Environmental Modelling & Software 25 (2010) 1630–1639

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

# Hybrid Bayesian network classifiers: Application to species distribution models

P.A. Aguilera [a], A. Fernández [b,*], F. Reche [b], R. Rumí [b]

[a] Informatics and Environment Research Group, Dept. of Ecology, University of Almería, La Cañada de San Urbano s/n, 04120 Almería, Spain
[b] Dept. of Statistics and Applied Mathematics, University of Almería, La Cañada de San Urbano s/n, 04120 Almería, Spain

ARTICLE INFO

*Article history:*
Received 16 December 2009
Received in revised form
20 April 2010
Accepted 22 April 2010

*Keywords:*
Hybrid Bayesian networks
Classification
Mixtures of truncated exponentials
Conservation planning

ABSTRACT

Bayesian networks are one of the most powerful tools in the design of expert systems located in an uncertainty framework. However, normally their application is determined by the discretization of the continuous variables. In this paper the naïve Bayes (NB) and tree augmented naïve Bayes (TAN) models are developed. They are based on Mixtures of Truncated Exponentials (MTE) designed to deal with discrete and continuous variables in the same network simultaneously without any restriction. The aim is to characterize the habitat of the spur-thighed tortoise (*Testudo graeca graeca*), using several continuous environmental variables, and one discrete (binary) variable representing the presence or absence of the tortoise. These models are compared with the full discrete models and the results show a better classification rate for the continuous one. Therefore, the application of continuous models instead of discrete ones avoids loss of statistical information due to the discretization. Moreover, the results of the TAN continuous model show a more spatially accurate distribution of the tortoise. The species is located in the Doñana Natural Park, and in semiarid habitats. The proposed continuous models based on MTEs are valid for the study of species predictive distribution modelling.



## 1. Introduction

Over the last decade, advances in species predictive distribution modelling have been paralleled by the evolution and the development of geographical information systems (GIS), remote sensing, statistical modelling and database management (Guisan and Zimmermann, 2000; Austin, 2002; Lehmann et al., 2002a; Segurado and Araújo, 2004). Statistical models relate observations of species, communities or diversity (Manel et al., 2001; Graham et al., 2004a; Brotons et al., 2004; Guisan and Thuiller, 2005; Wintle et al., 2005) to environmental predictors, and project the fitted relationships into geographical space to produce distribution maps (Maggini et al., 2009). The modelling of species distribution is a useful tool (Guisan and Thuiller, 2005) that is widely used in spatial ecology, biogeography and conservation biology. The models have contributed significantly to test biogeographical, ecological and evolutionary hypotheses (Anderson et al., 2002; Graham et al., 2004b), for assessing species invasion and proliferation (Peterson, 2003), for rare species distribution (Guisan et al., 2006), for supporting conservation planning and reserve selection (Ferrier, 2002; Araújo et al., 2004), and for the study of the impacts of global change (Peterson et al., 2002; Midgley et al., 2003; Thuiller, 2004; Araújo and Pearson, 2006). Many statistical techniques have been applied to modelling (Guisan and Zimmermann, 2000; Wintle et al., 2005; Burgmann et al., 2005; Pearson et al., 2006; Elith et al., 2006), including classical statistical models such as generalized linear regressions (Guisan et al., 1999), generalized additive models (Luoto et al., 2005), generalized regression analysis and spatial prediction (GRASP) (Lehmann et al., 2002b) or logistic regressions (Manel et al., 2001). Recently, machine learning methods such as classification and regression trees (Miller and Franklin, 2002; Dźeroski and Drumm, 2003) and neural networks (Moisen and Frescino, 2002; Dedecker et al., 2004) have also been applied.

Bayesian networks (Jensen and Nielsen, 2007) are one of the most powerful tools in the design of expert systems located in an uncertainty framework (probabilistic expert system). They have been applied in solving environmental problems such as eutrophication in an estuary (Borsuk et al., 2004), credal classification in agriculture (Zaffalon, 2005), management of endangered species (Borsuk et al., 2006; Pollino et al., 2007), water resources planning (Bromley et al., 2005) and conservation of dunnarts (Smith et al., 2007). Graphically, a Bayesian network is a directed acyclic graph in which the nodes represent variables and the arcs represent probabilistic dependence or independence between the nodes. Bayesian networks capture the uncertainty that an event occurs through a set of conditional probability distributions associated

* Corresponding author. Tel.: +34 950015748; fax: +34 950015167.
*E-mail addresses:* aguilera@ual.es (P.A. Aguilera), afalvarez@ual.es (A. Fernández), freche@ual.es (F. Reche), rrumi@ual.es (R. Rumí).

1364-8152/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2010.04.016

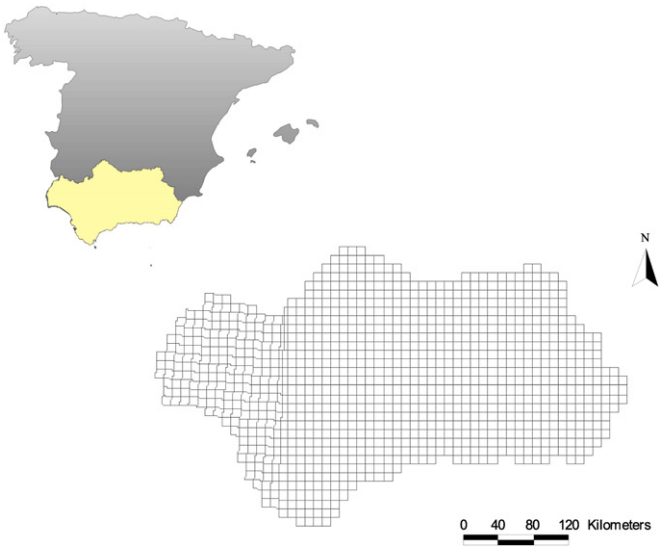



**Fig. 1.** Location of the study area. A 10 × 10 km grid was superimposed to calculate the values of the variables.

with the model. Several advantages are gained from this methodology (Uusitalo, 2007): suitability for incomplete data sets, possibility of structural learning, combination of different sources of knowledge, explicit treatment of uncertainty and support for decision analysis, and fast response. However, most environmental variables are continuous whilst Bayesian networks usually build the model over discrete domains, so that continuous variables need to be first discretized (Uusitalo, 2007). Discretization implies capturing only rough characteristics of the original distribution (Friedman and Goldszmidt, 1996) and loss of statistical information. Thus, there is a need to develop Bayesian networks that can work with continuous values.

The problem of using continuous and discrete variables simultaneously involves more complex mathematical models. There are several techniques in the literature to cope with hybrid variables. Two of the most common are the Conditional Gaussian model (CG) (Castillo et al., 1997; Cowell et al., 1999) and the MTE model (Mixtures of Truncated Exponentials) (Moral et al., 2001). These models avoid discretization of the variables. The Gaussian model imposes certain restrictions when working with discrete and continuous variables (Castillo et al., 1997; Cowell et al., 1999), namely that discrete variables cannot have continuous parents and that continuous variables (given the discrete ones) must have a multivariate Gaussian distribution. Implementation of the MTE model does not restrict either the graph structure or the distribution of the variables. This model has been studied for several years by our research group (Moral et al., 2001, 2002, 2003; Rumí, 2005; Rumí et al., 2006; Rumí and Salmerón, 2007; Cobb et al., 2006, 2007).

Bayesian networks can be used to solve classification problems. The most frequently used structures for this purpose are the naïve

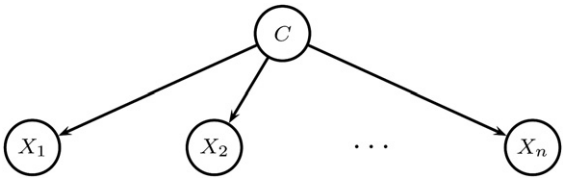



**Fig. 2.** Structure of a naïve Bayes classifier.

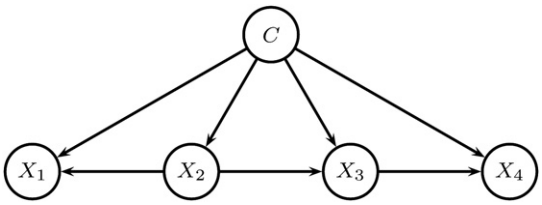



**Fig. 3.** A TAN structure with $X_2$ as root of the maximum spanning tree among the features ($C$: class; $X_i$: feature variables).

Bayes (NB) and the tree augmented naïve Bayes (TAN) (Friedman et al., 1997). These models have usually been applied only to discrete variables. Bayesian classifiers bring significant benefits over traditional statistical techniques. Mainly, accurate information about a target variable can be obtained without requiring complete observation of all the remaining variables.

The aim of this paper is to develop NB and TAN classifier structures based on the MTE model that allows the simultaneous use of continuous and discrete variables in the same network, without any pre-processing in either the variables or the data. Continuous environmental variables and presence/absence records of the spur-thighed tortoise (*Testudo graeca graeca*) were used to develop the models. The results are compared with discrete NB and TAN structures proposed by Aguilera et al. (2007) and used to characterize the habitat of the tortoise.

## 2. Methodology

### 2.1. Variables and data set description

The study area selected (Fig. 1) is located in the region of Andalusia (southern Spain). A set of thematic maps of vegetation and land use, lithology and soils were selected and incorporated into an automatic spatial representation system, ArcGis 9.2. A 10 × 10 km grid was superimposed over the thematic maps to calculate the percentage cover of each variable in each cell. Mean, maximum and minimum, height, slope, temperature and rainfall were also considered.

In this way, a matrix of 988 observations and 176 environmental variables was obtained. The data relating to the presence/absence of the spur-thighed tortoise (*T. graeca graeca*) for each cell were derived from the Atlas of Amphibians and Reptiles of Spain (Pleguezuelos et al., 2002). This tortoise is an endangered species (IUCN, 2009).

### 2.2. Selection of variables

Since the number of variables described in Section 2.1 is excessive, a selection of the most representative ones is needed (Aguilera et al., 2007). This selection can be done in different ways within the framework of Bayesian networks (Bell and Wang, 2000; Ben-Bassat, 1982; Inza et al., 2000; Mladenic, 2006). In this case, filter measures, based on information functions applied to discrete variables (qualitative or quantitative), were used. The selected measure was Kullback–Leibler (Kullback and Leibler, 1951; Kullback, 1959). This method is only defined for discrete variables, in fact there are still no validated methods for feature selection using filter measures in continuous variables. For this reason, the continuous variables were discretized using the $k$-means clustering algorithm. Three groups representing low, medium and high values for each variable were considered. This process was developed using Elvira GUI software (Elvira-Consortium, 2002).

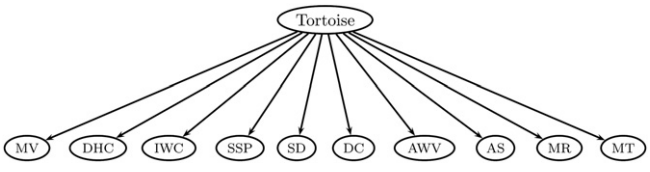



**Fig. 4.** Modelling using an NB structure. MV: marsh with vegetation; DHC: dry herbaceous crops; IWC: irrigated woody crops; SSP: sparse shrubland with pasture; SD: sand and dunes; DC: dry woody crops; AWV: areas with low vegetation; MR: mean rainfall; AS: aridisols; MT: mean temperature.

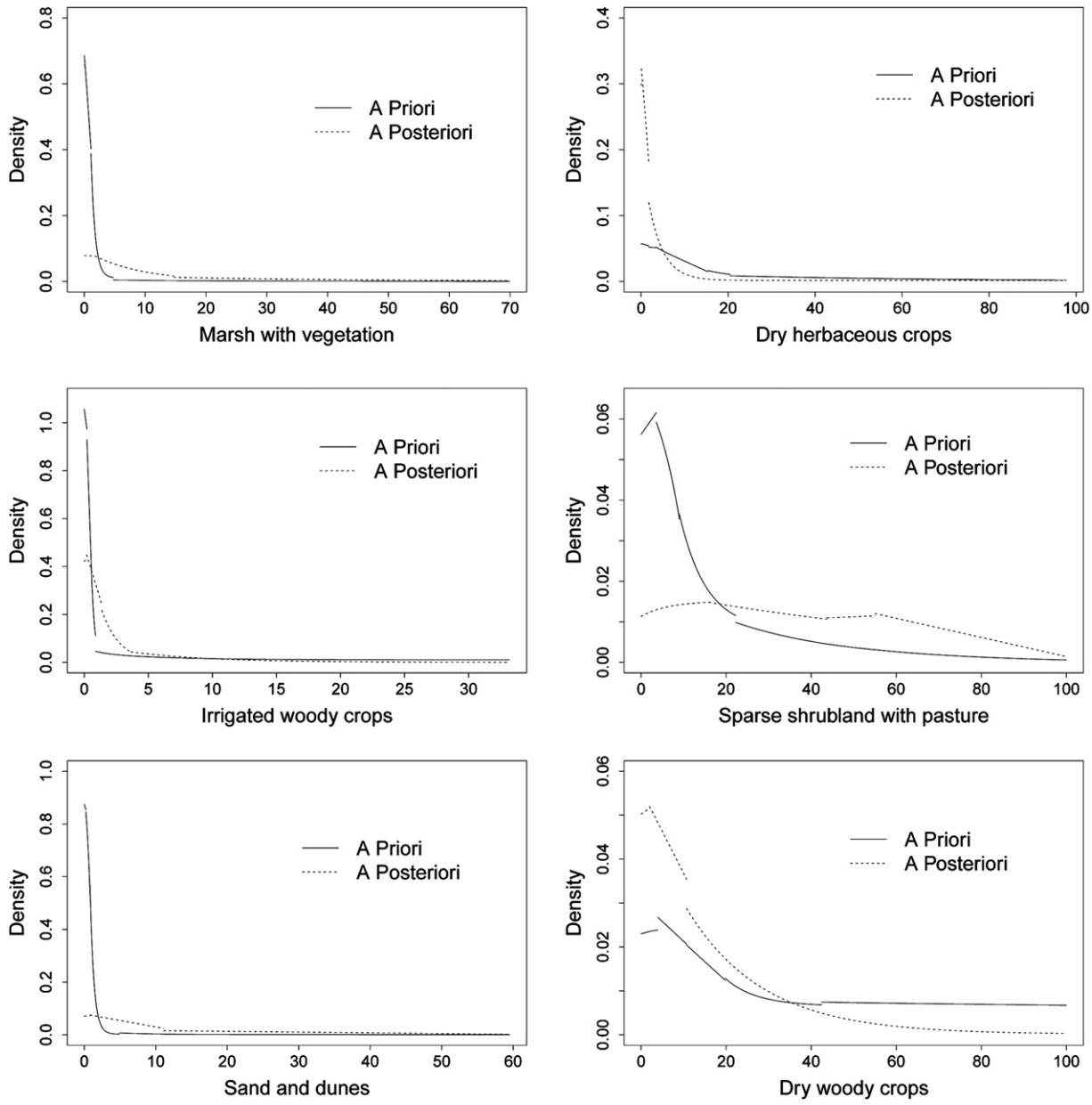



**Fig. 5.** Prior and posterior marginal probability distributions in the NB model for the variables marsh with vegetation, dry herbaceous crops, irrigated woody crops, sparse shrubland with pasture, sand and dunes and dry woody crops.

The discretization of the continuous variables was taken into account only to select the final set of variables in our study. Once obtained, they were treated as continuous variables in order to implement the NB and TAN models.

Once the discretization and Kullback–Leibler filter measure was applied, 10 variables were selected after consulting with an expert. In decreasing order they are: areas with low vegetation (vegetation cover <20%), sparse shrubland with pasture, aridisols soil type, mean rainfall, mean temperature, irrigated woody crops, marsh with vegetation, sand and dunes, dry woody crops and dry herbaceous crops.

### 2.3. Bayesian networks

A Bayesian network is a pair $(\mathcal{G}, P)$ where:

- $\mathcal{G}$ is a directed acyclic graph with a set of vertices **X** and links between them. The vertices represent random variables and the links relationships of dependence or independence located in the network.
- $P$ represents a set of conditional probability functions, one for each variable given its parents, i.e., $P = \{p(x_1|\pi_1), \ldots, p(x_n|\pi_n)\}$, where $\pi_i$ are the parents of $X_i$ in the graph and **x** is a configuration over variables **X**.

Let $(\mathcal{G}, P)$ be a Bayesian network, then the joint probability distribution $p(\mathbf{x})$ is defined by the following factorization:

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i|\pi) x \in \Omega_X. \quad (1)$$

where $\Omega_X$ is the support of variable $X$.

### 2.4. The MTE model

Having discrete and continuous variables simultaneously in the same network requires a model that can handle the network correctly. One of the proposed solutions is the MTE model (Moral et al., 2001).

During the probability inference process, where the posterior distributions of the variables are obtained given some evidence (Langseth et al., 2009), the intermediate functions are not necessarily density functions, therefore a general function called *MTE potential* needs to be defined as follows:

Let **X** be a mixed $n$-dimensional random vector. Let $\mathbf{Y} = (Y_1, \ldots, Y_d)$ and $\mathbf{Z} = (Z_1, \ldots, Z_c)$ be the discrete and continuous parts of **X**, respectively, with $c + d = n$. We say that

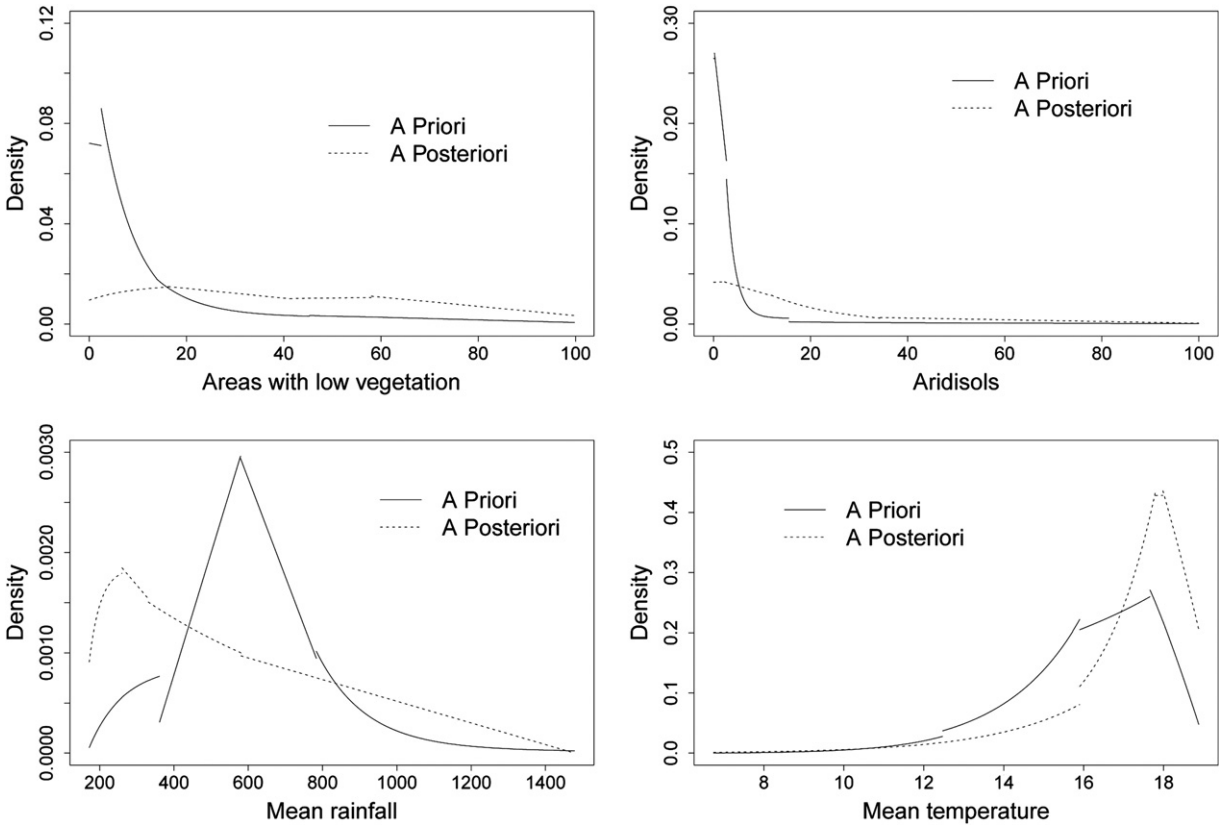



**Fig. 6.** Prior and posterior marginal probability distributions in the NB model for the variables areas with low vegetation, mean rainfall, aridisols and mean temperature.

a function $f : \Omega_{\mathbf{X}} \to \mathbb{R}_0^+$ is a *Mixture of Truncated Exponentials potential (MTE potential)* if one of the following conditions holds:

i. $\mathbf{Y} = \emptyset$ and $f$ can be written as

$$f(\mathbf{x}) = f(\mathbf{z}) = a_0 + \sum_{i=1}^{m} a_i \exp\left\{\sum_{j=1}^{c} b_i^{(j)} z_j\right\} \quad (2)$$

for all $\mathbf{z} \in \Omega_{\mathbf{Z}}$, where $a_i$, $i = 0, \ldots, m$ and $b_i^{(j)}$, $i = 1, \ldots, m, j = 1, \ldots, c$ are real numbers.

ii. $\mathbf{Y} = \emptyset$ and there is a partition $D_1, \ldots, D_k$ of $\Omega_{\mathbf{Z}}$ into hypercubes such that $f$ is defined as

$$f(\mathbf{x}) = f(\mathbf{z}) = f_i(\mathbf{z}) \quad \text{if } \mathbf{z} \in D_i,$$

where each $f_i$, $i = 1, \ldots, k$ can be written in the form of Equation (2).

iii. $\mathbf{Y} \neq \emptyset$ and for each fixed value $\mathbf{y} \in \Omega_{\mathbf{Y}}$, $f_{\mathbf{y}}(\mathbf{z}) = f(\mathbf{y},\mathbf{z})$ can be defined as in ii.

An MTE potential $f$ is an *MTE density* if

$$\sum_{\mathbf{y} \in \Omega_{\mathbf{Y}}} \int_{\Omega_{\mathbf{Z}}} f(\mathbf{y},\mathbf{z}) d\mathbf{z} = 1.$$

A *conditional MTE density* can be specified by dividing the domain of the conditioning variables and specifying an MTE density for the conditioned variable for each configuration of splits of the conditioning variables. Moral et al. (2001) proposed a data structure to represent MTE potentials, which is specially appropriate for this kind of conditional densities: the so-called *mixed probability trees*.

**Table 1**
Expected values of the probability distributions a priori and a posteriori for the NB model. The values represent percentage cover except for mean rainfall (mm) and mean temperature (°C).

| | A priori | A posteriori | % Change |
|---|---|---|---|
| Marsh with vegetation | 3.70 | 15.90 | 329 |
| Dry herbaceous crops | 22.54 | 10.06 | −56 |
| Irrigated woody crops | 6.83 | 3.39 | −50 |
| Sparse shrubland with pasture | 17.56 | 40.00 | 127 |
| Sand and dunes | 2.47 | 15.11 | 510 |
| Dry woody crops | 38.02 | 17.18 | −55 |
| Areas with low vegetation | 16.33 | 42.03 | 157 |
| Aridisols | 6.85 | 23.11 | 237 |
| Mean rainfall | 608.39 | 586.71 | −3.43 |
| Mean temperature | 16.10 | 16.92 | 5 |

### 2.5. Bayesian classifiers and calibration of models

A Bayesian network can be used for classification purposes if it contains a class variable, $C$, and a set of feature variables $X_1, \ldots, X_n$, where an object with observed features $x_1, \ldots, x_n$ will be classified as belonging to class $c^*$ obtained as

$$c^* = \text{argmax}_{c \in \Omega_C} p(c|x_1, \ldots x_n), \quad (3)$$

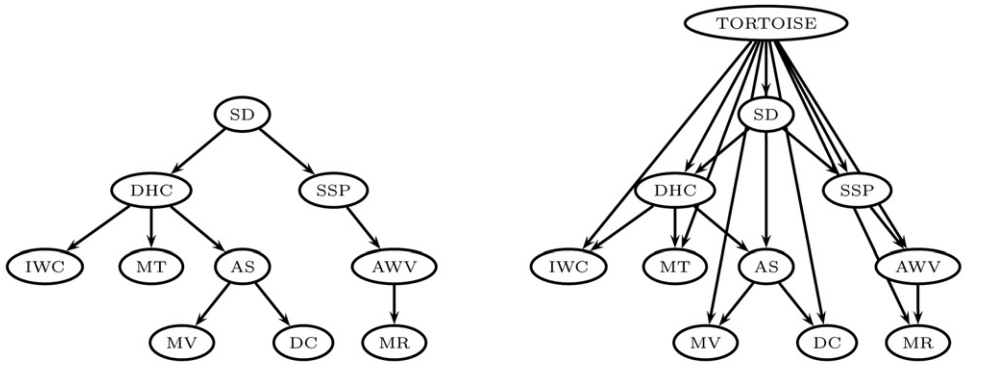
where $\Omega_C$ denotes the set of possible values of $C$.

Note that $p(c|x_1, \ldots, x_n)$ is proportional to $p(c) \times p(x_1, \ldots, x_n|c)$, and therefore, solving the classification problem would require a distribution to be specified over the $n$ feature variables for each value of the class. The associated computational cost can be very high. However, using the factorization determined by the network, the cost is reduced. Although the ideal would be to build a network without restrictions on the structure, this is not possible due to the limited data available. Therefore, networks with fixed and simple structures and specifically designed for classification have been used. The extreme case is the so-called *naïve Bayes (NB)* model (Duda et al., 2001).

This is based on the simplest structure from a computational point of view: therefore it has become the most widely used Bayesian classifier in the literature. Its name comes from the naive assumption that the feature variables are conditionally independent given the class variable.

The hypothesis of independence assumed by this model leads to a Bayesian network with a single root node and a set of attributes having only one parent (the root node). The NB model structure is shown in Fig. 2.

The strong assumption of independence underlying this model is compensated by the reduction in the number of parameters to be estimated, since in this case it holds that




**a** Tree showing the relations among the features.

**b** TAN model constructed adding to the previous tree the class variable and links from it to each feature.

**Fig. 7.** Sequence followed to obtain the TAN model. MV: marsh with vegetation; DHC: dry herbaceous crops; IWC: irrigated woody crops; SSP: sparse shrubland with pasture; SD: sand and dunes; DC: dry woody crops; AWV: areas with low vegetation; MR: mean rainfall; AS: aridisols; MT: mean temperature. (a) Tree showing the relations among the features. (b) TAN model constructed adding to the previous tree the class variable and links from it to each feature.

$$p(c|x_1,\ldots x_n) = p(c)\prod_{i=1}^{n} p(x|c), \tag{4}$$

and thus, instead of one $n$-dimensional conditional distribution, we have $n$ one-dimensional conditional distributions.

In Algorithm 1 (see Appendix) the steps for constructing an NB classifier with continuous features are shown. In essence, they consist of building a Bayesian network with an NB structure and estimating the marginal and conditional MTE densities for the variables (Moral et al., 2002).

The NB model assumes independence of the feature variables given the class. Despite this strong assumption, the results are amazing in many cases (Fernández et al., 2007, 2009; Fernández and Salmerón, 2008). However, some variables are highly correlated (Aguilera et al., 2007) and the accuracy of classification would improve if any dependence between them could be included in the network.

The tree augmented naïve Bayes (TAN) structure (Friedman et al., 1997) incorporates the following restriction: every feature variable except one must have the class variable and another feature as parents. The model is richer, since it allows arcs among features, but an increase of complexity is assumed instead, both in the learning of the graph structure and the associated probabilities. Fig. 3 shows an example of a TAN model with four features and one class variable.

The steps to build the TAN classifier with continuous features are shown in Algorithm 2 (see Appendix).

Elvira API software[1] (Elvira-Consortium, 2002) was used both for the learning of the models and their validation. It is remarkable that this is the only software in the literature dealing with hybrid Bayesian networks using MTEs.

### 2.6. Inference in Bayesian classifiers

The goal in this section is to determine all the probabilistic information, both a priori and a posteriori of the model. Thus, the model can be used to give a true reflection of the initial data set (model a priori) or to predict the impact (in terms of probability) of introducing evidence for certain variables (model a posteriori).

For example, if the evidence (the observation that the tortoise is present), $P$ (tortoise = presence) = 1, is set in the class variable, the density functions of the remaining environmental variables will be modified. In this way, an approximation to the most probable configuration for the presence of the tortoise can be obtained.

### 2.7. Validation of the models

The models were tested using $k$-fold cross validation (Stone, 1974). This technique is applied to the initial data set and used to evaluate the quality of a classification model.

A lazy choice would be to use holdout validation ($k = 1$). It is not considered cross validation as such, since the data never cross. The initial data set is randomly divided into two subsets: the first one ($D_a$) is devoted to the training phase of the model and the second one ($D_p$), to validating it. Usually, less than a third of the initial data set is used for $D_p$.

For a $k$-value greater than 1, the data set is split into $k$ subsets. In each step, one subset is assigned to $D_p$ and the remaining $k-1$ to $D_a$. Cross validation is repeated $k$ times, each time taking a different subset for $D_p$. This is the approach followed to test the classifiers presented in this paper.

The output model is constructed by including the entire database in $D_a$.

[1] It can be downloaded from http://leo.ugr.es/elvira.

## 3. Results and discussion

### 3.1. NB model

The resulting NB model is shown in Fig. 4. The introduction of the evidence "presence of tortoise", changes the probability distribution of the features because they are directly connected with the tortoise variable.

Figs. 5 and 6 show the density function for each variable, both without evidence of the tortoise being present (a priori) and with evidence (a posteriori). Table 1 shows the expected values of the marginal density function for each environmental variable both a priori and a posteriori.

The marginal probability distributions of the NB model show that it is likely to find the tortoise in areas with sand and dunes, marsh with vegetation, aridisols, areas with low vegetation and in sparse shrubland with pasture. The remaining variables vary by less than 100%, between the case of evidence to no evidence. The model also suggests that is likely to find the tortoise where mean rainfall is lower and where mean temperature is slightly higher.

### 3.2. TAN model

Fig. 7 shows the constructed TAN model. The main difference with respect to the NB is the relationship between the features. This increases the number of arcs in the structure and its complexity, but improves the accuracy and expressivity.

Fig. 7b shows the structure of the corresponding TAN model built from the tree shown in Fig. 7a. The procedure consists of adding the tortoise variable and drawing an arc from it to each environmental variable.

Figs. 8 and 9 show the density functions for each variable in case of there being no evidence of the tortoise (a priori) and with evidence (a posteriori). Table 2 shows the expected values of the

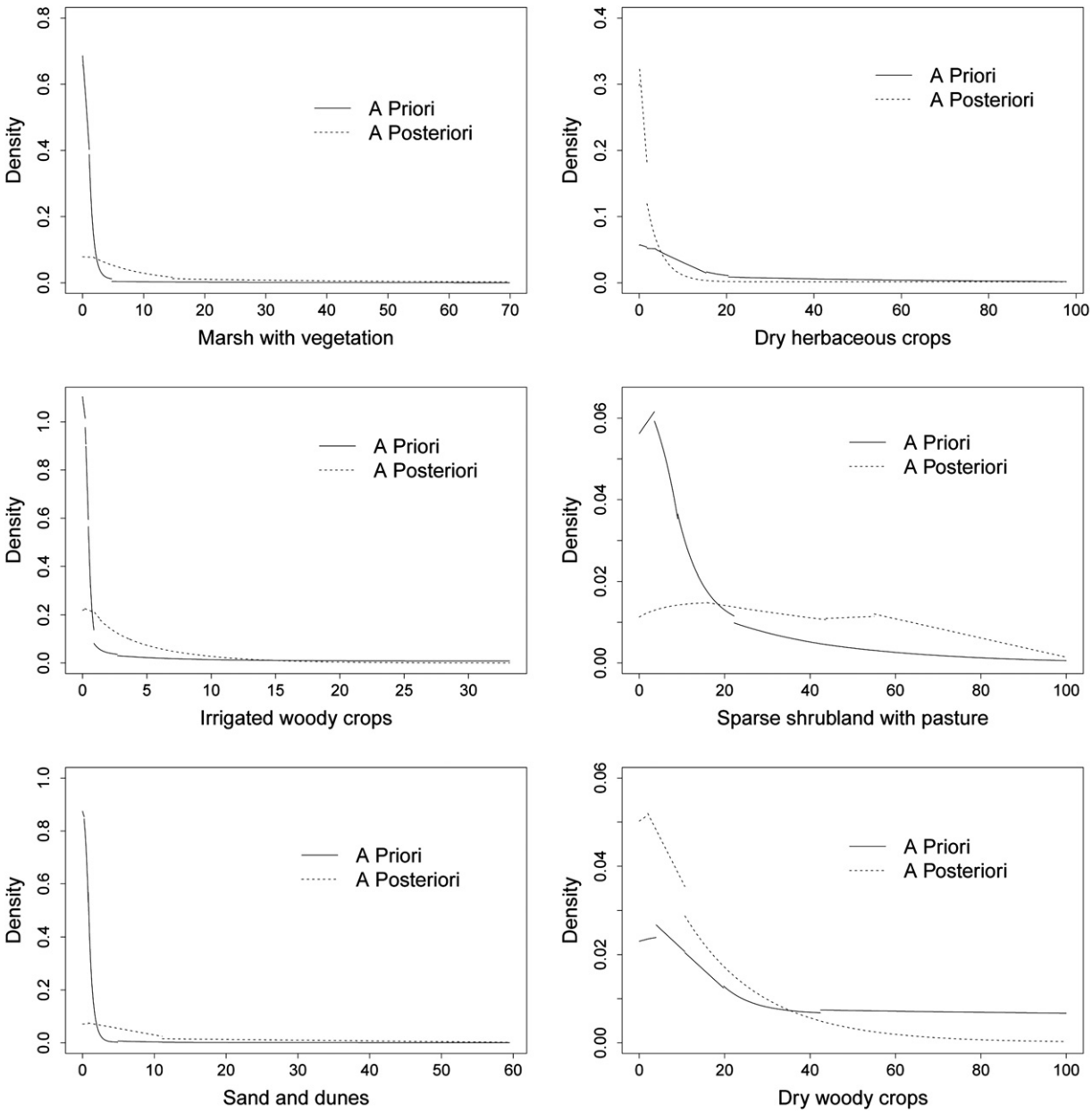



**Fig. 8.** Prior and posterior marginal probability distributions in the TAN model for marsh with vegetation, dry herbaceous crops, irrigated woody crops, sparse shrubland with pasture, sand and dunes and dry woody crops.

prior and posterior density function for each environmental variable.

Marsh with vegetation, dry herbaceous crops, sparse shrubland with pasture, sand and dunes, and dry woody crops show prior and posterior marginal probability functions similar to the NB model.

Thus, NB and TAN models show similar distributions both a priori and a posteriori, but the quantification varies (Tables 1 and 2). They vary in the definition of relationships between the features in the TAN model, so that each variable is influenced, not only by the class variable tortoise, but also by the variables directly connected with it in the network. Five probability distributions differ between TAN and NB, so the habitat description is slightly different: mean precipitation decreases by 33% (from 586.71 mm in NB to 392.56 mm in TAN), cover of irrigated woody crops increases by 20% (from 3.39% in NB to 4.08% in TAN), areas with low vegetation increases by 12% (from 42.03% in NB to 47.12% in TAN) and aridisols increases by 15.2% (from 23.11% in NB to 26.62% in TAN).

Table 2 identifies the representative variables related to the presence of the tortoise. In descending order, they are: sand and dunes, marsh with vegetation, areas with low vegetation and sparse shrubland with pasture. For these variables, evidence of the tortoise being present implies an increase of more than 100% in their mean cover.

Aridisols increases by only 72%. Mean rainfall and mean temperature are important climatic variables in the habitat characterization, and indicate that the tortoise's habitat has a lower mean rainfall and a higher mean temperature.

### 3.3. Validation

Table 3 shows the classification rate for the discrete (Aguilera et al., 2007) and the continuous models, using 10-fold cross validation. A classification rate grouping by NB, TAN, continuous and

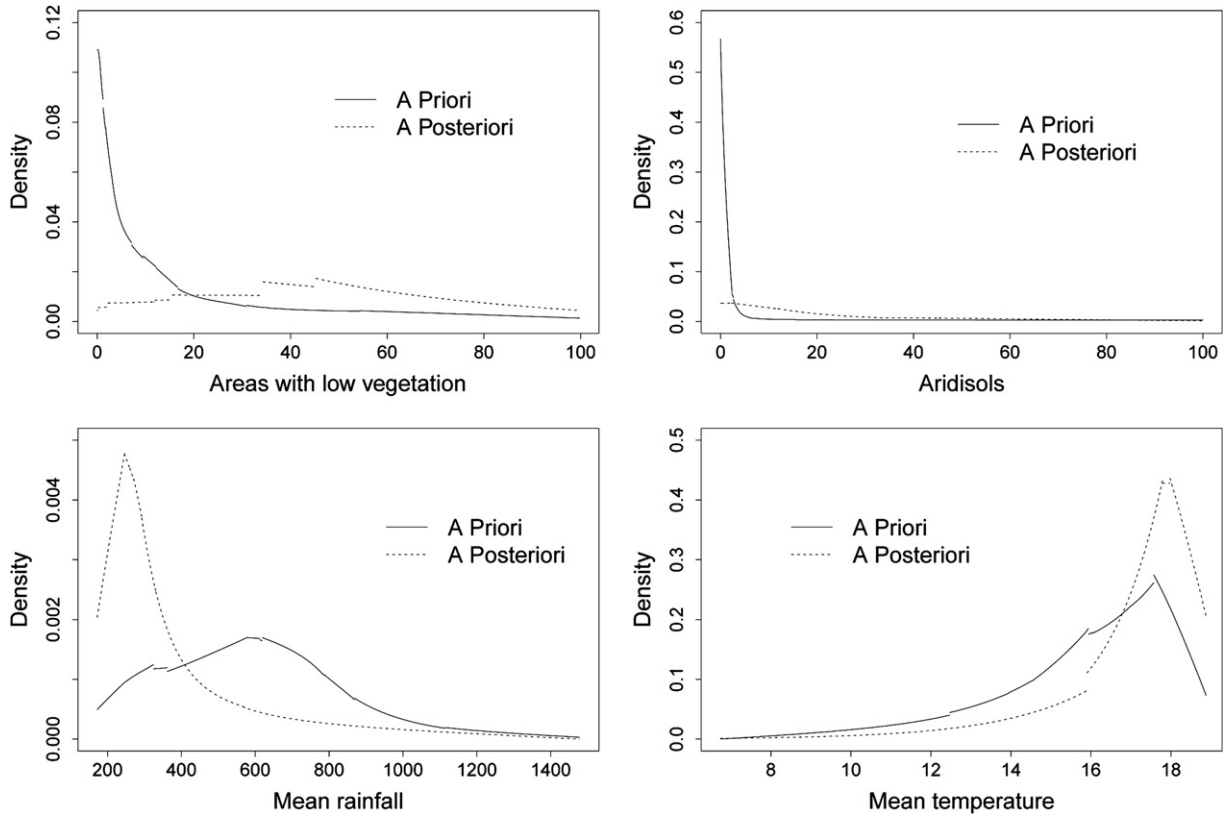



**Fig. 9.** Prior and posterior marginal probability distributions in the TAN model for areas with low vegetation, mean rainfall, aridisols and mean temperature.

discrete models, as well as the standard deviation for each value are also shown.

Fig. 10 shows two box plots. The first one represents the values of the classification rates for the continuous and discrete models. The second one shows the same values for the TAN and NB models.

After applying Lilliefors' test to check the normality of the data, the *t*-test to compare the experimental results was applied (see Table 4). There are significant differences between continuous and discrete models (*p*-value of 0.0021, $p < 0.05$). This difference is due to the loss of statistical information in the discretization process. On the other hand, there are no significant differences between TAN and NB models (*p*-value of 0.2531, $p > 0.05$). In general, it is demonstrated that TAN models are better for classification than NB models, but with scarce data (our case) the MTE learning process may cause a worse classification rate. This fact can modify slightly the results of TAN. In any case it seems from Fig. 10 that TAN outperforms NB but not statistically.

### 3.4. Spatial application of the models

Fig. 11a and b shows the probability of the tortoise being present in Andalusia according to the discrete models developed by Aguilera et al. (2007). The same is shown in Fig. 12a and b for the continuous models developed in this paper.

Figs. 11a,b and 12a,b clearly indicate the existence of two populations of tortoise in Andalusia: one located in the southwest and another in the southeast. The discrete models NB and TAN recognize this pattern, but show a more dispersed distribution in the region, locating the presence of tortoises in less likely inland habitats. The continuous NB model shows a better characterization of the habitat, however it includes an area close to the Strait of Gibraltar, determined by higher precipitation (mean value of 586.71 mm) with respect to the continuous TAN.

The continuous TAN model corresponds exactly to the presence of the tortoise in Andalusia. The probability distributions determined by this model characterize both habitats. In the southwest, the tortoises occur in areas of sandy substrate alternating with vegetation near to marshes. These environmental variables

**Table 2**
Expected values of the probability distributions a priori and a posteriori for the TAN model. The values represent percentage cover except for mean rainfall (mm) and mean temperature (°C).

| | A priori | A posteriori | % Change |
|---|---|---|---|
| Marsh with vegetation | 3.70 | 15.90 | 329 |
| Dry herbaceous crops | 22.54 | 10.06 | −56 |
| Irrigated woody crops | 6.16 | 4.08 | −34 |
| Sparse shrubland with pastures | 17.56 | 40.00 | 127 |
| Sand and dunes | 2.47 | 15.11 | 510 |
| Dry woody crops | 38.02 | 17.18 | −55 |
| Areas with low vegetation | 20.70 | 47.12 | 228 |
| Aridisols | 15.50 | 26.62 | 72 |
| Mean rainfall | 600.81 | 392.56 | −35 |
| Mean temperature | 15.85 | 16.92 | 7 |

**Table 3**
10-Fold cross validation for the discrete and continuous version of NB and TAN.

| | Classification rate | Standard deviation |
|---|---|---|
| Discrete NB | 0.9362 | 0.0165 |
| Continuous NB | 0.9707 | 0.0074 |
| Discrete TAN | 0.9493 | 0.0519 |
| Continuous TAN | 0.9707 | 0.0074 |
| NB models | 0.9535 | 0.0165 |
| TAN models | 0.9600 | 0.0377 |
| Continuous models | 0.9707 | 0.0072 |
| Discrete models | 0.9428 | 0.0381 |

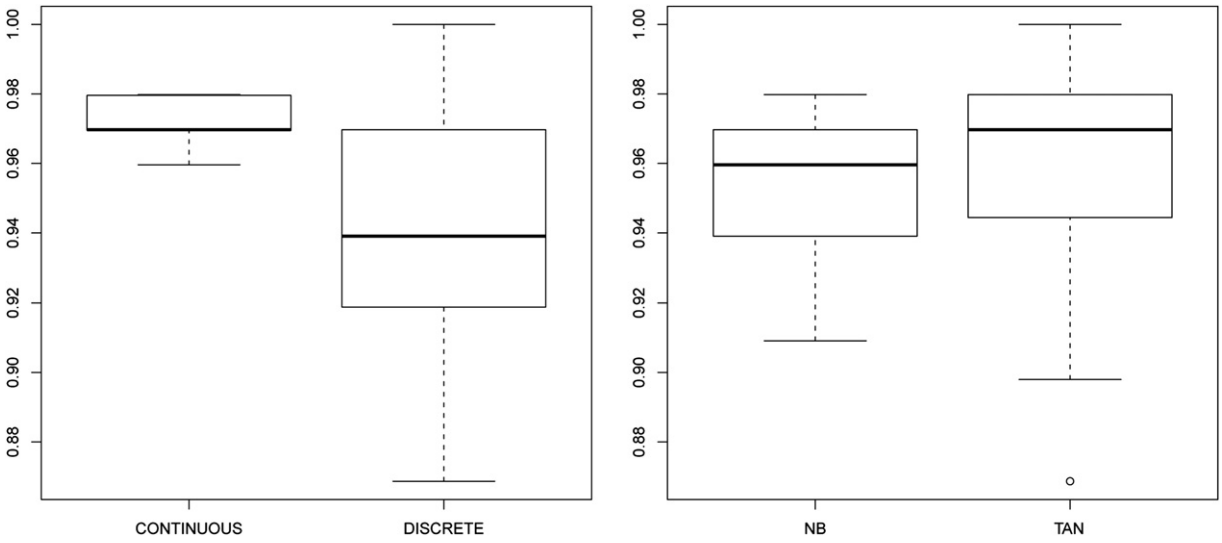
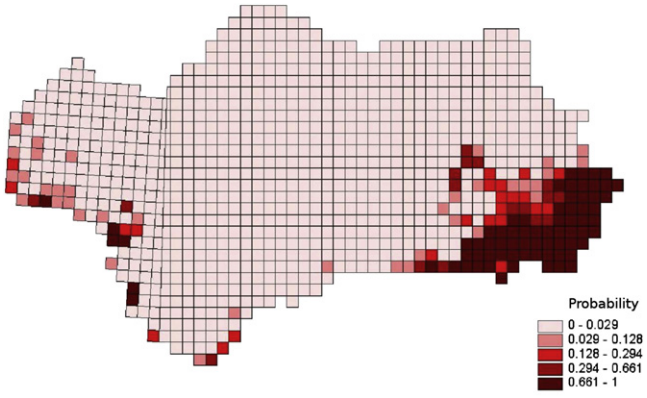
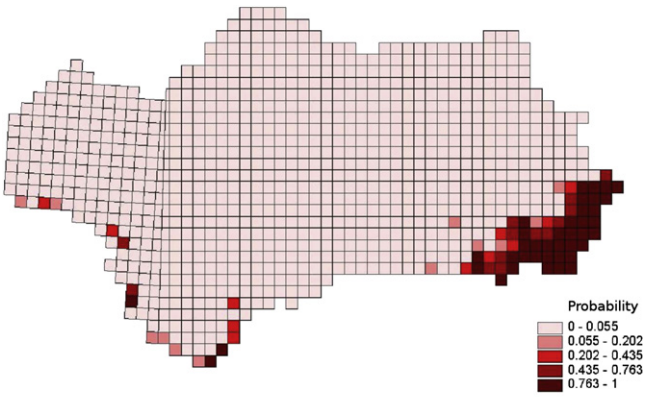


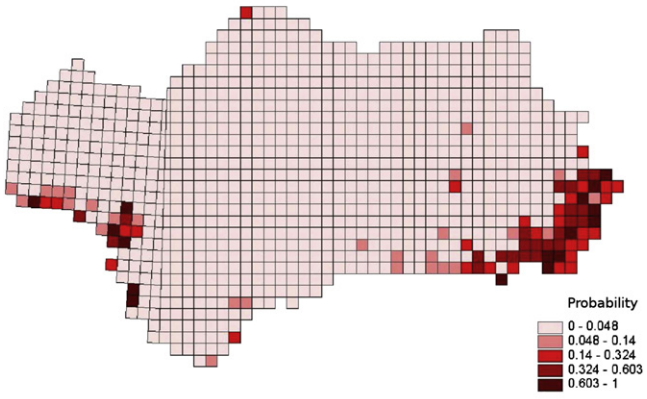

**Fig. 10.** Box plots comparing the classification rate for continuous against discrete models and NB against TAN models.



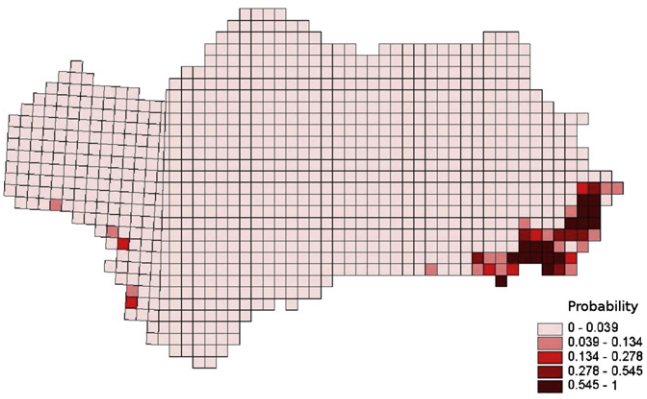

**a** Discrete NB




**a** Continuous NB




**b** Discrete TAN




**b** Continuous TAN

**Fig. 11.** Probability of presence of the tortoise in the region of Andalusia according to the discrete models (Aguilera et al., 2007).

**Fig. 12.** Probability of presence of the tortoise in the region of Andalusia according to the continuous models.

**Table 4**
Statistical differences between the models.

| | *p*-Value |
|---|---|
| Continuous – discrete | 0.0021 |
| TAN – NB | 0.2531 |

correspond spatially to the Doñana National Park. In the southeast the habitat is semiarid, with sparse shrubland with pasture, areas with low vegetation and an abundance of aridisol soil types. The model shows that the most important factors in the distribution of tortoises in the southeast are climate and vegetation type.

The results obtained in the characterization of tortoise habitat indicate that NB and TAN continuous models based on Mixtures of Truncated Exponentials (MTEs) can be applied to species distribution modeling, by allowing the simultaneous use of both discrete and continuous variables in the development of the models.

## Acknowledgement

This work has been supported by the Spanish Ministry of Science and Innovation through project TIN2007-67418-C03-02, by FEDER funds, and by the Junta de Andalucía through project P05-TIC-00276.

## Appendix. Algorithms

**Algorithm 1: naïve Bayes classifier with continuous features**

**Input**: A database $D$ with variables $X_1, \ldots, X_n, Y$.
**Output**: An NB model with root variable $C$ and features $X_1, \ldots, X_n$, with joint distribution of class MTE.
1. Construct a new network $\mathcal{G}$ with nodes $C, X_1, \ldots, X_n$.
2. Insert the links $C \rightarrow X_i$, $i = 1, \ldots, n$ en $\mathcal{G}$.
3. Estimate an MTE density for $C$, and a conditional MTE density for each $X_i$, $i = 1, \ldots, n$ given its parents in $\mathcal{G}$ (Rumí et al., 2006; Moral et al., 2003; Romero et al., 2006).
4. Let $P$ be the set of estimated densities.
5. Let NB be a Bayesian network with structure $\mathcal{G}$ and distributions $P$.
6. Return NB.

**Algorithm 2: TAN classifier with continuous features**

**Input:** A database $D$ with variables $X_1, \ldots, X_n, C$.
**Output**: A TAN model with root variable $C$ and features $X_1, \ldots, X_n$, with joint distribution of class MTE.
1. Construct a complete graph $\mathcal{C}$ with nodes $X_1, \ldots, X_n$.
2. Label each link $(X_i, X_j)$ with the conditional mutual information between $X_i$ and $X_j$ given $C$ (Fernández et al., 2007), i.e., $I(X_i, X_j|C) = \int\int\int\int f(x_i, x_j, c)\log(f(x_i, x_j|c)/f(x_i|c)f(x_j|c))\mathrm{d}x_i\mathrm{d}x_j\mathrm{d}c$.
3. Let $\mathcal{T}$ be the maximum spanning tree obtained from $\mathcal{C}$ using the algorithm 3.
4. Directs the links in $\mathcal{T}$ in such a way that no node has more than one parent.
5. Construct a new network $\mathcal{G}$ with nodes $C, X_1, \ldots, X_n$ and the same links as $\mathcal{T}$.
6. Insert the links $C \rightarrow X_i$, $i = 1, \ldots, n$ in $\mathcal{G}$.
7. Estimate an MTE density for $C$, and a conditional MTE density for each $X_i$, $i = 1, \ldots, n$ given its parents in $\mathcal{G}$ (Rumí et al., 2006; Moral et al., 2003; Romero et al., 2006).
8. Let $P$ be the set of estimated densities.
9. Let TAN be a Bayesian network with structure $\mathcal{G}$ and distributions $P$.
10. Return TAN.

**Algorithm 3: Maximum Spanning Tree (based on Kruskal's algorithm)**

**Input:** A graph $\mathcal{G} = (V, E)$, in which $V$ is the set of vertices and $E$ is the set of links.
**Output**: Maximum Spanning Tree $\mathcal{T}$.
1. Order the links of $E$ in decreasing order using its weight.
2. Let $A$ be a set of links empty initially.
3. $\mathcal{T} \leftarrow (V, A)$
4. **for** $i \rightarrow 0$ to $n - 2$ **do**
5. Add $i$-th link $(u,v) \in E$ to the set $A$, iff it doesn't cause a cycle in $\mathcal{T}$.
6. **end**
7. Return $\mathcal{T}$.

## References

Aguilera, P., Reche, F., López, E., Willaarts, B., Castro, A., Schmitz, M., 2007. Aplicación de las redes bayesianas a la caracterización del hábitat de la tortuga mora (*Testudo graeca graeca*) en Andalucía. In: Proceedings of the I Congreso Nacional de Biodiversidad.

Anderson, R., Peterson, A., Gómez-Laverde, M., 2002. Using niche-based GIS modelling to test geographic predictions of competitive exclusion and competitive release in South American pocket mice. Oikos 98, 3–16.

Araújo, M., Cabeza, M., Thuiller, W., Hannah, L., Williams, P., 2004. Would climate change drive species out of reserves? An assessment of existing reserve-selection models. Global Change Biology 10, 1618–1626.

Araújo, M., Pearson, W.T.R., 2006. Climate warming and the decline of amphibians and reptiles in Europe. Journal of Biogeography 33, 1712–1728.

Austin, M., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecological Modelling 157, 101–118.

Bell, D., Wang, H., 2000. A formalism for relevance and its application in feature subset selection. Machine Learning 41 (2), 175–195.

Ben-Bassat, M., 1982. Use of distance measures, information measures and error bounds in features evaluation. HandBook of Statistics 2, 773–791.

Borsuk, M., Reichert, P., Peter, A., Schager, E., Burkhardt-Holm, P., 2006. Assessing the decline of brown trout (*Salmo trutta*) in swiss rivers using Bayesian probability network. Ecological Modelling 192, 224–244.

Borsuk, M., Stow, C., Reckhow, K., 2004. A Bayesian network of eutrophication models for synthesis, prediction, and uncertainty analysis. Ecological Modelling 173, 219–239.

Bromley, J., Jackson, N., Clymer, O., Giacomello, A., Jensen, F., 2005. The use of Hugin to develop Bayesian networks as aid to integrated water resource planning. Environmental Modelling and Software 20, 231–242.

Brotons, L., Thuiller, W., Araujo, M., Hirzel, A., 2004. Presence–absence versus presence only modelling methods for predicting bird habitat suitability. Ecography 27, 437–448.

Burgmann, M., Lindenmayer, D., Elith, J., 2005. Managing landscapes for conservation under uncertainty. Ecology 86, 2007–2017.

Castillo, E., Gutiérrez, J., Hadi, A., 1997. Expert Systems and Probabilistic Network Models. Springer-Verlag.

Cobb, B., Shenoy, P., Rumí, R., 2006. Approximating probability density functions with mixtures of truncated exponentials. Statistics and Computing 16, 293–308.

Cobb, B.R., Rumí, R., Salmerón, A., 2007. Bayesian networks models with discrete and continuous variables. In: Advances in Probabilistic Graphical Models. Studies in Fuzziness and Soft Computing. Springer 81–102.

Cowell, R., Dawid, A., Lauritzen, S., Spiegelhalter, D., 1999. Probabilistic Networks and Expert Systems. In: Statistics for Engineering and Information Science. Springer.

Dedecker, A., Goethals, P., Gabriels, W., De Pauw, N., 2004. Optimization of Artificial Neural Network (ANN) model design for prediction of macroinvertebrates communities in the Zwalm river basin (Flanders, Belgium). Ecological Modelling 174, 161–173.

Duda, R., Hart, P., Stork, D., 2001. Pattern Classification. Wiley Interscience.

Dżeroski, S., Drumm, D., 2003. Using regression trees to identify the habitat preference of the sea cucumber (*Holothuria leucospilota*) on Rarotonga, Cook Islands. Ecological Modelling 170, 219–226.

Elith, J., Graham, C., Anderson, R., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R., Huettmann, F., Leathwick, J., Li, J., Lohmann, L., Loiselle, B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J., Peterson, A., Phillips, S., Richardson, K., Scachetti-Pereria, S., Schapire, R., Soberón, J., Williams, S., Wisz, M., Zimmermann, N., 2006. Novel methods to improve prediction of species' distribution from occurrence data. Ecography 29, 129–151.

Elvira-Consortium, 2002. Elvira: an environment for probabilistic graphical models. In: Proceedings of the First European Workshop on Probabilistic Graphical Models (PGM'02), pp. 222–230.

Fernández, A., Morales, M., Salmerón, A., 2007. Tree augmented naive Bayes for regression using mixtures of truncated exponentials: applications to higher education management. In: Lecture Notes in Computer Science, vol. 4723, IDA'07, pp. 59–69.

Fernández, A., Salmerón, A., 2008. Extension of Bayesian network classifiers to regression problems. In: Geffner, H., Prada, R., Alexandre, I.M., David, N. (Eds.), Advances in Artificial Intelligence – IBERAMIA 2008. Lecture Notes in Artificial Intelligence, vol. 5290. Springer, pp. 83–92.

Fernández, A., Salmerón, A., Nielsen, J.D., 2009. Learning Bayesian networks for regression from incomplete databases. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems 18 (1), 69–86.

Ferrier, S., 2002. Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? Systematic Biology 51, 331–363.

Friedman, Goldszmidt, M., 1996. Discretizing continuous attributes while learning Bayesian networks. In: Proceedings of the 13th International Conference on Machine Learning (ICML). Morgan Kaufmann Publishers, pp. 157–165.

Friedman, N., Geiger, D., Goldszmidt, M., 1997. Bayesian network classifiers. Machine Learning 29, 131–163.

Graham, C., Ferrier, S., Huettman, F., Moritz, C., Peterson, A., 2004a. New developments in museum-based informatics and applications in biodiversity analysis. Trends in Ecology and Evolution 19, 497–503.

Graham, C., Ron, S., Santos, J., Schneider, C., Moritz, C., 2004b. Integrating phylogenetics and environmental niche models to explore speciation mechanisms in dendrobatid frogs. Evolution 58, 1781–1793.
Guisan, A., Broennimann, O., Engler, R., Vust, M., Toccoz, N., Lehmann, A., Zimmermann, N., 2006. Using niche-based models to improve the sampling of rare species. Conservation Biology 20 (2), 501–511.
Guisan, A., Gueiss, S., Gueiss, A., 1999. GLM versus CCA spatial modeling of plant species distribution. Plant Ecology 143, 107–122.
Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitats models. Ecology Letters 8, 993–1009.
Guisan, A., Zimmermann, N., 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135, 147–186.
Inza, I., Larrañaga, P., Etxeberria, R., Sierra, B., 2000. Feature subselection by Bayesian networks based optimization. Artificial Intelligence 123 (1–2), 157–184.
IUCN, 2009. Red list of Threatened Species (Version 2009.1). www.iucnredlist.org.
Jensen, F.V., Nielsen, T.D., 2007. Bayesian Networks and Decision Graphs. Springer.
Kullback, S., 1959. Information Theory and Statistics. John Wiley & Son.
Kullback, S., Leibler, R., 1951. On information and sufficency. Annals of Mathematical Statistics 22, 79–86.
Langseth, H., Nielsen, T.D., Rumí, R., Salmerón, A., 2009. Inference in hybrid Bayesian networks. Reliability Engineering and Systems Safety 94, 1499–1509.
Lehmann, A., Overton, J.M., Austin, M., 2002a. Regression models for spatial prediction: their role for biodiversity and conservation. Biodiversity and Conservation 11, 2085–2092.
Lehmann, A., Overton, J.M., Leathwick, J., 2002b. Grasp: generalized regression analysis and spatial prediction. Ecological Modelling 160, 165–183.
Luoto, M., Pöyri, J., Heikkinen, R., Saarinen, K., 2005. Uncertainty of bioclimate envelope models based on the geographical distribution of species. Global Ecology and Biogeography 14, 575–584.
Maggini, R., Lehmann, A., Zimmermann, E., Guisan, A., 2009. Improving generalized regression analysis for the spatial prediction of forest communities. Journal of Biogeography 33, 1729–1749.
Manel, S., Williams, H.C., Ormerod, S., 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. Journal of Applied Ecology 38, 921–931.
Midgley, G., Hannah, L., Millar, D., Thuiller, W., Booth, A., 2003. Developing regional and species-level assessments of climate change impacts on biodiversity in the Cape floristic region. Biological Conservation 112, 87–97.
Miller, J., Franklin, J., 2002. Modelling distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence. Ecological Modelling 157, 227–247.
Mladenic, D., 2006. Feature selection for dimensionality reduction. In: Benferhat, S., Besnard, P. (Eds.), Subspace, Latent Structure and Feature Selection. Lecture Notes in Computer Science, vol. 3940. Springer, pp. 84–102.
Moisen, G., Frescino, T., 2002. Comparing five modeling techniques for predicting forest characteristics. Ecological Modelling 157, 209–225.
Moral, S., Rumí, R., Salmerón, A., 2001. Mixtures of truncated exponentials in hybrid Bayesian networks. ECSQARU'01. Lecture Notes in Artificial Intelligence 2143, 135–143.
Moral, S., Rumí, R., Salmerón, A., 2002. Estimating mixtures of truncated exponentials from data. In: Gámez, J., Salmerón, A. (Eds.), Proceedings of the First European Workshop on Probabilistic Graphical Models, pp. 156–167.
Moral, S., Rumí, R., Salmerón, A., 2003. Approximating conditional MTE distributions by means of mixed trees. In: Symbolic and Quantitative Approaches to Reasoning with Uncertainty. Lecture Notes in Artificial Intelligence, vol. 2711. Springer 197–183.
Pearson, R., Thuiller, W., Araújo, M., Martínez-Meyer, E., Brotons, L., McClean, C., Dawson, L.M.P.S.T., Lees, D., 2006. Model-based uncertainty in species range prediction. Journal of Biogeography 33, 1704–1711.
Peterson, A., 2003. Predicting the geography of species' invasions via ecological niche modelling. The Quarterly Review of Biology 78, 419–433.
Peterson, A., Ortega-Huerta, M., Bartley, J., Sánchez-Cordero, V., Buddmeier, J.S.R., Stockwell, D., 2002. Future projections for Mexican fauna under global climate change scenarios. Nature 416, 626–629.
Pleguezuelos, J., Márquez, R., Lizana, M. (Eds.), 2002. Atlas y libro rojo de los anfibios y reptiles de España, 2second Edition. Dirección General de la Conservación de la Naturaleza–Asociación Herpetológica Española, Madrid (in Spanish).
Pollino, C., White, A., Hart, B., 2007. Examination of conflicts and improved strategies for the management of an endangered eucalyp species using Bayesian networks. Ecological Modelling 201, 37–59.
Romero, V., Rumí, R., Salmerón, A., 2006. Learning hybrid Bayesian networks using mixtures of truncated exponentials. International Journal of Approximate Reasoning 42, 54–68.
Rumí, R., 2005. Kernel methods in Bayesian networks. In: Proceedings of the 1st International Mediterranean Congress of Mathematics, pp. 135–149.
Rumí, R., Salmerón, A., 2007. Approximate probability propagation with mixtures of truncated exponentials. International Journal of Approximate Reasoning 45, 191–210.
Rumí, R., Salmerón, A., Moral, S., 2006. Estimating Mixtures of Truncated Exponentials in Hybrid Bayesian networks. TEST 15 (2), 397–421.
Segurado, P., Araújo, M., 2004. An evaluation of methods for modelling species distribution. Journal of Biogeography 31, 1555–1568.
Smith, C., Howes, A., Price, B., McAlpine, C., 2007. Using Bayesian belief network to predict suitable habitat of an endangered mammal – the Julia Creek dunnart (*Sminthopsis douglasi*). Biological Conservation 139, 333–347.
Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society: Series B (Methodological) 36 (2), 111–147.
Thuiller, W., 2004. Patterns and uncertainties of species' range shifts under climate change. Global Change Biology 10, 2020–2027.
Uusitalo, L., 2007. Advantages and challenges of Bayesian networks in environmental modelling. Ecological Modelling 203, 312–318.
Wintle, B.A., Elith, J., Potts, J., 2005. Fauna habitat modelling and mapping: a review and case study in the Lower Hunter Central Coast region of NSW. Austral Ecology 30, 719–738.
Zaffalon, M., 2005. Credible classification for environmental problems. Environmental Modelling and Software 20 (8), 1003–1012.