"El GCGM (modelo gaussiano condicionado gráfico) como herramienta de análisis de las relaciones entre variables económicas y sociales en los municipios extremeños"

Miguel Ángel Fajardo Caldera (fajardo@unex.es)
Jesús Pérez Mayo (jperez@unex.es)
Lidia Andrades Caldito (andrade@unex.es)
Departamento Economía Aplicada. Universidad de Extremadura. Avda de Elvas, s/n 06071 Badajoz

RESUMEN

En este estudio, se utilizará el modelo gaussiano condicionado gráfico descrito por Lauritzen y Wermuth (1989) para analizar la relación existente entre una serie de variables económicas y sociales en los municipios extremeños. Entre otras variables, se consideran el número de empresas existente, la población de derecho, el número de empleados, la facturación de las empresas así como un índice de la presencia de un conjunto de servicios públicos considerados básicos construido por Pérez Mayo y Fajardo (1996)

1. Introducción

Entre otros muchos, un importante aspecto del análisis regional es la búsqueda de las estructuras de relación o asociación que determinan o describen la situación de un territorio.

Este trabajo pretende presentar una herramienta que consideramos puede ser útil para lograr el objetivo antes expuesto. Dicha herramienta es la modelización gráfica.

Saber qué variables están asociadas permite descubrir factores que influyen en un problema, implicaciones de algunas decisiones en el escenario contemplado, indicios o pistas para lograr algún objetivo. Estas cuestiones interesan no sólo en la economía, sino también en la geografía, sociología, política, ecología...

Como veremos más tarde, la modelización gráfica se basa en el uso de la teoría de grafos para simplificar la estructura de dependencia o independencia entre las variables consideradas. Esta teoría se aplica a modelos de variables discretas, continuas o mixtos, es decir, con variables de ambos tipos.

Para ilustrar esta teoría, presentaremos un ejemplo con variables discretas. Los modelos log-lineales son los más adecuados para medir la relación entre variables de este tipo.

Los datos que usaremos provienen de un estudio realizado dentro del subárea de Métodos Cuantitativos Aplicados a la Economía del Departamento de Economía Aplicada y Organización de Empresas de la Universidad de Extremadura.

Wermuth y Lauritzen (1990) introdujeron la familia de distribuciones Gaussianas condicionadas.

La distribución Gaussiana condicionada está definida de la siguiente forma: la distribución marginal de las variables discretas es multinomial y, condicionadas a estas variables, las variables continuas tienen una distribución normal multivariante. Esta distribución contienen como casos especiales las distribuciones puramente discretas y continuas.

Una característica importante de esta construcción es que los estadísticos suficientes para el modelo saturado son las frecuencias de las celdas y, dentro de cada celda, el vector de las medias y la matriz de covarianzas de las variables continuas. Dado un grafo, los modelos gráficos GC se definen como aquellas familias de distribuciones que satisfacen los

argumentos de independencia condicionada implícitos en el grafo de independencia asociado.

Un conjunto de modelos, los modelos de interacción jerárquica, Edwards (1990), se generan a partir de la distribución GC anulando algunos términos en la expansión de interacción jerárquica de los parámetros. El resultado es una generalización de las expansiones log-lineales para las variables discretas.

2. Modelo teórico

Es un tipo de técnicas estadísticas basadas en el ajuste de modelos gráficos a los datos. Sus grandes ventajas están a la hora de:

Interpretar el modelo: el objetivo principal de la modelización gráfica es describir las relaciones entre distintas variables así como explicarlas condicionándolas a otra(s) variable(s) o controladas por ésta(s) y,

Simplificar el escenario: se intentan condensar los datos sin eliminar cualquier asociación interesante.

Un *modelo gráfico* es una familia de funciones de densidad de probabilidad que incorporan un conjunto específico de restricciones de independencia condicionada presentadas en un grafo de independencia.

Este método simplifica la representación de la estructura de relación entre variables porque a través de la teoría de grafos y de la noción de *independencia condicionada* busca la forma más sencilla de exponer qué variables están asociadas. Como es natural, cualquier método que simplifique la representación de la realidad que se va a estudiar es interesante, ya que facilita el trabajo.

La herramienta aquí presentada puede utilizarse con variables aleatorias discretas, continuas o una mezcla de ambos tipos o modelos mixtos.

2.1 Conceptos básicos

- Independencia condicionada

Dos variables aleatorias X e Y son independientes condicionadas a otra variable aleatoria Z, si la función de densidad conjunta $f_{XYZ}\ (x, y, z)$ puede descomponerse en dos factores,

f_{XYZ} *(x, y, z) = g (x, z) h(y, z),* donde *g* y *h* son dos funciones.

•Teoría de grafos

Un *grafo* G es una estructura formada por un conjunto finito de *vértices* (K) y un conjunto finito de *arcos* (E) entre dichos vértices.

Los arcos pueden ser *líneas* cuando no existe una ordenación o *flechas*, en caso de que una variable sea causa de la otra.

Si no hay ninguna flecha en el grafo, se dice que es *no dirigido*. Si existen flechas y líneas, se llama *dirigido* y, finalmente, si sólo aparecen flechas, el grafo está *orientado*.

Dos vértices X, Y $\in$ K son *adyacentes* si existe un arco entre ellos, [XY] $\in$ K. Un grafo se llama *completo* si existe un arco entre cada par de vértices.

Cualquier subconjunto u $\subseteq$ K induce un *subgrafo* de G o grafo G_u=(u, F) cuyo conjunto de arcos F está formado por aquellos arcos del conjunto E con ambos extremos en el conjunto u.

FIGURA 1

En la figura 1, tenemos como ejemplo de subgrafo, el formado por los vértices WXY, que, al existir arcos entre ellos, es, además, completo.

Otro concepto importante es el de *clique*. Un subconjunto u $\subseteq$ K es calificado así si es maximalmente completo, es decir, u es completo y si w $\subseteq$ u, w no lo es.

Una secuencia de vértices distintos X_0,..., X_n tal que X_{i-1} sea adyacente a X_i para i=1,..., n es un *camino* entre los vértices X_0 y X_n de longitud n. Dos vértices, X e Y, están conectados si existe un camino desde X hasta Y y viceversa y si todos los pares de vértices de un grafo están conectados, se dice que el grafo está conectado.

Además, decimos que existe un *ciclo* si en un camino si los puntos de comienzo y final son el mismo vértice, es decir, X_0=X_n.

Si los n vértices de un ciclo son distintos y X_j es adyacente a X_k sólo si |j-k|=1 o n-1, dicho ciclo se llama ciclo *chordless*. Relacionada con este concepto, se halla la idea de grafo triangular. Esto sucede si no existen ciclos *chordless* de longitud mayor igual a 4. En la figura 1, el grafo de la derecha sería triangular.

Dado tres subconjuntos a, b y s del conjunto V, s separa a y b si cada camino que una a con b contiene al menos un vértice del subconjunto separador s.

Finalmente, la *frontera* de un subconjunto $u \subseteq V$ se define como aquellos vértices de V \ u adyacentes a un vértice de u.

Si representamos por vértices las variables del modelo, la interacción entre dos ellas se representa mediante un arco que uniese dichos puntos como vemos en la figura 1.

Los modelos de interés son aquellos modelos cuyo grafo de asociación es un grafo de independencia condicionada.

Dado un vector X de variables aleatorias, $X=(X_1,...,X_n)$ con su correspondiente conjunto de vértices K=(1,...,k), el grafo de independencia condicionada de X es el grafo no dirigido G=(K, E), donde el arco que une los vértices i y j no está en el conjunto de arcos E si y sólo si X_i es independiente de X_j condicionados a $X_{K\setminus\{i,j\}}$, es decir, al resto de las variables.

La interpretación de los grafos se apoya en las propiedades de Markov:

- por parejas: las parejas de variables no adyacentes son independientes condicionadas a las variables restantes.

- local: cualquier variable es independiente de las demás condicionada a sus adyacentes

- global: dos subconjuntos de variables separados por un tercero son independientes condicionados a las variables del tercer subconjunto.

Así, en el grafo podemos ver que si dos variables no son adyacentes, entonces son independientes condicionadas al resto.

Sin embargo, la clave para interpretar los grafos es la propiedad global de Markov. Permite convertir una propiedad de la teoría de grafos, la separación, en una propiedad estadística, la independencia condicionada.

Podemos crear el grafo a partir de las fórmulas de los modelos usados o, esto es lo más importante, conocer las interacciones existentes en el modelo a partir de su grafo. Es

decir, una vez construido el grafo, es mucho más fácil reconocer qué estructura de relación existe.

2.2 Distribución GC

El vector completo de las variables aleatorias es *X=(I,Y)*, donde es el vector p-dimensional de las variables discretas e Y es el vector q-dimensional de las variables continuas. El conjunto completo de los *k* vértices *K*, por tanto, se divide en $K=\Delta\cup\Gamma$. Un valor concreto tomado por *(I,Y)* se designa por (*i*,y) y un vector marginal de *I,Y*) por (I_a,Y_b) donde $a\subseteq\Delta$ y $b\subseteq\Gamma$.

Aquí *i* contiene los valores de las variables discretas e *y* es un vector real de longitud *q*.

Suponemos que la probabilidad de que *I= i* es p_i y que la distribución de *Y* dado *I= i* es normal multivariante $N(m_i,\Sigma_i)$ de manera que tanto la media como la covarianza condicionadas pueden depender de i. Esta distribución se conoce como distribución GC (Gaussiana condicionada). La densidad puede escribirse como

$$f(i,y)=p_i\,|\,2p\Sigma_i\,|^{-\frac{1}{2}}\exp\left\{-\tfrac{1}{2}(y-m_i)'\Sigma_i^{-1}(y-m_i)\right\}.$$

Los parámetros $\{p_i,m_i,\Sigma_i\}$ se conocen como los momentos.

A menudo se está interesado en los modelos homogéneos, esto es, aquéllos en los que la covarianza es contante para *i*. Sin embargo, como veremos más tarde, existen por lo general dos modelos gráficos correspondientes a un grafo dado: un modelo heterogéneo y un modelo homogéneo.

Podemos reescribir la expresión de la función de densidad de esta forma:

$$f(i,y)=\exp\left\{a_i+b_i'y-\tfrac{1}{2}y'\Omega_i y)\right\},$$

donde a_i es un escalar, b_i es un vector de orden *px1* y Ω_i es una matriz simétrica definido positiva. Estos parámetros se conocen como parámetros canónicos que se relacionan con los momentos mediante las siguientes expresiones:

$$\Omega_i=\Sigma_i^{-1},$$

$$b_i=\Sigma_i^{-1}m_i,$$

$$a_i=\ln(p_i)-\tfrac{1}{2}\ln|\,\Sigma_i\,|-\tfrac{1}{2}m_i'\Sigma m_i-\tfrac{q}{2}\ln(2p)$$

y

$$\Sigma_i = \Omega_i^{-1},$$

$$m_i = \Omega_i^{-1} b_i,$$

$$p_i = (2p)^{\frac{q}{2}} |\Omega_i|^{-\frac{1}{2}} \exp\left\{a_i + \tfrac{1}{2} b_i' \Omega_i^{-1} b_i\right\}$$

Los modelos de interacción jerárquicos se construyen restringiendo los parámetros canónicos de forma parecida a como se hace en los modelos log-lineales. Los parámetros canónicos se expanden como una suma de unos términos de interacción y un modelo se define anulando los términos de interacción de mayor orden.

2.3 Formulación del modelo

Supongamos que la fórmula del modelo tiene la forma

$$\underbrace{d_1,\ldots,d_r}_{discreta} / \underbrace{l_1,\ldots,l_s}_{continua} / \underbrace{q_1,\ldots,q_s}_{cuadrática}$$

Estas tres partes tienen las siguientes funciones:

1. los generadores discretos especifican la expansión de a_i

2. los generadores lineales especifican la expansión de b_i. Cada generador lineal contiene una variable continua. La expansión de b_i^g para cualquier $g \in \Gamma$ viene dada por los generadores lineales que contienen a g .

3. La parte cuadrática da la expansión de la inversa de la matriz de covarianzas. Ω_i Cada generador cuadrático debe contener al menos una variable continua. La expansión de w_i^{gz} para cualquier $g, z \in \Gamma$ viene dada por los generadores cuadráticos que contienen a g, z .

Además, la cantidad de posibles fórmulas se ve restringido ya que deben cumplir dos reglas:

1. El número de generadores lineales no puede ser mayor que el de generadores discretos.

2. El número de generadores cuadráticos no puede ser mayor que el de los correspondientes generadores lineales.

Estas reglas existen para asegurar que los modelos sean invariantes frente a cambios de escala y origen de las variables continuas.

2.4 Fórmulas y grafos

Para estudiar la correspondencia entre la formulación de los modelos y los grafos, se puede expandir la función de densidad como

$$f(i,y) = \exp\left\{a_i + \sum_{g\in\Gamma} b_i^g y_g - \tfrac{1}{2}\sum_{g\in\Gamma}\sum_{h\in\Gamma} w_i^{gh} y_g y_h)\right\},$$

y entonces aplicar el criterio de factorización para examinar las propiedades de Markov implicadas en un grafo dado.

Para dos variables discretas del modelo, sea A y B, A y B son independientes condicionadas al resto siempre que todos los términos de interacción que incluyen a A y B se anulan. Es decir, ninguna de las expansiones de a_i, b_i^g o w_i^{gh}, para cualquier, puede contener una interacción AB. En cuanto a la fórmula del modelo, se requiere que ningún generador discreto contenga AB puesto que las reglas de construcción implican que tampoco pueda conternerlo ningún generador lineal o cuadrático.

Si A es discreta y X continua, A es independiente de X dado el resto, si se anulan los términos de interacción que incluyen a A y X., por tanto, ninguna de las expansiones de b_i^X o w_i^{Xh} para cualquier $h \in \Gamma$ puede contener un término de interacción que incluya A. en cuanto a la fórmula del modelo, se requiere que ningún generador lineal pueda contener AX ya que las reglas de sintaxis implican que ningún generador cuadrático contendrá AX.

Para dos variables continuas, sean X e Y, X e Y son independientes dado el resto cuando $w_i^{XY} = 0$. En cuanto a la fórmula del modelo, esto significa que ningún generador cuadrático puede contener XY.

Estos resultados facilitan la construcción del grafo de independencia a partir de una fórmula. Simplemente se conectan los vértices que aparezcan en el mismo generador.

Para desarrollar la operación opuesta: encontrar la fórmula del modelo gráfico correspondiente a un grafo dado G, se necesita identificar las interacciones maximales consistentes con G.

Los generadores discretos vienen dados por los cliques de G_Δ, i.e., los subgrafos de *G* para las variables discretas.

Para la parte lineal de la fórmula, se necesita encontrar los cliques de $G_{\Delta\cup\{g\}}$ para cada $g \in \Gamma$.

Para la parte cuadrática, depende de si se busca el modelo gráfico homogéneo o heterogéneo. Para el modelo homogéneo, se necesita identificar los cliques de G_Γ. Para el modelo heterogéneo, se requiere hallar los cliques de *G* que cortan Γ.

2.5 Estimación por máxima verosimilitud

Los modelos necesitan datos. Supongamos que tenemos una muestra de N observaciones independientes e idénticamente distribuidas $(i^{(k)}, y^{(k)})$ para *k=1,...N*, donde *i* es un conjunto de niveles de las variables discretas e *y* es un vector de dimensión q. Sea $(n_j, t_j, \bar{y}_j, SS_j, S_j)$ las frecuencias observadas, los totales de las variables, las medias de las variables, las sumas de cuadrados y productos no corregidas y las varianzas de las celdas para cada celda j, es decir:

$$n_i = \#\left\{k : i^{(k)} = i\right\},$$

$$t_i = \sum_{k:i^{(k)}=i} y^{(k)},$$

$$\bar{y}_i = t_i \Big/ n_i,$$

$$SS_i = \sum_{k:i^{(k)}=i} y^{(k)}(y^{(k)})',$$

$$S_i = \sum_{k:i^{(k)}=i} (y^{(k)} - \bar{y}_i)(y^{(k)} - \bar{y}_i)' \Big/ n_i = SS_i \Big/ n_i - \bar{y}_i \bar{y}_i'$$

También es necesario establecer una notación para algunas cantidades marginales. Para $a \subseteq \Delta$, se escribe la celda marginal correspondiente a i como i_a y análogamente para $d \subseteq \Gamma$, se escribe el subvector de y como y^d. De igual forma, puede escribirse las frecuencias marginales como $\{n_{ia}\}$, los totales marginales de las variables como $\left\{t_{ia}^d\right\}$ y las sumas marginales no corregidas de cuadrados y productos como $\left\{SS_{ia}^d\right\}$.

Consideremos un modelo con fórmula $\underbrace{d_1,\ldots,d_r}_{\text{discreta}} / \underbrace{l_1,\ldots,l_s}_{\text{continua}} / \underbrace{q_1,\ldots,q_s}_{\text{cuadrática}}$. Por tanto, a partir de la ecuación, es fácil observar que un conjunto de estadísticos minimales suficientes viene dado por:

1. un conjunto de tablas de frecuencias marginales $\{n_{ia}\}$ correspondientes a los generadores discretos, es decir, para $a = d_1,\ldots,d_r$.

2. un conjunto de totales marginales de las variables $\{t_{ia}^d\}$, correspondientes a los generadores lineales, esto es, para $a = q_j \cap \Delta, g = l_j \cap \Gamma$, para $j = 1,\ldots,s$.

3. un conjunto de sumas marginales no corregidas de cuadrados y productos $\{SS_{ia}^d\}$ correspondientes a los generadores cuadráticos, i. e., para $a = q_j \cap \Delta$, y $b = q_j \cap \Gamma$, para $j = 1,\ldots,t$.

Como ya se ha visto, los modelos se construyen restringiendo los parámetros canónicos mediante las expansiones de los factores de interacción. Dado un conjunto de datos, se desea estimar los parámetros del modelo sujetos a estas restricciones mediante la estimación por máxima verosimilitud. A partir de la teoría de la familia exponencial, sabemos que los EMV (estimadores por máxima verosimilitud) pueden encontrarse igualando los estadísticos minimales suficientes con sus valores esperados. Esto es, para $a = d_1,\ldots,d_r$,

$$\{n_{ia}\} = \{m_{ia}\},$$

para, $a \cup g = l_1,\ldots,l_s$

$$\{t_{ia}^d\} = \left\{ \sum_{j:j_a = i_a} m_j m_j^g \right\}$$

y para, $a \cup d = q_1,\ldots,q_t$

$$\{SS_{ia}^d\} = \left\{ \sum_{j:j_a = i_a} m_j \left[\Sigma_j^{dd} + m_j^d (m_j^d)' \right] \right\}$$

Estas son las ecuaciones de verosimilitud. Cuando existen, los EMV son la única solución de las ecuaciones que, además, satisfacen las restricciones del modelo.

3. Aplicación y estimación del modelo

En este trabajo intentamos medir la relación existente entre la dotación de algunos servicios públicos, la población, el número de empresas, el número de empleados y la facturación para los municipios extremeños.

La asociación entre infraestructura y crecimiento ha sido estudiada sobre todo por la teoría del *potencial de crecimiento regional*, desarrollada por Biehl.

La idea básica de esta teoría es que las disparidades regionales son el resultado del desarrollo a largo plazo.

La infraestructura es uno de los factores regionales de potencialidad que determina las perspectivas de desarrollo regional. Por ello, merece la pena identificar la contribución relativa de la infraestructura al potencial de desarrollo.

En nuestro trabajo aparecen 5 variables: población, número de empresas, número de empleados y facturación y dotación de servicios públicos, datos referidos a 1998 que aparecen en el apéndice.

Un estudio realizado en 1996 mostró una relación entre algunas de estas variables tras haber sido categorizadas mediante un modelo log-lineal, trabajo que ahora se extiende aplicando un modelo gráfico gaussiano condicionado para observar las relaciones entre las variables.

En primer lugar, para evitar la influencia de las escalas de medida de las distintas variables, se estandarizaron las series de forma que en los resultados finales obtendremos las relaciones existentes entre las puntuaciones estandarizadas de cada variable y no entre un número determinado de trabajadores o de empresas. Además, se tomó esta decisión ya que los estadísticos descriptivos que requería el programa MIM (Edwards, 1990) con el que se realizó loa estimación se veían muy afectados por las escalas de medida y provocaban problemas de estimación.

El procedimiento de estimación utilizado es el algoritmo MIPS (*modified iterative proportional system*), descrito por Frydenberg y Edwards. En concreto, consiste en ir contrastando los modelos resultantes tras añadir o eliminar un arco del grafo con el modelo base, ya sea una selección del modelo hacia arriba o hacia abajo.

Los modelos se contrastan mediante el cálculo de la desviación. Dado un modelo la desviación de su razón de verosimilitud viene dada, con respecto con el modelo heterogéneo saturado, por

$$2\sum_i \ln\left(n_i \big/ m_i\right) - \sum_i n_i \ln|S_i\hat{\Sigma}_i^{-1}| + \sum_i n_i\left\{tr(S_i\hat{\Sigma}_i^{-1}) - q\right\} + \sum_i n_i(\bar{y}_i - \hat{m}_i)'\hat{\Sigma}_i^{-1}(\bar{y}_i - \hat{m}_i), \qquad \text{y}$$

para el modelo homogéneo saturado, por

$$2\sum_i \ln\left(n_i \big/ m_i\right) - N\ln|S\hat{\Sigma}^{-1}| + N\left\{tr(S\hat{\Sigma}^{-1}) - q\right\} + \sum_i (\bar{y}_i - \hat{m}_i)'\hat{\Sigma}^{-1}(\bar{y}_i - \hat{m}_i)$$

El primer modelo que estimamos es la estructura de relación entre todas las variables, donde A representa la **dotación de servicios públicos** y tiene 6 niveles (desde 0 a 5), W es la serie de puntuaciones estandarizadas de la **población**, X del **número de empresas**, Y, del **número de empleado** y Z, de la **facturación**.

Para estimar el modelo, podemos partir del modelo saturado e ir eliminando los arcos, es decir, las relaciones no significativas (estimación hacia atrás) o partir del modelo de efectos principales e ir incorporando los arcos significativos (estimación hacia delante). Hemos optado por realizar una estimación hacia atrás ya que parte de un modelo consistente con los datos y los modelos que va contrastando contra éste también son consistentes. Sin embargo, como Edwards (1990) muestra, el proceso inverso de estimación al comenzar con un modelo que no es necesariamente consistente con los datos, contrasta sucesivamente modelos que tampoco lo son.

Además, hemos decidido restringir los posibles modelos que expliquen la relación a modelos separables (*decomposable models*), es decir, a modelos que pueden descomponerse en otros más simples para explicar las relaciones entre grupos de variables. Tienen la característica de que existen formas cerradas de EMV para los submodelos marginales.

En primer lugar, vamos a contrastar si existe homocedasticidad o no, es decir, si las varianzas son homogéneas.

```
MIM->model a/aw,ax,ay,az/wxyz; fit; boxtest
Deviance: 44.7195         DF: 50
Test of H0: A/AW,AX,AY,AZ/WXYZ
against H: A/AW,AX,AY,AZ/AWXYZ
Box's test: 4179.1322     DF: 50  P: 0.0000
```

Por tanto, se rechaza la hipótesis de homogeneidad de varianzas. Podemos llegar a nuestra primera conclusión: la dispersión de los valores de las variables depende del nivel de dotación que contemplemos. Además, ya sabemos que el modelo del que partiremos será el modelo saturado heterogéneo A/AW,AX,AY,AZ/AWXYZ

MIM->StepWise
Coherent Backward Selection
Decomposable models, chi-squared tests.
DFs adjusted for sparsity.
Critical value: 0.0500
Initial model: A/AW,AX,AY,AZ/AWXYZ
Model: A/AW,AX,AY,AZ/AWXYZ
Deviance: 0.0000 DF: 0 P: 1.0000

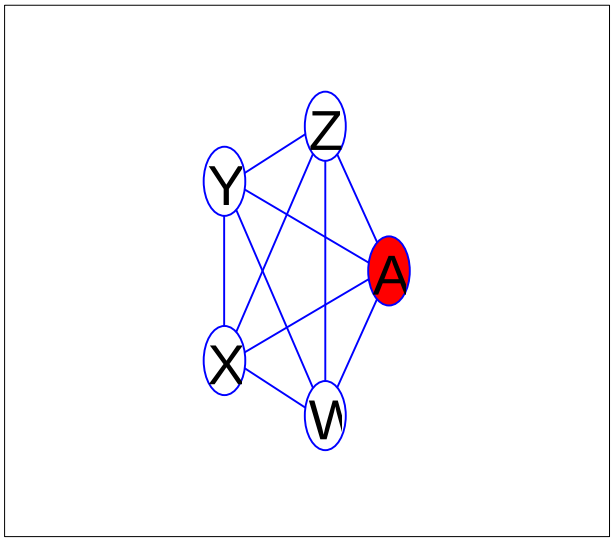
Edge excluded	Test Statistic	DF	P
[AW]	568.8029	25	0.0000 +
[AX]	1314.0914	25	0.0000 +
[AY]	1305.7590	25	0.0000 +
[AZ]	1267.6195	25	0.0000 +
[WX]	168.9971	6	0.0000 +
[WY]	99.1070	6	0.0000 +
[WZ]	80.0278	6	0.0000 +
[XY]	327.1732	6	0.0000 +
[XZ]	224.5272	6	0.0000 +
[YZ]	515.0527	6	0.0000 +

No change.
Selected model: A/AW,AX,AY,AZ/AWXYZ

Como vemos, el modelo que se elige es el modelo saturado, ya que la eliminación de cualquier arco daría lugar a un modelo no significativo. La representación gráfica de dicho modelo aparece en la figura 2. Dicho modelo refleja la existencia de una relación entre la dotación de servicios públicos y el resto de variables, relación que ya comentamos al hablar de la homogeneidad de varianzas, ya que no eran iguales las varianzas para todos los municipios, sino que varían para cada conjunto de municipios agrupados según su nivel de servicios. Además, las medias de cada celda varían por lo que la puntuación media obtenida también varía según el nivel de dotación que seleccionemos.

La estimación de medias y coeficientes de correlación para cada nivel pueden observarse en el apéndice 3. Podemos comprobar cómo las medias de las variables son mayores a medida que la clase de la variable A es mayor, es decir, cuánto mayor sea la dotación en servicios públicos, mayor es el valor esperado de las puntuaciones estandarizadas de población, número de empresas, número de empleados y facturación.

FIGURA 2



En consecuencia, mayores son los valores de cada variable para su media y su dispersión.

De igual forma, los coeficientes de correlación entre las variables varían según la clase. Por lo general, para la clase 3 y superiores, existe una gran relación entre las variables. Esto parece indicar que una alta dotación de servicios públicos se sitúa en municipios que poseen más población, mayor número de empresas, mayor número de empleados y empresas con mayor facturación. No hablamos de una relación causal ya que no hemos realizado una regresión, sino que simplemente hemos efectuado un análisis exploratorio de las relaciones existentes.

Las variables más relacionadas son el número de empresas con el número de empleados y el número de empleados con el volumen de facturación. Están relacionadas directamente con coeficientes muy altos incluso en las dos primeras clases. De hecho son las únicas variables con coeficientes superiores al 0.7. La razón es que presentan valores muy bajos que no tiene nada que ver con la población, ya que son municipios con tamaños similares que presentan valores para el resto de las variables también parecidos.

Podemos concluir asimismo que los municipios que pertenecen a la clase superior de la dotación de servicios públicos, presentan las medias mayores así como coeficientes de correlación entre las variables continuas superiores a1 90%, es decir, que son municipios

situados en los puestos más altos de las correspondientes distribuciones.

Dijimos que era una ventaja el hecho de que el modelo fuera separable. Así es, ya que se podría descomponer la función de densidad conjunta en la función de densidad marginal de un conjunto de variables y la función de densidad condicionada del resto de variables dado el conjunto anterior. En definitiva, no sería más que un conjunto de ecuaciones de regresión. Dichas ecuaciones explicarían la influencia de una variable o un conjunto de variables sobre otra u otras.

4. Conclusiones

Hemos visto cómo el método propuesto, la modelización gráfica, puede ser útil a la hora de construir un modelo que represente la realidad. La aplicación al ejemplo muestra cómo las relaciones entre las variables se representan de una manera más clara. En nuestro caso, el grafo indica la estructura de asociación entre las variables definidas en el ejemplo. La conclusión más llamativa que se podría extraer es que, según los datos de la tabla, todas las variables están relacionadas y que el nivel de dotación de servicios públicos influye en la relación existente entre las demás.

Sin embargo, la utilización posible de este método trasciende el ejemplo expuesto. Donde realmente podemos descubrir su potencial es en escenarios donde intervienen muchas variables, ya que permite agruparlas en función de su independencia condicionada. De ahí que las variables que separan los distintos grupos o bloques de variables cobren una importancia esencial en el análisis. Son las variables clave para entender el comportamiento y evolución del modelo.

Bibliografía

Andersen, E.B. (1984)
The Statistical Analysis of Categorical Data, Springer-Verlag, Nueva York.
Bishop, Y.M., S.E. Fienberg y P.W. Holland (1975)
Discrete Multivariate Analysis, MIT Press, Cambridge, MA.
Christensen, R. (1990)
Log-linear Models, Springer-Verlag, Nueva York.

Edwards, D. (1995)

Introduction to Graphical Modelling, Springer-Verlag, Nueva York.

Pérez Mayo, J. (1996)

Ensayo sobre la distribución de algunos servicios públicos en Extremadura, Dept. Economía Aplicada. Universidad de Extremadura, Badajoz. (inédito)

APÉNDICE 1

Fitted counts, means and correlations.

A						
1	W	1.000				
	X	0.447	1.000			
	Y	-0.000	0.707	1.000		
	Z	0.071	0.474	0.894	1.000	
	Means	-0.276	-0.208	-0.185	-0.184	98.000
		W	X	Y	Z	Count
2	W	1.000				
	X	0.239	1.000			
	Y	0.170	0.735	1.000		
	Z	0.105	0.412	0.812	1.000	
	Means	-0.160	-0.163	-0.155	-0.159	198.000
		W	X	Y	Z	Count
3	W	1.000				
	X	0.871	1.000			
	Y	0.562	0.642	1.000		
	Z	0.802	0.965	0.696	1.000	
	Means	-0.004	-0.045	-0.021	-0.075	37.000
		W	X	Y	Z	Count
4	W	1.000				
	X	0.723	1.000			
	Y	0.746	0.871	1.000		
	Z	0.624	0.963	0.828	1.000	
	Means	0.123	0.035	-0.066	-0.071	13.000
		W	X	Y	Z	Count

5	W	1.000				
	X	0.769	1.000			
	Y	0.693	0.800	1.000		
	Z	0.751	0.820	0.973	1.000	
	Means	0.341	0.207	0.055	0.076	15.000
		W	X	Y	Z	Count

6	W	1.000				
	X	0.991	1.000			
	Y	0.970	0.978	1.000		
	Z	0.935	0.951	0.984	1.000	
	Means	2.485	2.417	2.360	2.483	21.000
		W	X	Y	Z	Count