de la variable respuesta

Aplicando el algoritmo EH, a los datos contenidos en la tabla de contingencia de seis dimensiones referente al estudio de opinión, y después de un proceso iterativo en siete etapas, se obtienen 6 modelos d-aceptados (ver Tabla 5 y Figura 5)

.

De los 6 modelos considerados como modelos consistentes con los datos, podemos decir que todos tienen en común una *p* baja y ello hace que debamos ser cuidadosos en su evaluación, además todos lo modelos registran una fuerte relación entre las variables explicativas del primer bloque, los dos primeros al realizarles una búsqueda paso a paso la interacción entre ***ads*** y ***bds*** se rechaza, por lo tanto estos modelos los descartamos como opción de estimación para los datos. De los cuatro modelos restantes el modelo cinco es el que tiene un mayor *p* valor, al realizar la búsqueda paso a paso ($p = 0.521$), es un modelo mas simple que el inicial y por ello podemos considerarlo como una mejor opción para describir la estructura de asociación subyacente en la tabla de contingencia.

Respecto de las variables consideradas para el análisis podemos decir que la variable ***s*** no parece una buena variable explicativa para la variable respuesta principal, ya que la estructura de asociación entre las otras variables explicativas y la respuesta es fuerte y bien se puede prescindir de esta variable.

Modelo 1		*Deviance*: **118.01**	**g.l. 96**	*p*: **0.063**
Bloque	**1**	**2**	**3**	**4**
	[***abd***]	[***abd***][***bds***]	[***abds***][***absx***]	[***abdsx***][***absxm***]
Modelo 2		*Deviance*: **110.91**	**g.l. 90**	*p*: **0.067**
Bloque	**1**	**2**	**3**	**4**
	[***abd***]	[***abd***][***ads***]	[***abds***][***absx***]	[***abdsx***][***asxm***]
Modelo 3		*Deviance*: **116.48**	**g.l. 96**	*p*: **0.076**
Bloque	**1**	**2**	**3**	**4**
	[***abd***]	[***abd***][***ds***]	[***abds***][***sx***]	[***abdsx***][***adsxm***]
Modelo 4		*Deviance*: **118.98**	**g.l. 96**	*p*: **0.056**
Bloque	**1**	**2**	**3**	**4**
	[***abd***]	[***abd***][***bs***]	[***abds***][***sx***]	[***abdsx***][***adsxm***]
Modelo 5		*Deviance*: **114.19**	**g.l. 96**	*p*: **0.099**
Bloque	**1**	**2**	**3**	**4**
	[***abd***]	[***abd***][***ds***]	[***abds***][***sx***]	[***abdsx***][***absxm***]
Modelo 6		*Deviance*: **116.69**	**g.l. 96**	*p*: **0.074**
Bloque	**1**	**2**	**3**	**4**
	[***abd***]	[***abd***][***bs***]	[***abds***][***sx***]	[***abdsx***][***absxm***]

Tabla 5: Modelos en bloques encadenados aceptados

5 CONCLUSIÓN

En este artículo, hemos mostrado los modelos gráficos encadenados, como una potente herramienta para el análisis de tablas de contingencia multivariantes. Consideramos que en el contexto de un análisis de datos cualitativos, la modelización gráfica posee un gran potencial, en el sentido que el análisis de la estructura de las interrelaciones entre las variables, permite el uso de variables respuesta intermedias a una variable de interés principal. Esto es crucial para un mejor conocimiento de una problemática bajo estudio.

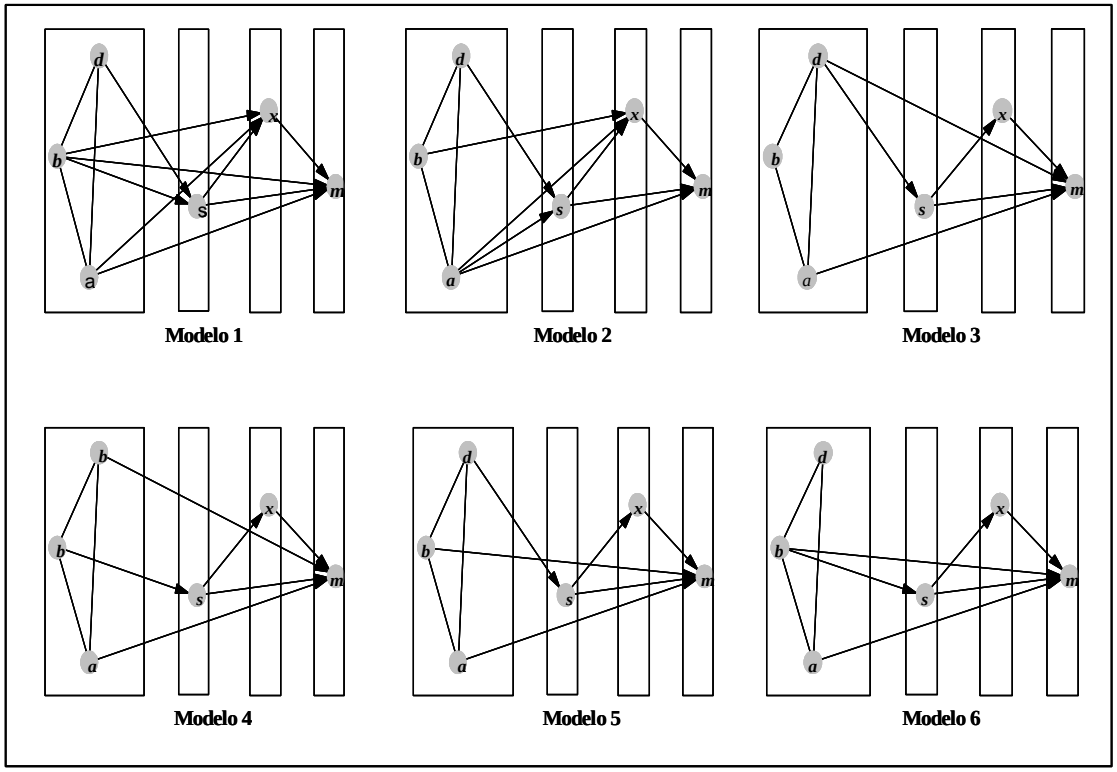

Figura 5: Modelos Gráficos encadenados aceptados

Además hemos realizado una aplicación práctica y se comentan las características del software mayormente utilizado en la modelización gráfica.

REFERENCIAS:

AGRESTI, A. (1990): Categorical Data Análisis. **Wiley**, New York.

ASMUSSEN, S & EDWARDS, (1983): **Collapsibility and Response Variables in Contingency tables**. Biometrika. 70, 567-578.

BADSBERG, J. H. (1995): An Enviroment for Graphical Models. **PhD Thesis, Aalborg University.**

CHRISTENSEN, R. (1990): Log-linear Models and Logistic Regression. **Springer Verlag**, Berlin.

DARROCH, J.N.; LAURITZEN, S.L. & SPEED, T.P. (1980): **Markov Fields and Log-linear Interaction Models for Contingency Tables**. Annals of Statistics. 8, 522-539.

DEMPSTER, A.P., LAIRD, N.M. and RUBIN, D.B. (1977): Maximun Likelihood from Incomplete Data Via the EM Algorithm**. Journal of the Royal Statistical Society. B 39**, 1-38

EDWARDS, D. and HAVRANEK, T. (1985): A Fast Procedure for Model Search in Multidimensional Contingency Tables. **Biometrika. 72**, 339-351.

EDWARDS, D. and HAVRÁNEK (1987): A fast model selection procedure for large families of models. **Journal American Statistical Association. 82**, 205-213.

EDWARDS, D. (2000): **Introduction to Graphical Modelling**. Springer Verlag. New York.

FRYDENBERG, M (1990): Marginalization and collapsibility in graphical interaction models. **Annals of Statistics. 18**, 790-805.

FRYDENBERG, M. and EDWARDS, D. (1989): A modified Iterative Scaling Algorithm for Estimation in regular Exponential Families. **Computational Statistics and Data Analysis. 8,** 142-153.

HOJSGAARD, S. (2004): The mim**R** Package for Graphical Modelling in **R**. **Journal of Statistical Software. 11**, issue 6.

LAURITZEN, S.L. and WERMUTH, N. (1989): Graphical Models for Associations between Variables, Some of Which are Qualitative and Quantitative. **Annals of Statistics. 17,** 31-57.

LAURITZEN, S.L. (1996): Graphical Models. **Clarendon Press**. Oxford.

LAURITZEN (1999): Causal Inference from Graphical Models. **Research report R-99-2021. Aalborg University.**

LAURITZEN,S.L. (2002): Lectures On Contingency Tables. **Electronic Edition**.

PEARL,J.(1986): Fusion, Propagation and Structuring in Belief Networks. **Artificial Inteligence, 29**: 241-88.

PEARL,J. 1990: Aspects Of Graphical Models Connected With Causality. **Computer Science Department. University of California,** Los Angeles California.

STANGHELLINI E. (2003): Monitoring the Behaviour of Credit Card Holders with Graphical Chain Models. **Journal of Business Finance Accounting.**

STUDENY M. (1997): On separation Criterion and recovery Algorithm for Chain Graphs. **Institute of Information Theory and Automation** Academy of Sciences of Czech Republic, Prague.

SONDEO DE OPINION UNIVERSIDAD VERACRUZANA (2000): Facultad de Estadística e Informática, Laboratorio de Investigación y Asesoría Estadística Universidad Veracruzana. **Reporte de Investigación.**

TIERNY, L. (1990): LISP-STAT, **Wiley and Sons Inc**., New York.

VERMA, T. and PEARL, J. (1990): Causal Networks: Semantics and Expressiveness. Uncertainy in **Artificial Intelligence IV.** R. D. Scachter. 66-76.

WERMUTH, N (2003): Analyzing Social Science Data with Graphical Markov Models. **University of Mainz.**

WERMUTH, N and COX D.R. (1998): On the application of conditional independence to ordinal data. **International Statistical Review. . 66** , 181-199. .

WHITTAKER, J. (1990): Graphical Models in Applied Multivariate Statistics **Wiley**, New York.

Received June 2005
Revised July 2006