Geographical Clustering of Cancer Incidence by Means of Bayesian Networks and Conditional Gaussian Networks

J. M. Peña[1] **I. Izarzugaza**[2] **J. A. Lozano**[1] **E. Aldasoro**[2] **P. Larrañaga**[1]
[1]Dept. of Computer Science and Artificial Intelligence
University of the Basque Country, Donostia-San Sebastián, Spain
[2]Basque Health Service
Basque Government, Vitoria-Gasteiz, Spain
[1]{ccbpepaj, ccploalj, ccplamup}@si.ehu.es, [2]{info5-san, osasunka-san}@ej-gv.es

Abstract

With the aim of improving knowledge on the geographical distribution and characterization of malignant tumors in the Autonomous Community of the Basque Country (Spain), age-standardized cancer incidence rates of the 6 most frequent cancer types for patients of each sex between 1986 and 1994 are analyzed, in relation to the towns of the Community. Concretely, we perform a geographical clustering of the towns of the Community by means of Bayesian networks and conditional Gaussian networks. We present several maps that show the clusterings encoded by the learnt models. In addition to this, we outline the cancer incidence profile for each of the obtained clusters.

1 INTRODUCTION

One of the basic problems that arises in a great variety of fields, including pattern recognition, machine learning and statistics, is the so-called *data clustering problem* (Anderberg 1973, Hartigan 1975, Kaufman and Rosseeuw 1990). Despite the different interpretations and expectations it gives rise to, the generic data clustering problem involves the assumption that, in addition to the observed variables or predictive attributes, there is a *hidden* variable. This last unobserved variable would reflect the cluster membership for every case in the database. Thus, the data clustering problem is also referred to as an example of learning from *incomplete data* due to the existence of such a hidden variable. Incomplete data represents a special case of *missing data* where all the missing entries are concentrated in a single variable: the hidden cluster variable. That is, we refer to a given database as incomplete when all the cases are unlabeled.

From the point of view adopted in this paper, the data clustering problem may be defined as the inference of the joint generalized probability distribution for a given database. Concretely, we focus on the unsupervised learning of *Bayesian networks* (Pearl 1988, Peña et al. 2000a, 2000b) and *conditional Gaussian networks* (Lauritzen and Wermuth 1989, Lauritzen 1992, 1996, Peña et al. 2000c, 2000d) to obtain a geographical clustering of the towns of the Autonomous Community of the Basque Country (ACBC) according to their cancer incidence rates. Then, the learning is performed from a set of databases where every town is described by the *age-standardized cancer incidence rates* of the 6 most frequent cancer types for patients of each sex between 1986 and 1994.

The remainder of this paper is organized as follows. In Section 2, we introduce Bayesian networks and conditional Gaussian networks applied to data clustering. Section 3 is dedicated to explain the problem of cancer incidence in the ACBC and the construction of the databases used in the subsequent learning process. We present some experimental results in Section 4. Basically, the results consist of some coloured maps representing the clusterings encoded by the learnt models. Finally, we draw conclusions in Section 5.

2 BAYESIAN NETWORKS AND CONDITIONAL GAUSSIAN NETWORKS FOR DATA CLUSTERING

In this section, we introduce two classes of probabilistic graphical models applied to data clustering: Bayesian networks and conditional Gaussian networks.

All through this paper we follow the usual convention of denoting variables by upper-case letters and their states by the same letters in lower-case. We use a letter or letters in bold-face upper-case to designate a set of variables and the same bold-face lower-case letter or letters to denote an assignment of state to each vari-

able in a given set. The joint generalized probability distribution of $\mathbf{X}$ is represented as $\rho(\mathbf{x})$. Additionally, $\rho(\mathbf{x} \mid \mathbf{y})$ denotes the generalized conditional probability distribution of $\mathbf{X}$ given $\mathbf{Y} = \mathbf{y}$. If all the variables in $\mathbf{X}$ are discrete, then $\rho(\mathbf{x}) = p(\mathbf{x})$ is the joint probability mass function of $\mathbf{X}$. Thus, $p(\mathbf{x} \mid \mathbf{y})$ denotes the conditional probability mass function of $\mathbf{X}$ given $\mathbf{Y} = \mathbf{y}$. If all the variables in $\mathbf{X}$ are continuous, then $\rho(\mathbf{x}) = f(\mathbf{x})$ is the joint probability density function of $\mathbf{X}$. Thus, $f(\mathbf{x} \mid \mathbf{y})$ denotes the conditional probability density function of $\mathbf{X}$ given $\mathbf{Y} = \mathbf{y}$.

When facing a data clustering problem it is assumed the existence of a $(n+1)$-dimensional random variable $\mathbf{X}$ partitioned as $\mathbf{X} = (\mathbf{Y}, C)$ into a n-dimensional observed variable $\mathbf{Y}$ and a unidimensional discrete hidden variable C. In the particular case of every component Y_i of $\mathbf{Y}$ being discrete, the probabilistic graphical models that we aim to learn are called Bayesian networks. On the other hand, if every component Y_i of $\mathbf{Y}$ being continuous, then the probabilistic graphical models are named conditional Gaussian networks.

2.1 BAYESIAN NETWORKS

Given a discrete random variable $\mathbf{X} = (\mathbf{Y}, C) = (Y_1, \ldots, Y_n, C)$, a Bayesian network (BN) for $\mathbf{X}$ (Pearl 1988, Peña et al. 2000a, 2000b) is a graphical factorization of the joint probability distribution of $\mathbf{X}$. When applied to data clustering, a BN is defined by a directed acyclic graph $\mathbf{s}$ (model structure) determining the conditional (in)dependencies among the variables of $\mathbf{Y}$ and a set of local probability distributions. The model structure yields to a factorization of the joint probability distribution for $\mathbf{X}$ as follows:

$$p(\mathbf{x}) = p(c)p(\mathbf{y} \mid c) = p(c) \prod_{i=1}^{n} p(y_i \mid \mathbf{pa}(\mathbf{s})_i, c) \qquad (1)$$

where $\mathbf{pa}(\mathbf{s})_i$ denotes the state of the parents of Y_i in $\mathbf{s}$, $\mathbf{Pa}(\mathbf{s})_i$, consistent with $\mathbf{x}$.

The local probability distributions of the BN are those in Equation 1 and we assume that they depend on a finite set of parameters $\boldsymbol{\theta}_\mathbf{s} \in \boldsymbol{\Theta}_\mathbf{s}$. Moreover, let $\mathbf{s}^h$ denote the hypothesis that the conditional (in)dependence assertions implied by $\mathbf{s}$ hold in the true joint probability distribution of $\mathbf{X}$. Therefore, Equation 1 can be rewritten as follows:

$$p(\mathbf{x} \mid \boldsymbol{\theta}_\mathbf{s}, \mathbf{s}^h) = p(c \mid \boldsymbol{\theta}_\mathbf{s}, \mathbf{s}^h) p(\mathbf{y} \mid \boldsymbol{\theta}^c_\mathbf{s}, \mathbf{s}^h)$$

$$= p(c \mid \boldsymbol{\theta}_\mathbf{s}, \mathbf{s}^h) \prod_{i=1}^{n} p(y_i \mid \mathbf{pa}(\mathbf{s})_i, \boldsymbol{\theta}^c_i, \mathbf{s}^h) \qquad (2)$$

where $\boldsymbol{\theta}^c_\mathbf{s} = (\boldsymbol{\theta}^c_1, \ldots, \boldsymbol{\theta}^c_n)$ denotes the parameters for the local probability distributions when $C = c$.

In this paper, we limit our discussion to the case in which the local probability distributions of each variable of the BN consist of a set of multinomial distributions, one for each configuration of the parents and the cluster variable C.

2.2 CONDITIONAL GAUSSIAN NETWORKS

A random variable $\mathbf{X} = (\mathbf{Y}, C) = (Y_1, \ldots, Y_n, C)$, being $\mathbf{Y}$ continuous and C discrete, is said to have a *conditional Gaussian distribution* (Lauritzen and Wermuth 1989, Lauritzen 1992, 1996) if the distribution of $\mathbf{Y}$, conditioned on each state of C, is a multivariate normal distribution:

$$f(\mathbf{y} \mid C = c) \equiv \mathcal{N}(\mathbf{y}; \boldsymbol{\mu}(c), \boldsymbol{\Sigma}(c)) \qquad (3)$$

whenever $p(c) = p(C = c) > 0$. Given $C = c$, $\boldsymbol{\mu}(c)$ is the n-dimensional mean vector, and $\boldsymbol{\Sigma}(c)$, the $n \times n$ variance matrix, is positive definite.

We define a conditional Gaussian network (CGN) for $\mathbf{X}$ (Lauritzen and Wermuth 1989, Lauritzen 1992, 1996, Peña et al. 2000c, 2000d) as a probabilistic graphical model that encodes a conditional Gaussian distribution for $\mathbf{X}$. Thus, a CGN is defined by a directed acyclic graph $\mathbf{s}$ (model structure) determining the conditional (in)dependencies among the variables of $\mathbf{Y}$, a set of local probability density functions and a multinomial distribution for the variable C. The model structure yields to a factorization of the joint generalized probability density function for $\mathbf{X}$ as follows:

$$\rho(\mathbf{x}) = p(c) f(\mathbf{y} \mid c) = p(c) \prod_{i=1}^{n} f(y_i \mid \mathbf{pa}(\mathbf{s})_i, c) \qquad (4)$$

where $\mathbf{pa}(\mathbf{s})_i$ denotes the state of the parents of Y_i in $\mathbf{s}$, $\mathbf{Pa}(\mathbf{s})_i$, consistent with $\mathbf{x}$.

The local probability density functions and the multinomial distribution of the CGN are those in the previous equation and we assume that they depend on a finite set of parameters $\boldsymbol{\theta}_\mathbf{s} \in \boldsymbol{\Theta}_\mathbf{s}$. Moreover, let $\mathbf{s}^h$ denote the hypothesis that the conditional (in)dependence assertions implied by $\mathbf{s}$ hold in the true joint generalized probability density function of $\mathbf{X}$. Therefore, Equation 4 can be rewritten as follows:

$$\rho(\mathbf{x} \mid \boldsymbol{\theta}_\mathbf{s}, \mathbf{s}^h) = p(c \mid \boldsymbol{\theta}_\mathbf{s}, \mathbf{s}^h) f(\mathbf{y} \mid \boldsymbol{\theta}^c_\mathbf{s}, \mathbf{s}^h)$$

$$= p(c \mid \boldsymbol{\theta}_\mathbf{s}, \mathbf{s}^h) \prod_{i=1}^{n} f(y_i \mid \mathbf{pa}(\mathbf{s})_i, \boldsymbol{\theta}^c_i, \mathbf{s}^h) \qquad (5)$$

where $\boldsymbol{\theta}^c_\mathbf{s} = (\boldsymbol{\theta}^c_1, \ldots, \boldsymbol{\theta}^c_n)$ denotes the parameters for the local probability density functions when $C = c$.

In order to encode a conditional Gaussian distribution for $\mathbf{X}$, each local probability density function of the CGN should be the linear-regression model. Thus, when $C = c$:

$$f(y_i \mid \mathbf{pa}(\mathbf{s})_i, \boldsymbol{\theta}_i^c, \mathbf{s}^h)$$

$$\equiv \mathcal{N}(y_i; m_i^c + \sum_{y_j \in \mathbf{pa}(\mathbf{s})_i} b_{ji}^c (y_j - m_j^c), v_i^c) \qquad (6)$$

where $\mathcal{N}(y; \mu, \sigma^2)$ is a univariate normal distribution with mean μ and standard deviation σ ($\sigma > 0$). Given this form, a missing arc from Y_j to Y_i implies that $b_{ji}^c = 0$ in the linear-regression model. When $C = c$, the local parameters are $\boldsymbol{\theta}_i^c = (m_i^c, \mathbf{b}_i^c, v_i^c)$, $i = 1, \ldots, n$, where $\mathbf{b}_i^c = (b_{1i}^c, \ldots, b_{i-1i}^c)^t$ is a column vector.

The interpretation of the components of the local parameters $\boldsymbol{\theta}_i^c$, $i = 1, \ldots, n$, is as follows: given $C = c$, m_i^c is the unconditional mean of Y_i, v_i^c is the conditional variance of Y_i given $\mathbf{Pa}(\mathbf{s})_i$, and b_{ji}^c, $j = 1, \ldots, i-1$, is a linear coefficient reflecting the strength of the relationship between Y_j and Y_i.

2.3 UNSUPERVISED LEARNING OF BNS AND CGNS

In order to perform the unsupervised learning of BNs and CGNs, we consider two techniques: the well-known *Bayesian Structural EM (BS-EM) algorithm* (Friedman 1998) to learn CGNs, and the *Bayesian Structural BC+EM (BS-BC+EM) algorithm* (Peña et al. 2000a) to learn BNs. Whereas the former algorithm has received special attention in literature and it has motivated several variants of itself due to its good performance, the latter is an improved version of the BS-EM algorithm for discrete domains.

When applying the BS-EM and BS-BC+EM algorithms to a data clustering problem, we assume that we have a database of N cases, $\mathbf{d} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, where every case is represented by an assignment to n of the $n+1$ variables involved in the problem domain. So, there are $(n+1)N$ random variables that describe the database. Let $\mathbf{O}$ denote the set of observed variables, that is, the nN variables that have assigned values. Similarly, let $\mathbf{H}$ denote the set of hidden or unobserved variables, that is, the N variables that reflect the unknown cluster membership of each case of $\mathbf{d}$.

Both algorithms perform a search over the space of models based on the well-known *EM algorithm* (Dempster et al. 1977, McLachlan and Krishnan 1997) and the direct optimization of the Bayesian score. This results in an attempt to maximize the expected Bayesian score at each iteration instead of the true Bayesian score. As it is shown in Figure 1, both algorithms are comprised of two steps: an optimization of the model parameters and a structural search for model selection.

loop $l = 0, 1, \ldots$
1. Compute the MAP parameters $\widehat{\boldsymbol{\theta}}_{\mathbf{s}_l}$ for $\mathbf{s}_l$ given $\mathbf{o}$
2. Perform search over model structures, evaluating each model structure by:
   $Score(\mathbf{s} : \mathbf{s}_l) = E[\log \rho(\mathbf{h}, \mathbf{o}, \mathbf{s}^h) \mid \mathbf{o}, \widehat{\boldsymbol{\theta}}_{\mathbf{s}_l}, \mathbf{s}_l^h]$
   $= \sum_{\mathbf{h}} p(\mathbf{h} \mid \mathbf{o}, \widehat{\boldsymbol{\theta}}_{\mathbf{s}_l}, \mathbf{s}_l^h) \log \rho(\mathbf{h}, \mathbf{o}, \mathbf{s}^h)$
3. Let $\mathbf{s}_{l+1}$ be the model structure with the highest score among these encountered in the search
4. **if** $Score(\mathbf{s}_l : \mathbf{s}_l) = Score(\mathbf{s}_{l+1} : \mathbf{s}_l)$ **then** return $(\mathbf{s}_l, \widehat{\boldsymbol{\theta}}_{\mathbf{s}_l})$

Figure 1: The BS-EM and BS-BC+EM algorithms.

Concretely, the optimization of the parameters (step 1 in Figure 1) consists of the search for the maximum a posterior parameters (MAP) for the current model. The difference between the BS-EM and BS-BC+EM algorithms is that, whereas the former makes use of the EM algorithm to perform such a search, the latter takes advantage of the *BC+EM method* (Peña et al. 2000a, 2000b). The BC+EM method represents an alternative technique to perform the parameter search step in discrete domains. It exhibits a faster convergence rate as well as a more effective and robust behaviour than the EM algorithm. Basically, the BC+EM method is comprised of an alternation between the *Bound and Collapse method* (Ramoni and Sebastiani 1998, 1999) and the EM algorithm.

To completely specify the BS-EM and BS-BC+EM algorithms, we have to decide on the structural search procedure (step 2 in Figure 1). The usual approach is to perform a greedy hill-climbing search over model structures considering all possible additions, removals and reversals of one arc at each point in the search. This structural search procedure is desirable as it exploits the decomposition properties of BNs and CGNs, and the factorization properties of the Bayesian score for complete data. However, any structural search procedure that exploits these referred properties can be used. The log marginal likelihood of the expected complete data is usually chosen as the score to guide the structural search.

The direct application of the learning algorithms as they appear depicted in Figure 1 may result in unrealistic and inefficient solutions due to the fact that the computation of $Score(\mathbf{s} : \mathbf{s}_l)$ implies a huge computational expense as it takes account of every possible completion of the database. It is common to use re-

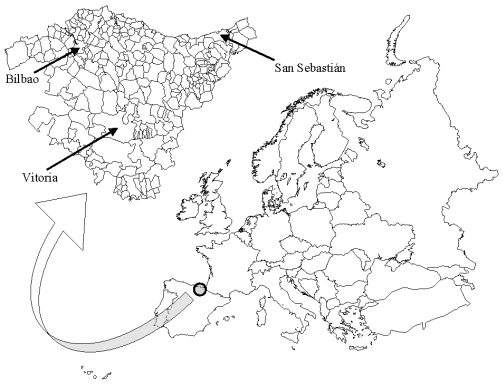

Figure 2: Map of the towns of the ACBC (the 3 major towns appear indicated).

laxed versions of these algorithms that only consider the most likely completion of the database to compute $Score(\mathbf{s} : \mathbf{s}_l)$ instead of considering every possible completion. Thus, these relaxed version of the BS-EM and BS-BC+EM algorithms are comprised of the iteration of a parametric optimization for the current model, and a structural search once the database has been completed with the most likely completion by using the best estimate of the joint generalized probability distribution of the data so far (current model). The completion is achieved by calculating the posterior probability distribution of the cluster variable C for each case of the database, $p(c \mid \mathbf{y}_i, \widehat{\boldsymbol{\theta}}_{\mathbf{s}_l}, \mathbf{s}_l^h)$. The case is assigned to the cluster where the maximum of this posterior probability distribution of C is reached. We use these relaxed versions in our experiments.

3 CANCER INCIDENCE IN THE ACBC BETWEEN 1986 AND 1994

The ACBC is located in the north of Spain (Figure 2) and it covers some 7,234 sq km (1.4 % of the total area of Spain). According to the 1986-1994 census, the population of the ACBC was around 2,100,000 inhabitants on average (5 % of the population of Spain).

With the purpose of obtaining a geographical clustering of the towns of the ACBC, we created one database for patients of each sex from the cancer registry of the ACBC between 1986 and 1994. The male database summarized all the records of male patients in 231 cases, while the female database summarized all the records of female patients in 221 cases. Each case in the male and female databases represented one town of the ACBC with more than 100 male inhabitants for the former database and with more than 100 female inhabitants for the latter database. Each town of each of the 2 databases was characterized by 6 attributes representing the age-standardized cancer incidence rates per 100,000 of the 6 most frequent cancer types for patients of that concrete sex. Table 1 shows these 6 cancer types for patients of each sex by their body sites encoded following the International Classification of Diseases for Oncology (ICD-O) (WHO, 1976).

Table 1: The 6 most frequent cancer types for patients of each sex by their body sites encoded according to the ICD-O.

Male patients	
Cancer type	ICD-O
m1	140-149
m2	150-159
m3	160-165
m4	185-189
m5	169, 196
m6	170, 171, 173, 175, 190-195, 199

Female patients	
Cancer type	ICD-O
f1	140-149, 160-165
f2	150-159
f3	174
f4	169, 196
f5	179-184, 188, 189
f6	170, 171, 173, 190-195, 199

Age-standardized cancer incidence rates may be defined as the hypothetical rates that would be observed in a population with a standard age distribution referred to as *standard population*. They allow comparisons between populations with different age distributions. In our case, the 1991 census of the ACBC was used as the standard population. Each of the 6 age-standardized cancer incidence rates, here denoted as $ASCIR_i$, $i = 1, \ldots, 6$, was calculated as follows:

$$ASCIR_i = \frac{\sum_{j=1}^{A} a_{ij} q_j}{\sum_{j=1}^{A} q_j} \tag{7}$$

where $A = 19$ represents the number of age-groups and q_j, $j = 1, \ldots, A$, the population of the j-th age-group in the standard population ($\sum_{j=1}^{A} q_j = 100,000$). Additionally, a_{ij}, $j = 1, \ldots, A$, denotes the *age-specific cancer incidence rate* per 100,000 of the i-th cancer type for the j-th age-group calculated as:

$$a_{ij} = \frac{t_{ij}}{r_j} 10^5 \tag{8}$$

being t_{ij} the number of patients suffering from the i-th cancer type who belong to the j-th age-group, and r_j the risk population for the j-th age-group.

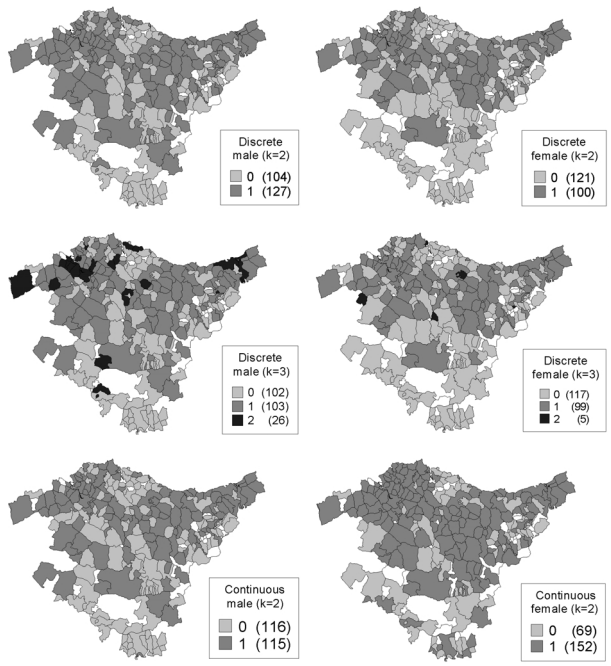

Figure 3: Maps showing the geographical clusterings encoded by the learnt models (white towns are those towns excluded from out study).

As the 2 databases constructed were essentially continuous, every age-standardized cancer incidence rate was discretized into 4 equal frequency intervals in order to perform the unsupervised learning of BNs.

3.1 RESULTS

In this subsection, some experimental results are presented. In order to do this, the models learnt by the BS-EM and BS-BC+EM algorithms from the constructed databases are analyzed to obtain several coloured maps of the ACBC showing the geographical clusterings encoded by these models (see Figure 3). The analysis of the elicited models also helps us to characterize each of the obtained clusters in relation to the rest of them by their cancer incidence profile.

It should be noticed that the learnt models do not provide us with explicit partitions of the databases into clusters but with an encoding of the joint generalized probability distribution of each of the databases previously constructed. However, every learnt model allows us to obtain a clustering that partitions the towns into clusters by assigning each of them to the cluster with the highest posterior probability given that concrete model.

From a structural point of view, the models elicited by the BS-EM and BS-BC+EM algorithms are very simple, i.e., very few arcs between predictive attributes are learnt. Thus, these learnt BNs and CGNs indicate us that the majority of the cancer incidence rates is conditionally independent of the rest given a value for the cluster variable C. This fact is considered reasonable by the experts.

Figure 3 shows the clusters for patients of each sex encoded by the learnt BNs and CGNs when assuming the existence of 2 ($k = 2$) and 3 ($k = 3$) clusters. For the continuous databases, we only report the results achieved when $k = 2$ as those are almost exactly repeated when $k = 3$.

It is interesting to notice that there are 2 main clusters almost independently of both the database used in the learning and the assumed number of clusters. The cluster labeled with 1 (dark grey) groups the most densely populated and industrial towns of the ACBC, while the cluster labeled with 0 (light grey) represents the small villages. This result is much more noticeable for the male than for the female databases because the cluster labeled with 1 groups the most densely populated and industrial towns of the ACBC more accurately for the former database than for the latter.

It is also worth paying attention to the clusters obtained for the discrete male database when $k = 3$: the cluster that represents the most densely populated area in the clustering obtained for the same database with $k = 2$ appears divided into 2 sub-clusters (dark grey and black) showing a different cancer incidence profile for the 3 major towns of the ACBC: San Sebastián and Bilbao share the same profile as they belong to the same cluster (black), whereas Vitoria appears to have a different one as it is grouped under the cluster labeled with 1 (dark grey).

Roughly speaking about the characterization of each cluster according to the BNs learnt from the discrete male ($k = 2$) and discrete female ($k = 2, 3$) databases, the cluster representing the most densely populated area of the ACBC shows higher cancer incidence rates than the cluster that groups the small villages in the 6 cancer types considered. Similar results are observed according to the CGNs learnt from the continuous male database. However, the results for the continuous female database do not follow this tendency: the most densely populated area exhibits higher cancer incidence rates than the least densely populated area in 3 out of the 6 cancer types (f2, f4 and f6), and lower rates in the remaining 3. We attribute this disparity between the learnt BNs and CGNs to the normality assumption made by the latter models and the discretization that the former models require.

It is specially appealing the fact that, when learning BNs from the discrete male database with $k = 3$, the

cluster labeled with 2 (black) shows higher cancer incidence rates than the other 2 clusters in the 6 cancer types considered. We should recall that San Sebastián and Bilbao belong to this referred cluster, while Vitoria appears in the cluster labeled with 1.

In spite of the disrupting effects that the required discretization process may imply, BNs appear to be more appropriate than CGNs when applied to the data clustering problem depicted in this paper. In our opinion, the assumption of normality made when learning CGNs is too strong for the continuous databases considered (few data and too sparse), whereas the assumption of multinomiality implied when learning BNs is a more realistic approach.

4 CONCLUSIONS

With the aim of improving knowledge on the geographical distribution and characterization of the 6 most frequent cancer types in the ACBC, BNs and CGNs have been elicited from continuous and discrete databases. The learnt models have helped to present several geographical clusterings of the towns of the ACBC as well as to characterize each of the clusters obtained according to the cancer incidence profile of the towns belonging to it. The classes of probabilistic graphical models considered have exhibited a promising behaviour in the problem domain depicted in this paper as they have provided us with valuable insights. However, the results reported in this study should be considered as a preliminary stage for further investigation in this field due to the fact that this work has been carried out by taking into account a very short number of towns and some of them had short population for providing us with reliable estimates.

Acknowledgements

J. M. Peña wishes to thank Dr. Steve Ellacott for his interest in this work and his useful comments. He also made it possible to visit the School of Computing and Mathematical Sciences of the University of Brighton, Brighton, United Kingdom.

This work was supported by the Spanish Ministry of Education and Culture under AP97 44673053 grant.

References

M. R. Anderberg (1973). *Cluster Analysis for Applications*. Academic Press, New York, NY.

A. Dempster, N. Laird and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B* **39**, 1-38.

N. Friedman (1998). The Bayesian Structural EM algorithm. *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, Inc., San Francisco, CA, 129-138.

J. A. Hartigan (1975). *Clustering Algorithms*. John Wiley & Sons, Inc., Canada.

L. Kaufman and P. Rousseeuw (1990). *Finding Groups in Data*. John Wiley & Sons, Inc., New York, NY.

S. L. Lauritzen (1992). Propagation of probabilities, means and variances in mixed graphical association models. *Journal of the American Statistical Association* **87**, 1098-1108.

S. L. Lauritzen (1996). *Graphical Models*. Clarendon Press, Oxford, United Kingdom.

S. L. Lauritzen and N. Wermuth (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics* **17**, 31-57.

G. J. McLachlan and T. Krishnan (1997). *The EM Algorithm and Extensions*. John Wiley & Sons, Inc., New York, NY.

J. Pearl (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, Inc., Palo Alto, CA.

J. M. Peña, J. A. Lozano and P. Larrañaga (2000a). An improved Bayesian Structural EM algorithm for learning Bayesian networks for clustering. *Pattern Recognition Letters* **21**, 779-786.

J. M. Peña, J. A. Lozano and P. Larrañaga (2000b). Learning recursive Bayesian multinets for data clustering by means of constructive induction. *Machine Learning*, accepted.

J. M. Peña, J. A. Lozano and P. Larrañaga (2000c). Performance evaluation of compromise conditional Gaussian networks for data clustering. Submitted.

J. M. Peña, J. A. Lozano and P. Larrañaga (2000d). Learning conditional Gaussian networks for data clustering via edge exclusion tests. Submitted.

M. Ramoni and P. Sebastiani (1998). Parameter Estimation in Bayesian Networks from Incomplete Databases. *Intelligent Data Analysis* **2**.

M. Ramoni and P. Sebastiani (1999). Learning Conditional Probabilities from Incomplete Data: An Experimental Comparison. *Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics*, Morgan Kaufmann, Inc., San Mateo, CA.

WHO (1976). *International Classification of Diseases for Oncology*. 1st edition. World Health Organization, Geneva, Switzerland.