27 Congreso Nacional de Estadística e Investigación Operativa
Lleida, 8–11 de abril de 2003

Análisis del sector agrario del poniente almeriense mediante redes bayesianas

Antonio J. Céspedes[1], Rafael Rumí[2], Antonio Salmerón[2], Francisco J. Soler[2],

[1]Estación Experimental de Las Palmerillas,
Apartado 250, 04080 Almería, España
E-mail: ajcespedes@cajamar.es
[2]Departamento de Estadística y Matemática Aplicada
Universidad de Almería, 04120 Almería, España
E-mail: {rrumi,Antonio.Salmeron,fsoler}@ual.es

RESUMEN

En este trabajo analizamos los resultados de una encuesta realizada sobre el sector agrario del poniente almeriense. Proponemos distintos modelos basados en redes bayesianas en cada uno de los aspectos estudiados, todos ellos referentes a las características de los invernaderos.

Palabras y frases clave: Redes bayesianas, anaálisis multivariante.
Clasificación AMS: 62H99.

1 Introducción

Las redes bayesianas son una representación compacta de una distribución de probabilidad multivariante. Formalmente, una *red bayesiana* es un grafo dirigido acíclico donde cada nodo representa una variable aleatoria y las dependencias entre las variables quedan codificadas en la propia estructura del grafo según el criterio de d-separación (Pearl 1988). Asociada a cada nodo de la red hay una distribución de probabilidad condicionada a los padres de ese nodo, de manera que la distribución conjunta factoriza como el producto de las distribuciones condicionadas asociadas a los nodos de la red.
Diferentes tipos de inferencias pueden llevarse a cabo sobre estos modelos. La tarea más frecuente es la llamada *propagación de probabilidad*, para la cual existen diversos algoritmos que sacan partido de las independencias codificadas por la red para realizar los cálculos de manera eficiente (Cano, Moral, and Salmerón 2000; Lauritzen and Spiegelhalter 1988; Madsen and Jensen 1999). La propagación de probabilidad

consiste en la obtención de las probabilidades a posteriori de ciertas variables de la red dado que se conoce el valor que toman algunas otras variables observadas.
Otra tarea muy habitual con redes bayesianas consiste en la extración de la explicación o explicaciones más probables a una determinada observación (Gámez 1998; Nilsson 1998).
Desde el punto de vista del análisis de datos, las redes bayesianas son una potente herramienta por varios motivos:

1. No suponen un determinado modelo subyacente.
2. Son fácilmente interpretables.
3. Son adaptables y permiten la incorporación de conocimiento a priori de forma cualitativa.

En este trabajo estudiamos el uso de las redes bayesianas para analizar los datos sobre los invernaderos de la provincia de Almería recopilados por la Estación Experimental de Las Palmerillas (El Ejido, Almería). Este trabajo, restringido a los invernaderos, es un paso previo para un análisis más detallado de todo el tejido agrícola de la provincia, que comprenderá aspectos como los productores, las fincas, los cultivos, etc.
El principal problema desde el punto de vista del análisis radica en la heterogeneidad de las variables (continuas, discretas, cualitativas) que aparecen simultáneamente en una misma base de datos. A esto hay que añadir, ya a nivel computacional, las dificultades del manejo conjunto de un alto número de variables.

2 Descripción de los datos

El estudio se ha realizado sobre tres bases de datos con características de los invernaderos analizados (I1, I2 e I3). Pasamos a describir cada una de estas bases de datos.
Base de datos I1:

- Número de variables: 7.
- Número de registros:
- Descripción de las variables:
  - PREPSUELO: Tipo de preparación del suelo. Cualitativa.
  - SISTCULTIVO: Sistema de cultivo. Cualitativa.
  - SISTRIEGO: Sistema de riego. Cualitativa.
  - ESTRUCTURA: Estructura del invernadero. Cualitativa.

- ALTRASPA: Altura de la raspa. Continua.
- APOYOS_PER: Material de los apoyos perimetrales. Cualitativa.
- APOYOS_INT: Material de los apoyos interiores. Cualitativa.
- ANTIGÜEDAD: Años de antiguedad. Discreta.

Base de datos I2:

- Número de variables: 17.
- Número de registros:
- Descripción de las variables:
  - ESTRUCTURA: Estructura del invernadero. Cualitativa.
  - ALTRASPA: Altura de la raspa. Continua.
  - ANTIGÜEDAD: Años de antiguedad. Discreta.
  - SITRPLU: Sistema de recogida de pluviales. Cualitativa.
  - HORPASI: Pasillos hormigonados. Cualitativa.
  - MALLV: Tipo de malla en las ventanas. Cualitativa.
  - MALLS: Tipo de malla de sombreo. Cualitativa.
  - TVC: Tipo de ventana cenital.Cualitativa.
  - TVL: Tipo de ventana lateral.
  - SUPERFICIE: Superficie en metros cuadrados. Continua.
  - INSTRAFI: Instalación fija de tratamientos. Cualitativa.
  - PT: Tipo de pantalla térmica. Cualitativa.
  - VT: Tipo de ventilador. Cualitativa.
  - SC: Tipo de sistema de calefacción. Cualitativa.
  - SH: Sistema de humidificación. Cualitativa.
  - DBPUERTA: Indica la presencia de doble puerta. Cualitativa.
  - SICO2: Sistema de inyección de CO_2. Cualitativa.

Base de datos I3:

- Número de variables: 15.
- Número de registros:
- Descripción de las variables:
  - ESTRUCTURA: Estructura del invernadero. Cualitativa.

- SITRPLU: Sistema de recogida de pluviales. Cualitativa.
- HORPASI: Pasillos hormigonados. Cualitativa.
- TVC: Tipo de ventana cenital.Cualitativa.
- TVL: Tipo de ventana lateral.
- SUPERFICIE: Superficie en metros cuadrados. Continua.
- INSTRAFI: Instalación fija de tratamientos. Cualitativa.
- PT: Tipo de pantalla térmica. Cualitativa.
- VT: Tipo de ventilador. Cualitativa.
- SC: Tipo de sistema de calefacción. Cualitativa.
- SH: Sistema de humidificación. Cualitativa.
- DBPUERTA: Indica la presencia de doble puerta. Cualitativa.
- SICO2: Sistema de inyección de CO_2. Cualitativa.
- MALLV: Tipo de malla en las ventanas. Cualitativa.
- MALLS: Tipo de malla de sombreo. Cualitativa.

3 Modelos basados en redes bayesianas

Una *red bayesiana* representa una distribución de probabilidad multivariante, de manera que las relaciones de independencia entre las variables que la forman quedan identificadas de forma gráfica mediante el concepto de *d*-separación.

Definición 1. *(Pearl 1988) Dos variables A y B en una red bayesiana se dice que están* **d-separadas** *si todos los caminos entre A y B son como los que aparecen en la figura 1. Se dice además que C d-separa a A y B.*

El concepto de d-separación se corresponde con el de independencia condicional, de manera que dos variables (o conjuntos de variables) X e Y serán condicionalmente independientes dada una tercera variable (o conjunto de variables) Z si y sólo si Z d-separa a X e Y.

3.1 Construcción de las redes a partir de los datos

Existen diversas técnicas para construir redes bayesianas a partir de una base de datos. En este trabajo hemos empleado el algoritmo K2 (Cooper and Herskovits 1992) y el de búsqueda estocástica de vecindad variable (Puerta 2001). Pasamos a describir brevemente estos algoritmos.

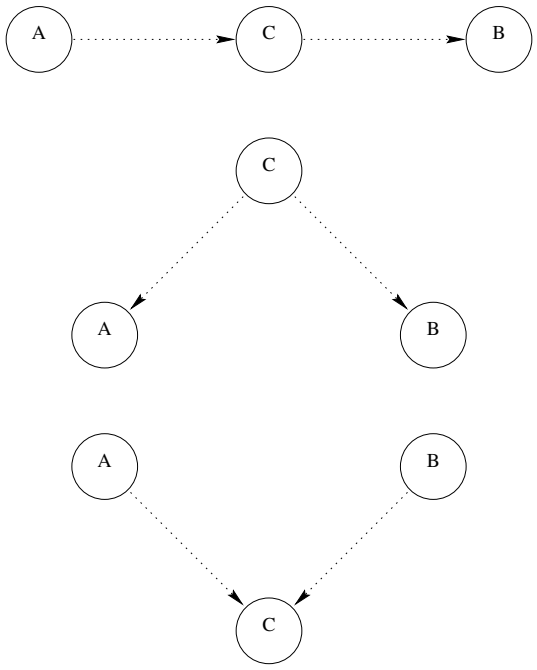

Figura 1: Caracterización gráfica del concepro de d-separación.

El algoritmo K2

El algoritmo K2 está basado en la optimización de una medida. Esa medida se usa para explorar, mediante un algoritmo de ascensión de colinas, el espacio de búsqueda formado por todas las redes que contienen las variables de la base de datos. Se parte de una red inicial y ésta se va modificando (añadiendo arcos, borrándolos o cambiándolos de dirección) obteniendo una nueva red con mejor medida. En concreto, la medida K2 (Cooper and Herskovits 1992) para una red G y una base de datos D es la siguiente:

$$f(G:D) = \log P(G) + \sum_{i=1}^{n} \left[\sum_{k=1}^{s_i} \left[\log \frac{\Gamma(\eta_{ik})}{\Gamma(N_{ik}+\eta_{ik})} + \sum_{j=1}^{r_i} \log \frac{\Gamma(N_{ijk}+\eta_{ijk})}{\Gamma(\eta_{ijk})} \right] \right] \quad , \quad (1)$$

donde N_{ijk} es la frecuencia de las configuraciones encontradas en la base de datos D de las variables x_i, donde n es el número de variables, tomando su j-ésimo valor y sus padres en G tomando su k-ésima configuración, donde s_i es el número de configuraciones posibles del conjunto de padres y r_i es el número de valores que puede tomar la variable x_i. Además, $N_{ik} = \sum_{j=1}^{r_i} N_{ijk}$ y Γ es la función Gamma.

El algoritmo de búsqueda estocástica con vecindad variable

Este algoritmo difiere del K2 en dos aspectos. Por un lado, el algoritmo de búsqueda no es una simple ascensión de colinas, sino que se agrupan los puntos del espacio de búsqueda por vecindades y se explora de forma local en dichas vecindades. Por otro lado, la medida de calidad de las redes no es necesariamente la K2. En este trabajo hemos probado el algoritmo con la medida K2 y también con la medida BIC, definida como sigue:

$$f(G:D) = \sum_{i=1}^{n} \sum_{j=1}^{r_i} \sum_{k=1}^{s_i} N_{ijk} \log \frac{N_{ijk}}{N_{ik}} - \frac{1}{2} C(G) \log N \ , \tag{2}$$

donde N es el número de registros de la base de datos y $C(G)$ es una medida de complejidad de la red G, definida como

$$C(G) = \sum_{i=1}^{n} (r_i - 1) s_i \ .$$

4 Estudio de independencias

Los modelos presentados en esta sección han sido construidos usando el programa Elvira (Elvira consortium 2002).

5 Agradecimientos

Este trabajo ha sido parcialmente subvencionado por el Ministerio de Ciencia y Tecnología a través del proyecto TIC2001-2973-C05-02 y por la Junta de Andalucía, grupo FQM244 del Plan Andaluz de Investigación.

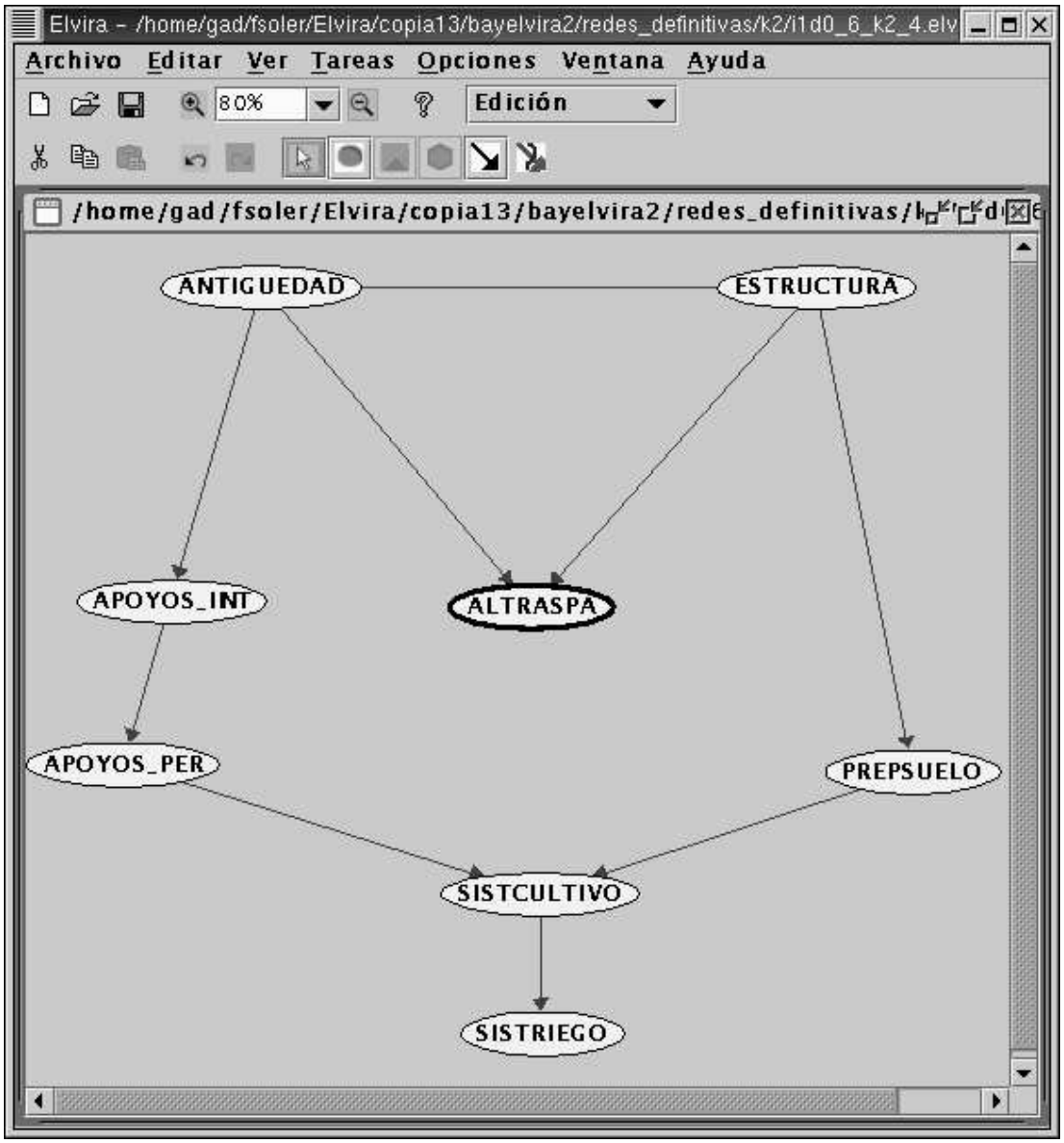

Figura 2: Red obtenida para la base de datos I1 con el algoritmo K2

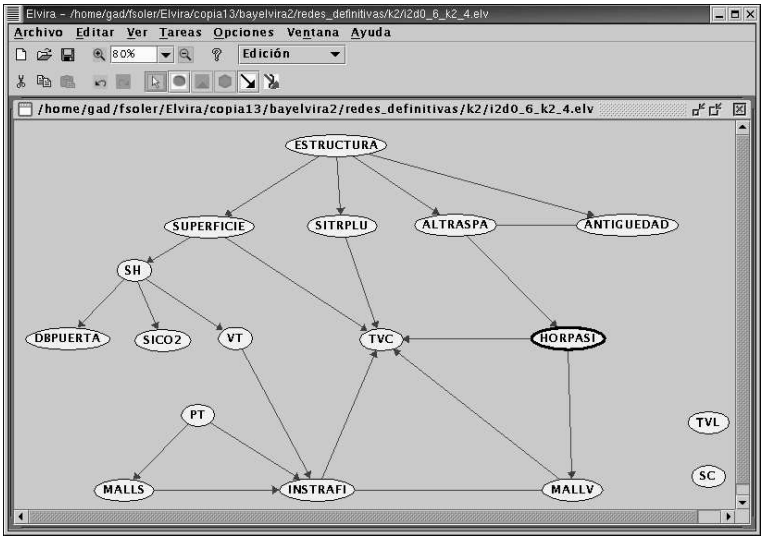

Figura 3: Red obtenida para la base de datos I2 con el algoritmo K2

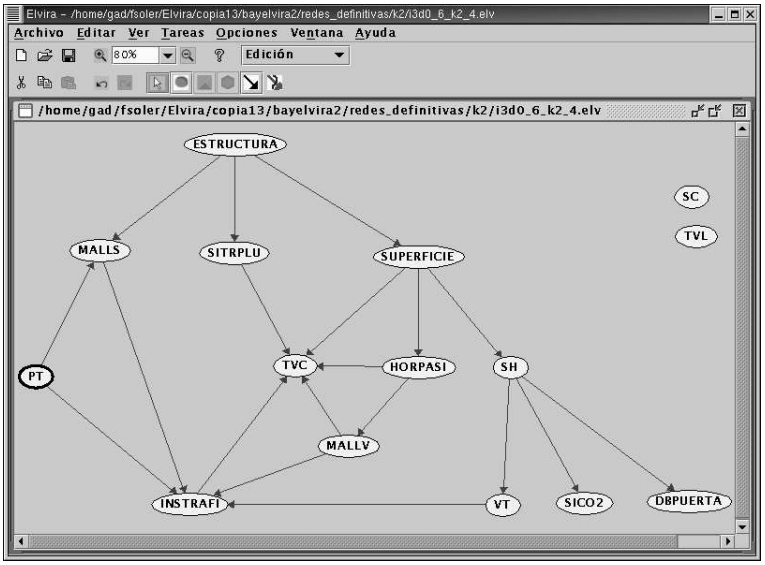

Figura 4: Red obtenida para la base de datos I3 con el algoritmo K2

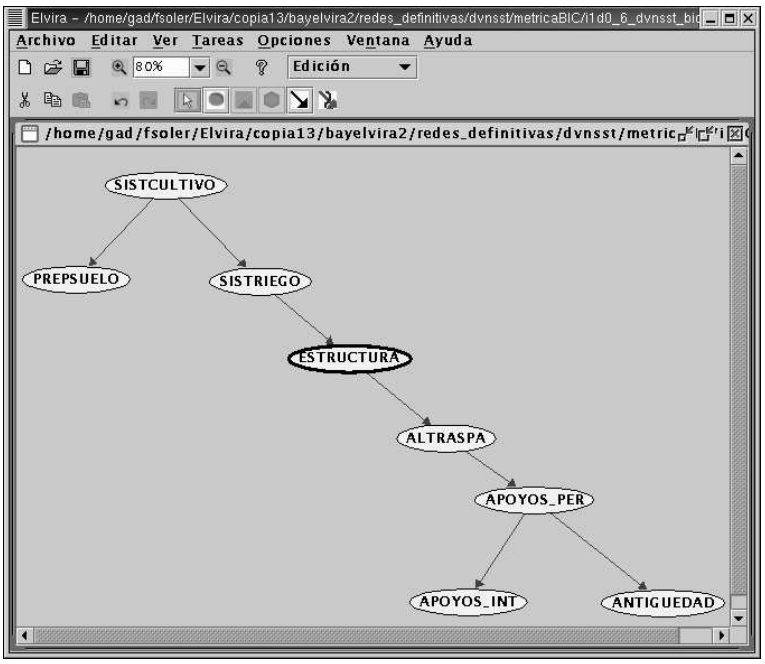

Figura 5: Red obtenida para la base de datos I1 con el algoritmo de búsqueda por vecindad variable con métrica BIC

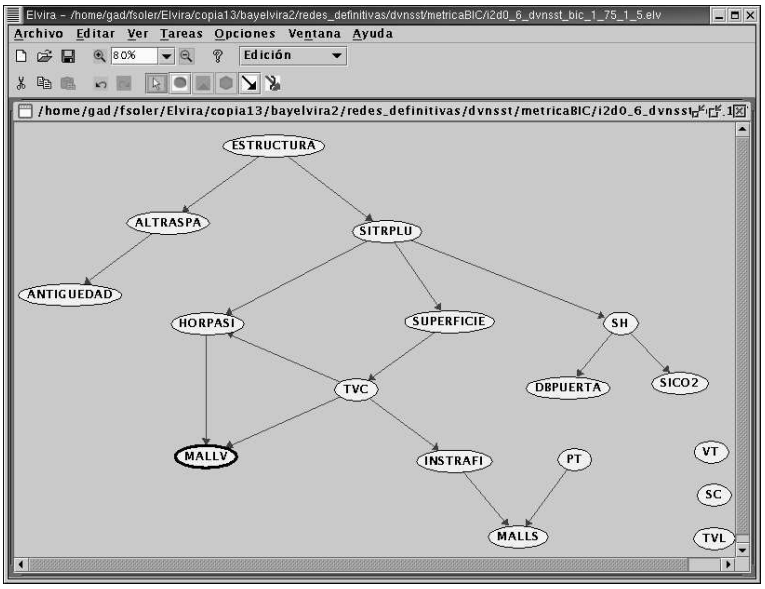

Figura 6: Red obtenida para la base de datos I2 con el algoritmo de búsqueda por vecindad variable con métrica BIC

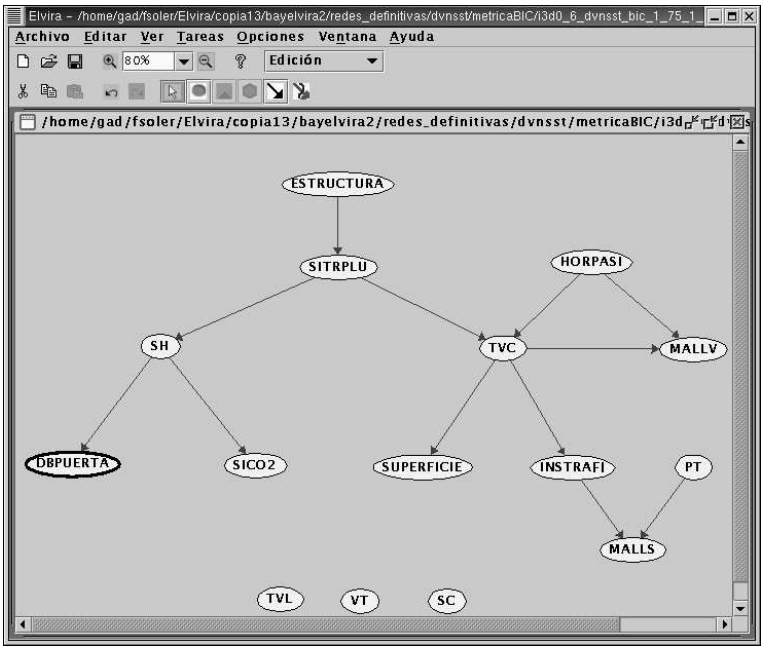

Figura 7: Red obtenida para la base de datos I3 con el algoritmo de búsqueda por vecindad variable con métrica BIC

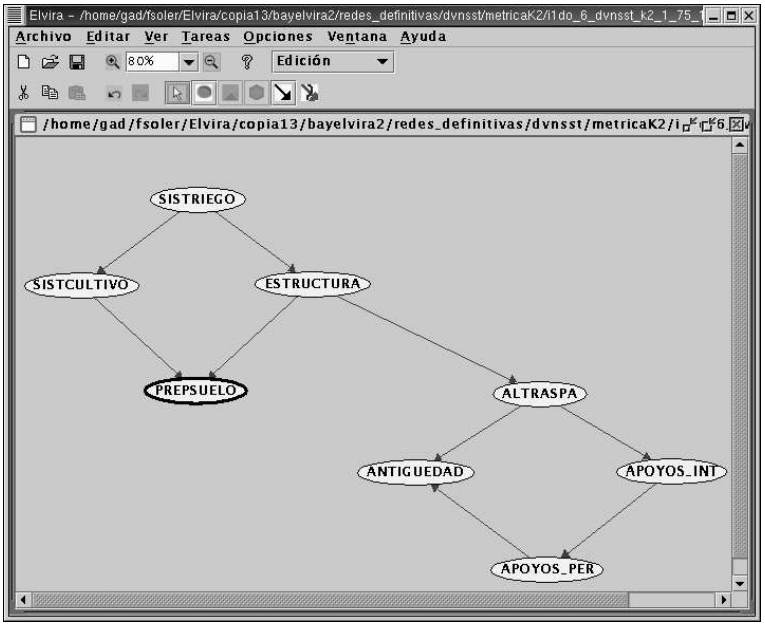

Figura 8: Red obtenida para la base de datos I1 con el algoritmo de búsqueda por vecindad variable con métrica K2

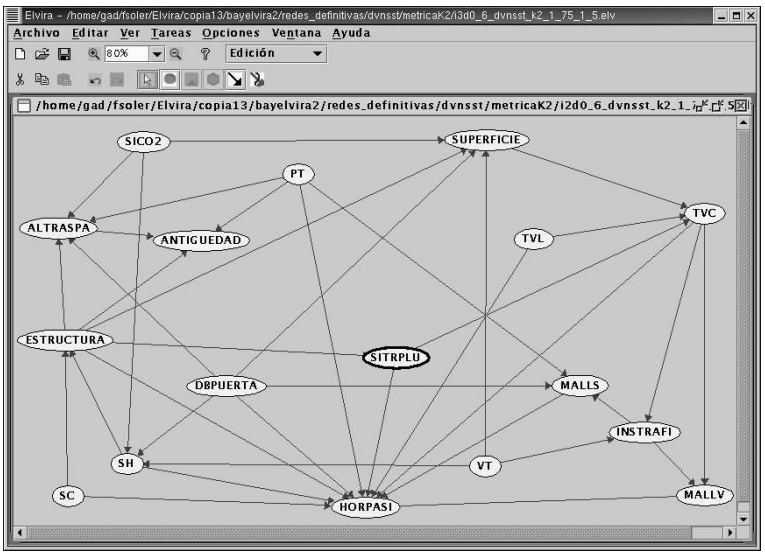

Figura 9: Red obtenida para la base de datos I2 con el algoritmo de búsqueda por vecindad variable con métrica K2

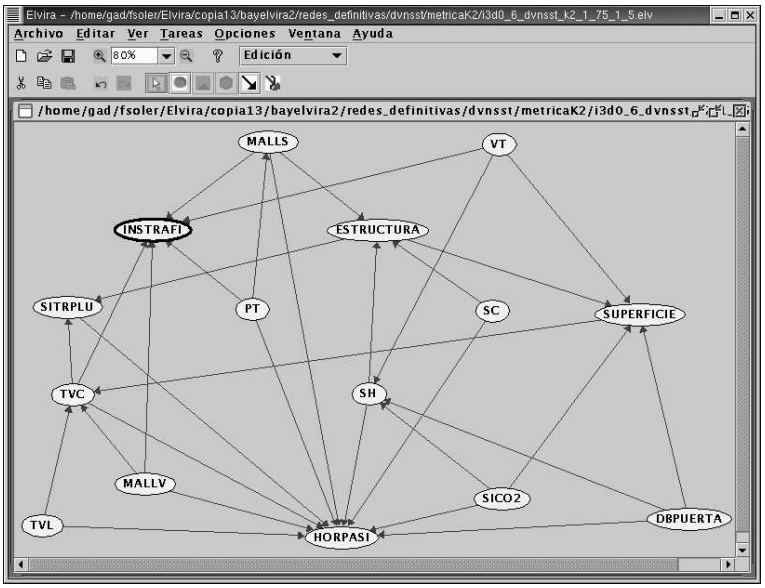

Figura 10: Red obtenida para la base de datos I3 con el algoritmo de búsqueda por vecindad variable con métrica K2

Referencias

Cano, A., S. Moral, and A. Salmerón (2000). Penniless propagation in join trees. *International Journal of Intelligent Systems 15, 1027–1059.*

Cooper, G. and E. Herskovits (1992). A Bayesian method for the induction of probabilistic networks from data. Machine Learning 9, 309–348.

Elvira consortium (2002). Elvira: An environment for probabilistic graphical models. In J. Gámez and A. Salmerón (Eds.), *Proceedings of the First European Workshop on Probabilistic Graphical Models (PGM'02)*, pp. 222–230.

Gámez, J. (1998). Abducción en modelos gráficos. In J. Gámez and J. Puerta (Eds.), *Sistemas expertos probabilísticos*, pp. 113–140.

Lauritzen, S. and D. Spiegelhalter (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B 50, 157–224.*

Madsen, A. and F. Jensen (1999). Lazy propagation: a junction tree inference algorithm based on lazy evaluation. Artificial Intelligence 113, 203–245.

Nilsson, D. (1998). An efficient algorithm for finding the m most probable configurations in probabilistic expert systems. *Statistics and Computing 8, 159–173.*

Pearl, J. (1988). Probabilistic reasoning in intelligent systems. *Morgan-Kaufmann (San Mateo).*

Puerta, J. (2001). Métodos locales y distribuidos para la construcción de redes de creencia estáticas y dinámicas. *Ph. D. thesis, Universidad de Granada.*