Modeling zero-inflated explanatory variables in hybrid Bayesian network classifiers for species occurrence prediction

Ana D. Maldonado, Pedro A. Aguilera, Antonio Salmerón

Published in:
Environmental Modelling & Software

DOI (link to publication from publisher):
https://doi.org/10.1016/j.envsoft.2016.04.003

Publication date:
2016

Document version:
Accepted author manuscript, peer reviewed version

Citation for published version:
Maldonado, A. D., Aguilera, P. A., & Salmerón, A. (2016). Modeling zero-inflated explanatory variables in hybrid Bayesian network classifiers for species occurrence prediction. *Environmental Modelling & Software*, *82*, 31-43.
https://doi.org/10.1016/j.envsoft.2016.04.003

Modeling zero-inflated explanatory variables in hybrid Bayesian network classifiers for species occurrence prediction

A.D. Maldonado[a,*], P.A. Aguilera[b], A. Salmerón[a]

[a] *Department of Mathematics, University of Almería, Almería, Spain*
[b] *Informatics and Environment Laboratory, Department of Biology and Geology, University of Almería, Almería, Spain*

Abstract

Datasets with an excessive number of zeros are fairly common in several disciplines. The aim of this paper is to improve the predictive power of hybrid Bayesian network classifiers when some of the explanatory variables show a high concentration of values at zero. We develop a new hybrid Bayesian network classifier called *zero-inflated tree augmented naive Bayes* (Zi-TAN) and compare it with the already known *tree augmented naive bayes* (TAN) model. The comparison is carried out through a case study involving the prediction of the probability of presence of two species, the fire salamander (*Salamandra salamandra*) and the Spanish Imperial Eagle (*Aquila adalberti*), in Andalusia, Spain. The experimental results suggest that modeling the explanatory variables containing many zeros following our proposal boosts the performance of the classifier, as far as species distribution modeling is concerned.

Keywords: hybrid Bayesian networks, mixtures of truncated exponentials, zero excess treatment, species distribution modeling

---

*Corresponding author

Email addresses: `ana.d.maldonado@ual.es` (A.D. Maldonado), `aguilera@ual.es` (P.A. Aguilera), `antonio.salmeron@ual.es` (A. Salmerón)

Software availability

The algorithms introduced in this paper have been implemented within the Elvira environment for probabilistic graphical models (Elvira Consortium, 2002), which is a free open source software programmed in Java. The software, including the necessary scripts for replicating the experiments reported in this paper, can be downloaded from the website
`http://www.ual.es/personal/amg457/downloads`
where the datasets used in the paper are also available for download. Both the software and the data are contained in a single zip file, which includes README files with the necessary instructions. The size of the zip file is 6.5 MB.

The software has been compiled with Oracle Java™ SE version 1.8.0_45 build 14. It is ready to work in Windows, Mac and Linux platforms. The datasets provided are in `dbc` format, which is a plain text format used by the Elvira software, which provides facilities for exporting it to `csv` format.

1. Introduction

Environmental datasets tend to present a number of problems, which must be detected and solved in order to obtain plausible results (Ancelet et al., 2010; Lecomte et al., 2013). One of these problems is the presence of data with highly skewed frequency distributions containing an excessive number of zeros. As a consequence, the data do not follow a standard distribution and the application of the usual analysis techniques may yield inaccurate parameter estimates and misleading inferences (Martin et al., 2005). Examples of data with many zeros often occur in different fields, including environmental sciences (Potts and Elith, 2006; Kamarianakis et al., 2008; Dorevitch et al., 2011), ecology (Damgaard, 2008; Wenger and Freeman, 2008; Calama et al., 2011), epidemiology (Böhning et al., 1999; Ngatchou-Wandji and Paris, 2011), genetics (Varona and Sorensen, 2010), biochemistry (Nie et al., 2006; McDavid et al., 2013) or economy (Edmeades and Smale, 2006; Solé-Auró et al., 2012).

The algorithms developed to deal with zero excess are typically focused on the dependent variable. The most popular models are usually extensions of the *Generalized Linear Models*, comprising *zero-inflated Binomial* (ZIB) model (Hall, 2000) for binary variables; *zero-inflated Poisson* (ZIP) (Lambert, 1992), *zero-inflated Negative Binomial* (ZINB) (Greene, 1994), *Poisson* *hurdle* and *Negative Binomial hurdle* (Cragg, 1971; Mullahy, 1986) for discrete variables; and *delta models* or *compound Poisson process* for continuous variables (Ancelet et al., 2010; Lecomte et al., 2013). Generalizing, these models are combinations of probability distributions which separately model the occurrence of zeros and the rest of the domain of the variable of interest. These models are appropriate for handling response variables with high concentration of zeros and have been shown to outperform methodologies that assume the dependent variable to have a standard distribution (Martin et al., 2005).

However, the distribution of the explanatory variables has not typically been of concern and therefore methodologies for dealing with explanatory variables containing high concentration of zeros have not been studied so far. Notwithstanding, accurately modeling the distribution of the explanatory variables is crucial in models such as Bayesian network classifiers, that have been successfully utilized in species distribution analysis (Aguilera et al., 2010). Unfortunately, the methods described above for handling zero excess are not directly applicable to Bayesian network classifiers because they are not designed for modeling conditional distributions and in the case of continuous variables, they rely on distributions that are not compatible with Bayesian network algorithms, as is the case of the Gamma distribution.

Bayesian networks (BNs) belong to the so-called *probabilistic graphical models* and roughly speaking they are compact representations of joint probability distribution over a set of variables whose independence relations are encoded by the structure of an underlying directed acyclic graph (Pearl, 1988). When a BN hosts discrete and continuous variables simultaneously, it is called a *h*ybrid BN. However, not every kind of distribution is compatible with the factorization encoded in a hybrid Bayesian network. One of the most flexible models is based on the use of *mixtures of truncated exponentials* (MTEs), introduced by Moral et al. (2001), generalized later by Shenoy and West (2011) and Langseth et al. (2012).

A hybrid BN classifier is just a BN where one of the variables is the *class* (which is discrete) while the others (discrete or continuous) are the explanatory variables, also called *features* (Aguilera et al., 2011). Typically, when facing classification problems only restricted network structures are considered, such as naive Bayes (NB) or tree augmented naive Bayes (TAN) (Friedman et al., 1997). The NB model assumes that the explanatory variables are independent of each other given the class variable, while the TAN model relaxes that assumption by allowing some dependencies among the features. Within the Environmental Sciences area, the NB model appears to be more popular (Markus et al., 2010; Aguilera et al., 2013; Fytilis and Rizzo, 2013; Ropero et al., 2014, 2015) than the TAN model (Aguilera et al., 2010; Maldonado et al., 2015).

In this paper we address the problem of having a high concentration of zeros in explanatory variables of hybrid BN classifiers. More precisely, we introduce a new model called *zero-inflated* TAN (Zi-TAN) that extends the hybrid BN classifier proposed by Aguilera et al. (2010) by explicitly modeling the zero values. We show how the new model outperforms the formerly used hybrid BN classifier in a case study related to Species Distribution Models (SDM). In the case study, environmental variables are used as explanatory variables of the species occurrence, including climate, land use, soil and lithology. Depending on the scale, these variables may contain a large proportion of zeros which justifies the development of the new model.

The remainder of the paper is organized as follows. We describe Bayesian network classifiers and our baseline model, the TAN, in Section 2. Section 3 is devoted to the methodological aspects of our new proposal. The performance of the new model is analyzed in a case study involving two species in Section 4. The paper ends with conclusions in Section 5.

2. Bayesian networks for classification

A Bayesian network (BN) is a statistical multivariate model for a set of variables $\mathbf{X} = \{X_1, \ldots, X_n\}$, which is defined in terms of two components:

- Qualitative component: A directed acyclic graph (DAG) where each vertex represents one of the variables in the model, and so that the presence of an edge linking two variables indicates the existence of statistical dependence between them.

- Quantitative component: A conditional distribution $p(x_i|pa(x_i))$ for each variable X_i, $i = 1, \ldots, n$ given its parents in the graph, denoted as $pa(X_i)$.

The joint distribution of the variables in the network is therefore represented in a factorized way as

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|pa(x_i)) \qquad \forall x_1, \ldots, x_n \in \Omega_{X_1,\ldots,X_n} \tag{1}$$

where Ω_{X_i} represents the set of all possible values of variable x_i and $pa(x_i)$ denotes an instantiation of the parents of X_i.

Hybrid BNs can handle both discrete and continuous data without imposing restrictions on the interactions among the variables thanks to the development of models such as the *Mixtures of Truncated Exponentials* (MTEs) developed by Moral et al. (2001). The MTE model is characterized by a function defined as follows.

Definition 1. (MTE potential) *Let* $\mathbf{X}$ *be a mixed* n*-dimensional random vector. Let* $\mathbf{W} = (W_1, \ldots, W_d)$ *and* $\mathbf{Z} = (Z_1, \ldots, Z_c)$ *be the discrete and continuous parts of* $\mathbf{X}$*, respectively, with* $c + d = n$*. We say that a function* $f : \Omega_{\mathbf{X}} \mapsto \mathbb{R}_0^+$ *is a* Mixture of Truncated Exponentials potential (MTE potential) *if for each fixed value* $\mathbf{w} \in \Omega_{\mathbf{W}}$ *of the discrete variables* $\mathbf{W}$*, the potential over the continuous variables* $\mathbf{Z}$ *is defined as:*

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

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. We also say that* f *is an MTE potential if there is a partition* $D_1, \ldots, D_k$ *of* $\Omega_{\mathbf{Z}}$ *into hypercubes and in each* D_i*,* f *is defined as in Eq. (2).*

An MTE function is an *MTE density* if it integrates to 1. A *conditional MTE density* can be specified by dividing the domain of the conditioning variables and giving an MTE density of the conditioned variable for each configuration of splits of the other variables. The more the intervals used to divide the domain of the continuous variables, the better the MTE model accuracy but in exchange of a higher number of parameters. To estimate the parameters of MTE densities, we followed the approach recently introduced by Langseth et al. (2014), which is based on least squares optimization, but limiting the number of exponential terms to 2, i.e., $m = 2$ in Eq. (2), in order to keep the complexity of the models moderate.

Hybrid BNs can also be modeled by discretizing the continuous variables, so that all the existing methodology for discrete BNs can be applied with no further modification. The most prominent proposal in this direction is the so-called *dynamic discretization* (Neil et al., 2007) which seeks for better representations of high density areas throughout the inference process. The problem with discretization is to balance the desire for high accuracy in the approximations with a reasonable complexity of the resulting models. A study of the complexity of the MTE approach versus discretization can be found in (Rumí and Salmerón, 2007; Langseth et al., 2009).

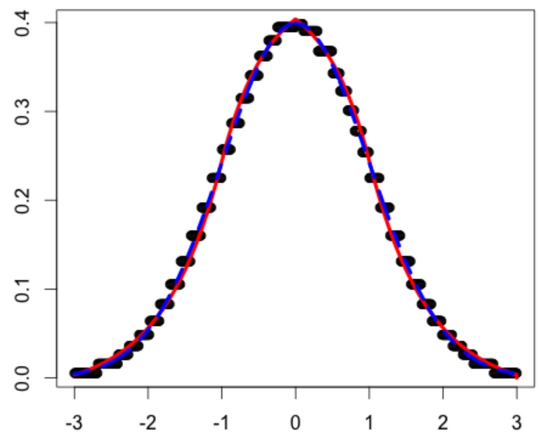


Figure 1: Plot of a standard normal density (dashed blue line) overlaid on an approximation using MTEs (solid red line) and dynamic discretization (solid black line).

As an illustration of the potential advantages of the MTE approach versus dynamic discretization, consider the problem of approximating a standard normal density using both approaches. An approximation using MTEs is given by Cobb et al. (2006) as

$$f(x) = \begin{cases} -0.017203 + 0.9309604e^{1.27x} & \text{if } -3 \leq x < -1, \\ 0.442208 - 0.038452e^{-1.64x} & \text{if } -1 \leq x < 0, \\ 0.442208 - 0.038452e^{1.64x} & \text{if } 0 \leq x < 1, \\ -0.017203 + 0.9309604e^{-1.27x} & \text{if } 1 \leq x < 3, \end{cases} \tag{3}$$

An approximation of the standard normal density using dynamic discretization can be obtained using the AgenaRisk software.[1] Figure 1 shows both approximations overlaid on the plot of the standard normal density between -3 and 3. The plot illustrates how using MTEs the approximation is

[1]http://www.agenarisk.com

smooth while the discretized version is a staircase function. Hence, it is possible to obtain more accurate approximations using fewer parameters with MTEs in general. In this case, the MTE approximation in Equation (3) has 12 parameters while the discretized approximation provided by AgenaRisk has 50 parameters. We have also computed the mean absolute error of both approximations, obtaining a value of 0.0045 for the MTE model and 0.0055 for the discretized one.

Hence, the potential benefits of using MTEs instead of discretized models are: (i) they provide, in general, more accurate approximations using fewer parameters, which leads to more compact models that require fewer parameters to be estimated from data, (ii) they can easily represent variables whose nature is not discrete nor continuous, as we will discuss in Section 3. Furthermore, a discretized model can be seen as a particular case of an MTE where only parameter a_0 in Equation (2) is different from 0.

A Bayesian network can be used as a classifier if it contains a class variable C and a set of continuous or discrete explanatory variables $X_1, \ldots, X_n$, where an object with observed features $x_1, \ldots, x_n$ will be classified as belonging to class $c^* \in \Omega_C$ obtained as

$$c^* = \arg\max_{c \in \Omega_C} p(c|x_1, \ldots, x_n),$$

where Ω_C denotes the set of all posible values of C.

Considering that $p(c|x_1, \ldots, x_n)$ is proportional to $p(c) \times p(x_1, \ldots, x_n|c)$, the specification of an n dimensional distribution for $X_1, \ldots, X_n$ given C is required in order to solve the classification problem, which implies a considerable computational cost, as the number of parameters necessary to specify a joint distribution is exponential in the number of variables, in the worst case. However, this problem is simplified if we take advantage of the factorization encoded by the BN. Since building a network without any structural restriction is not always feasible (they might be as complex as the above mentioned joint distribution), networks with fixed or restricted and simple structures are utilized instead when facing classification tasks. The extreme case is the naive Bayes (NB) structure, where all the feature variables are considered independent given C, as depicted in Fig. 2(a). The strong assumption of independence behind NB models is somewhat compensated by the reduction in the number of parameters to be estimated from data, since in this case, it holds that

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

which means that, instead of one n-dimensional conditional density, n one-dimensional conditional densities must be estimated.

In TAN models, more dependencies are allowed, expanding the NB structure by permitting each feature to have one more parent besides C. It is illustrated in Fig. 2(b). The increase in complexity, in both the structure and the number of parameters, results in richer and more accurate models in general (Friedman et al., 1997).

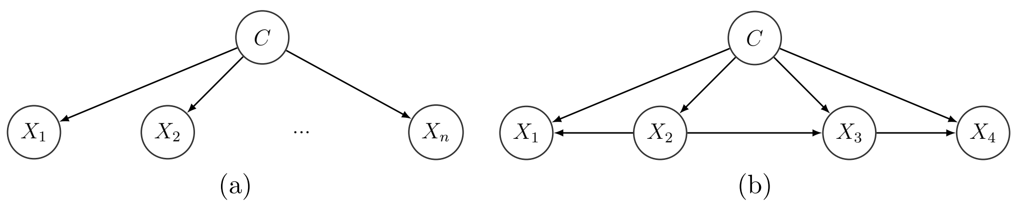


Figure 2: Structure of naive Bayes (a) and TAN (b) classifiers.

In general, there are several possible TAN structures for a given set of variables. The way to choose among them is to construct a maximum weight spanning tree containing the features, where the weight of each edge is the mutual information between the linked variables, conditional on the class (Friedman et al., 1997; Fernández et al., 2007). The mutual information between features X_i and X_j given the class is defined as

$$I(X_i, X_j|C) = \sum_{x_i,x_j,c} \log \frac{p(x_i, x_j|c)}{p(x_i|c)p(x_j|c)} \ . \tag{5}$$

The details for constructing a TAN classifier model are given in Algorithm 1.

3. Zero-inflated TAN based on mixtures of truncated exponentials

In environmental datasets, it is common to find variables with a high concentration of observations at a single repeated value. This makes the modeling of the probability distribution for such variables a problematic task. As an example, consider the histogram on the left panel of Fig. 3. It represents the distribution of *eutric regosols*, used in the case study in Section 4,

Algorithm 1: TAN classifier

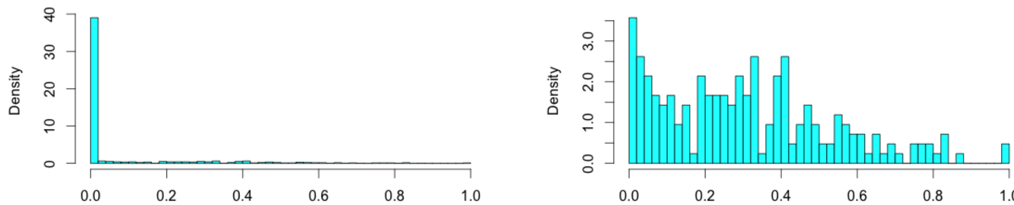
Input: A dataset D with variables $X_1, \ldots, X_n, C$.
Output: A TAN classifier with root variable C and features $X_1, \ldots, X_n$.

1. Calculate the conditional mutual information $I(x_i, x_j|c)$ between each pair of attributes, i $\neq$j.
2. Construct a complete undirected graph with nodes $X_1, \ldots, X_n$ and label each link connecting X_i to X_j by $I(x_i, x_j|c)$.
3. Build a maximum weighted spanning tree $\mathcal{T}$.
4. Transform $\mathcal{T}$ into a directed tree by choosing a root variable, C, and setting the direction of every link to be outward from it.
5. Construct a new network $\mathcal{G}$, with node C being connected to each X_i and nodes $X_1, \ldots, X_n$ having the same links as in $\mathcal{T}$.
6. Estimate an MTE density for C, and a conditional MTE density for each X_i, $i = 1, \ldots, n$ given its parents in $\mathcal{G}$.
7. Let P be a set of estimated densities.
8. Let TAN be a Bayesian network with structure $\mathcal{G}$ and distribution P.
9. **return** TAN.

including values equal to zero, which represent 667 out of 887 values. The distribution is so concentrated at zero that the histogram provides no valuable information for values above zero. By excluding the zero values, the resulting histogram is represented on the right panel of Fig. 3. It is apparent that the distribution of the values greater than zero is far from being uniform, and therefore modeling it accurately can provide benefits in prediction tasks.

Actually, the situation described above is somehow motivated by the fact that the variable is not really discrete nor continuous. Instead, one can consider that there is some probability mass allocated at 0, and the rest of the probability mass is described by a density function. Formally, a variable that is not discrete nor continuous is called a *mixed variable* in Statistics. More precisely, a random variable is mixed if its distribution function has discontinuity jumps at a countable number of points, and it is continuously increasing at least in one interval of values of the variable.

As an example, let $g(x) \geq 0$ for $0 < x \leq 1$ be any non-negative real function such that $\int_0^1 g(x)dx = 1 - p$, with $0 < p < 1$. Then, a random variable X with density function



(a) (b)

Figure 3: Histogram for the proportion of *eutric regosols* including (a) and excluding (b) values equal to zero.

$$f(x) = \begin{cases} p & \text{if } X = 0 \\ g(x) & \text{if } 0 < x \leq 1, \end{cases} \tag{6}$$

is a mixed random variable.

Considering the variable represented in Fig. 3, p would correspond to the fraction of observations allocated at 0 (i.e. the leftmost bar in the left panel), while $g(x)$ would correspond to the rest of the histogram or, equivalently, to the histogram on the right panel of the figure, which can be considered as the result of zooming in the initial histogram for the values of X strictly greater than 0. From now on, we will say that a mixed random variable whose density can be written as in Eq. (6) is a *zero-inflated random variable*. Note that we are considering, without loss of generality, that zero-inflated variables take values on $[0, 1]$. Variables with a different support can be re-scaled.

Zero-inflated random variables have not previously been considered in hybrid Bayesian network literature in general, nor in MTEs in particular. However, they can be easily accommodated within MTE models by incorporating *artificial variables*. More precisely, our proposal consists in including an artificial variable X^* for each mixed variable X in the network, where X^* has no parents and X is its only child. The artificial variable is defined as follows:

$$X^* = \begin{cases} 0 & \text{if } X = 0 \\ 1 & \text{otherwise }, \end{cases} \tag{7}$$

and its probability function is

$$f(x^*) = P(X^* = x^*) = \begin{cases} p & \text{if } x^* = 0 \\ 1-p & \text{if } x^* = 1, \end{cases} \tag{8}$$

where p is as in Eq. (6). Note that $f(x^*)$ is trivially an MTE, according to Eq. (2).

The conditional distribution of X given X^* is

$$f(x|x^*) = \begin{cases} 1 & \text{if } x^* = 0, x = 0 \\ \dfrac{1}{1-p} g(x) & \text{if } x^* = 1, 0 < x \leq 1, \end{cases} \tag{9}$$

with p and $g(x)$ as in Eq. (6). Again, $f(x|x^*)$ is an MTE whenever $g(x)$ is an MTE as well. Note that so far we have made no assumptions about $g(x)$ beyond those required for the corresponding density function being well defined - see Eq. (6). We will see later in Definition 2 that $g(x)$ plays the role of the conditional MTE distributions in a TAN classifier. The following proposition states that the introduction of artificial variables does not modify the marginal distribution of the zero-inflated variable.

Proposition 1. *Let X^* be a binary random variable with probability function as in Eq.* (8) *and let X be a random variable whose distribution conditional on X^* is as given in Eq.* (9)*. Then, X is a zero-inflated random variable with marginal distribution as in Eq.* (6)*.*

Proof. The joint distribution of X and X^* is $f(x, x^*) = f(x|x^*)f(x^*)$, which can be written as

$$f(x, x^*) = \begin{cases} p \times 1 & \text{if } x = 0, x^* = 0 \\ \dfrac{1}{1-p} g(x) \times (1-p) & \text{if } 0 < x \leq 1, x^* = 1 \end{cases}$$

which amounts to

$$f(x, x^*) = \begin{cases} p & \text{if } x = 0, x^* = 0 \\ g(x) & \text{if } 0 < x \leq 1, x^* = 1. \end{cases}$$

Therefore, the marginal distribution for X is obtained by marginalizing out X^* as follows:

$$f(x) = \sum_{x^*=0}^{1} f(x, x^*) = \begin{cases} p & \text{if } x = 0 \\ g(x) & \text{if } 0 < x \leq 1, \end{cases}$$

which matches Eq. (6). $\square$

Our methodological proposal consists of including zero-inflated random variables in TAN classifiers, resulting in a new Bayesian network classifier formally defined as follows.

Definition 2. *Let $\mathcal{T}$ be a TAN classifier over class variable C and features $X_1, \ldots, X_n$. Let X_i, $i \in I \subset \{1, \ldots, n\}$ be a set of zero-inflated random variables. A* zero-inflated TAN *(Zi-TAN) classifier $\mathcal{T}^*$ is obtained from $\mathcal{T}$ by:*

1. *Inserting, for each variable X_i, $i \in I$, an artificial variable X_i^* as in Eq. (7) and a link $X_i^* \rightarrow X_i$.*
2. *Attaching to each node $X_i^*, i \in I$ a distribution as in Eq. (8).*
3. *Attaching to each node $X_i, i \in I$ with parents $\{Y_1, \ldots, Y_m\}$ in $\mathcal{T}$, and conditional distribution $f(x_i|y_1, \ldots, y_m)$ in $\mathcal{T}$, a new conditional distribution*

$$f(x_i|x_i^*, y_1, \ldots, y_m) = \begin{cases} 1 & \text{if } x_i^* = 0, x_i = 0 \\ \dfrac{1}{1-p} g(x_i|y_1, \ldots, y_m) & \text{if } x_i^* = 1, 0 < x_i \leq 1, \end{cases} \tag{10}$$

where p is the proportion of values of X_i equal to 0 and $g(x_i|y_1, \ldots, y_m)$ is a conditional MTE density for X_i given $\{Y_1, \ldots, Y_m\}$ learnt from the same sample as $f(x_i|y_1, \ldots, y_m)$ but excluding the values where $X_i = 0$.

Notice that the new conditional distributions defined in Eq. (10) are of class MTE as long as the distributions in the original TAN model are MTEs as well. Also, the role of $g(x_i|y_1, \ldots, y_m)$ corresponds to that of $g(x)$ in Eq.(6). Such conditional distributions are learnt making use of the procedure introduced by Langseth et al. (2014).

Example 1. *Consider the TAN structure in Figure 2(b). Assume that X_1 and X_4 are zero-inflated random variables. The corresponding Zi-TAN structure, according to Definition 2, is shown in Figure 4.*

The insertion of the artificial variables when constructing the Zi-TAN means that the new model can be factorized as a sum of TAN models, one per each combination of values of the artificial variables. However, from a practical point of view it is not a problem, as the joint distribution over the

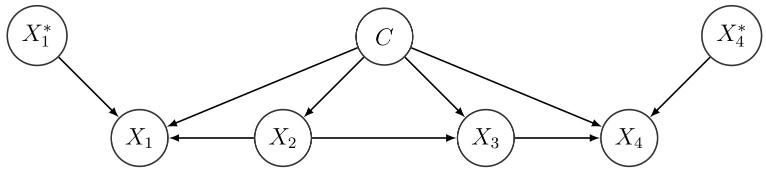


Figure 4: An example of a Zi-TAN classifier structure, obtained from Figure 2(b) assuming that X_1 and X_4 are zero-inflated random variables. X_1^* and X_4^* are their respective artificial variables.

class variables and the features is not affected, as shown in Proposition 2. Recall that the aim of the Zi-TAN model is not to modify the underlying distribution over the variables in the domain being analyzed, but rather to express it in a way that permits overcoming the problem of high concentration of values at zero.

Proposition 2. *Let $\mathcal{T}$ be a TAN classifier over class variable C and features $X_1, \ldots, X_n$, and $\mathcal{T}^*$ be a Zi-TAN classifier constructed as in Definition 2. Then, $\mathcal{T}^*$ encodes the same probability distribution as $\mathcal{T}$ over variables $\{C, X_1, \ldots, X_n\}$.*

Proof. According to Proposition 1, marginalizing out each artificial variable X_i^* in $\mathcal{T}^*$ yields a conditional distribution for X_i exactly equal to the one it had in $\mathcal{T}$. Therefore, after removing all the artificial variables in $\mathcal{T}^*$, both models become the same. □

The details on how to build a Zi-TAN classifier from data are given in Algorithm 2. It relies on Definition 2 and Algorithm 1.

4. Case study

In this section, the methodology explained above is applied to SDMs. More precisely, we considered two case studies involving the Fire Salamander and the Spanish Imperial Eagle.

4.1. Study area

The study area is Andalusia, a region in southern Spain which occupies an area of 87 000 km^2 and whose latitude and longitude is between 36°N - 38°44'N and 3°50'W - 0°34'E. As far as elevation is concerned, the study area ranges from 0 to 3460 meters above the sea level. The main mountain ranges

Algorithm 2: Zi-TAN classifier

Input: A dataset D with variables $X_1, \ldots, X_n, C$. A set of indices $I \subset \{1, \ldots, n\}$ of zero-inflated variables.

Output: A Zi-TAN classifier with root variable C and features $\{X_1, \ldots, X_n\} \cup \{X_i | i \in I\}$.

1 Build a TAN model, $\mathcal{T}$, from dataset D, for variables $X_1, \ldots, X_n, C$ using Algorithm 1.

2 $\mathcal{T}^* \leftarrow \mathcal{T}$.

3 **for** $i \in I$ **do**

4 Create a new binary variable

$$X_i^* = \begin{cases} 0 & \text{if } X_i = 0 \\ 1 & \text{otherwise .} \end{cases}$$

5 Add a new column to dataset D, corresponding to X_i^*

6 Insert a new link $X_i^* \rightarrow X_i$ in $\mathcal{T}^*$.

7 Let p be the proportion of values $X_i^* = 0$ in D.

8 Attach to X_i^* in $\mathcal{T}^*$ the distribution

$$f(x_i^*) = \begin{cases} p & \text{if } x_i^* = 0 \\ 1 - p & \text{if } x_i^* = 1, \end{cases}$$

9 Let $\{Y_1, \ldots, Y_m\}$ be the parents of X_i in $\mathcal{T}$.

10 Let $f(x_i | y_1, \ldots, y_m)$ be the conditional distribution of X_i given its parents in $\mathcal{T}$.

11 Estimate a new density $g(x_i | y_1, \ldots, y_m)$ from the same data used to learn $f(x_i | y_1, \ldots, y_m)$, but excluding the elements in the sample where $X_i = 0$.

12 Attach to X_i in $\mathcal{T}^*$ a new conditional distribution

$$f(x_i | x_i^*, y_1, \ldots, y_m) = \begin{cases} 1 & \text{if } x_i^* = 0, x_i = 0 \\ \dfrac{1}{1-p} g(x_i | y_1, \ldots, y_m) & \text{if } x_i^* = 1, 0 < x_i \leq 1, \end{cases}$$

13 **return** $\mathcal{T}^*$.

of Andalusia are the Sierra Morena mountain range (in the North) and the Baetic systems (in the South), which are separated by the Baetic depression, the lowest territory in Andalusia (Figure 5). The flattest areas correspond to the littoral and the Baetic depression, through which the Guadalquivir river runs, and the steepest ones to the Baetic Systems, comprising the Prebaetic, Subbaetic and Pennibaetic systems.

The geographic location determines Andalusia's climate, which belongs to the Mediterranean domain. The Mediterranean climate alternates mild, rainy and humid winters with dry and warm summers. The average annual temperature usually does not drop below 15°C, as a consequence of the ocean influence. On the other hand, precipitation (P) shows a high spatial variability, ranging from 170 mm/year to 2180 mm/year. In addition, potential evapotranspiration (PET) ranges from about 300 mm in the eastern Baetic systems to more than 1000 mm/year in both the Guadalquivir river area and the eastern coast. As a result, the quotient of precipitation divided by PET, i.e. the humidity index, varies along the study area from surplus or areas with hydric excess ($P > PET$) to deficit or areas lacking in water resources ($P < PET$).

Regarding land use, half of the study area corresponds to natural vegetation, followed by croplands (44%), urbanized areas (3%) and bodies of water (3%). With reference to soil classification, cambisols are the most common soil group (41%), followed by regosols (19%), vertisols (9%), litosols (8.2%), luvisols (8.15%), fluvisols (5.5%), planosols (2.5%), xerosols (2%), solonchaks (2%), arenosols (1%), dunes (0.12%) and histosols (0.01%). On the subject of lithology, 64% of the Andalusia's crust consists of sedimentary rocks, followed by metamorphic rocks (26%), plutonic rocks (6%) and volcanic rocks (4%).

4.2. Data description

Data from different thematic maps (Table 1) were incorporated into a geographic information system - ArcGis (ESRI® ArcMap™10.2.2). A 10 x 10 km grid with presence records of different species was superimposed on the thematic maps in order to build a matrix composed of a number of explanatory variables and 1 target variable. The coordinate system for all the datasets is based on the European Terrestrial Reference System 1989 (ETRS89).

The percentage of each land use, soil and lithology variable within each cell was calculated by dividing the area of each variable by the total cell

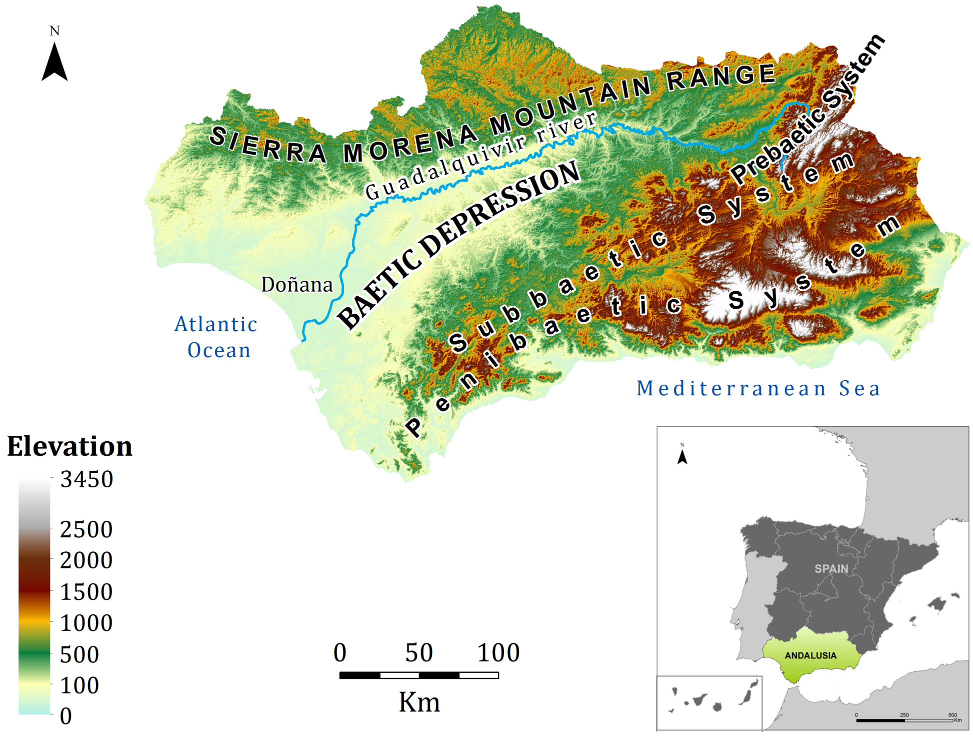


Figure 5: Study area, Andalusia, Spain (37°23'00"N 5°59'00"W).

area. In addition, the average of annual mean temperature, annual rainfall, annual PET and annual humidity index were calculated for the 30 year period 1971-2000 for each cell in the grid. Moreover, the average elevation, slope and aspect of each cell was calculated from the Andalusian Digital Terrain Model (DTM). The average elevation and slope in each cell were determined by means of their arithmetic means. Since aspect is measured clockwise from 0 to 360 degrees from the north, describing a full circle, its average ($\bar{A}$) was calculated as in Davis (1986)

$$\bar{A} = \arctan\left(\frac{\sum_{i=1}^{n} \sin\theta_i}{\sum_{i=1}^{n} \cos\theta_i}\right), \tag{11}$$

where θ_i is the aspect angle in each DTM pixel and n is the number of pixels.

Once the value of each variable was calculated for the 10 x 10 km grid, those cells occupied by less than 50% of terrestrial surface were removed. Afterwards, a matrix composed of 156 variables taking values over 887 cells was obtained, where the class variable is the presence of a particular species and the remaining 155 variables are the explanatory variables (Table 1). Finally, the data were rescaled to interval [0,1] in order to prevent numerical instability problems.

We have considered an explanatory variable to be zero-inflated if more than 50% of its observations are zeros. In this regard, 128 explanatory variables out of 155 came out to be zero-inflated. In order to test the *zero-inflated* TAN (Zi-TAN) classifier, 2 species were chosen: 1) the Fire Salamander (*Salamandra salamandra*; from now on referred to as salamander), which is present in 300 out of 887 cells; therefore it was neither too present nor too absent; and 2) the Spanish Imperial Eagle (*Aquila adalberti*; from now on referred to as eagle), whose presence in the study area is imbalanced and scarce, occurring in 54 out of 887 cells. Therefore, the 2 species are different from the point of view of occurrence.

4.3. Variable selection

An excessive number of variables may decrease both the generalization of the model by overfitting and its performance by introducing noise. Since the dataset contained a high number of variables, a variable selection process was carried out. For both species, the selection process was entirely carried out by experts.

In the case of salamander, since the species prefers dark moist areas, with rainfall being abundant, and is widely found in the Mediterranean forest at medium-high elevations (Degani and Warburg, 1978), the selected variables were rainfall, humidity index, dense oak woodlands, oak woodlands with shrub, oak woodlands with herbaceous crops, woodlands with herbaceous crops, grassland, olive groves, eutric regosols, calcaric regosols, eutric cambisols, sand-silt-clay-gravel lithology type, slate-greywacke-sandstone lithology type and volcano-sedimentary complex lithology type. Within this selection, 7 out of 14 were zero-inflated explanatory variables.

Regarding eagle, the 3 main populations in Andalusia differ in terms of environmental conditions, ranging from the mediterranean forest to low-lying marshes near the ocean (González et al., 2008). Therefore, the selected variables were temperature, rainfall, evapotranspiration, rainfed herbaceous crops, oak woodlands with shrub, oak woodlands with herbaceous crops, marshes, eutric regosols, albic arenosols, solonchaks, eutric cambisols, slate-shale-greywacke-quartzite lithology type, slate-greywacke-sandstone lithology type, sand lithology type and silt-clay lithology type. Within this selection, 10 out of 15 were zero-inflated explanatory variables.

4.4. Learning species distribution models

Once the variables were selected, we used Algorithm 1 for learning the TAN model and Algorithm 2 for learning the Zi-TAN model.

A powerful advantage of Bayesian networks is their capability of predicting a full specification of the posterior distribution of the class variable. In the case of species distribution modeling, we are interested in determining the probability of occurrence of a particular species, rather than giving a fixed binary prediction.

The classifiers described in Sections 2 and 3 can be used to depict the potential distribution area, defined as the *probability of presence of the species* in each 10x10 km cell given the explanatory variables $(X_1, \ldots, X_n)$ included in the model. Each cell in the map can be colored across a gradual white to red color ramp, where low probability of presence (0-0.2) is represented with the white color and high probabilities (0.8-1) with dark red.

We have distinguished two learning scenarios, one for validation (see section 4.5) and the other for plotting the species distribution maps described above. The models used to plot the maps were learnt from a random subsample containing an 80% of the original dataset, aiming at avoiding overfitting.

4.5. Model validation

The models were validated by means of the k-fold cross validation technique (Stone, 1974), obtaining a confusion matrix in each fold. This technique randomly splits the complete dataset into k subsets, using *k-1* of them for learning the model and one for validation. The method is repeated k times and the confusion matrix of the model is computed in each step. In order to compute a confusion matrix, the output of the model, given in terms of "probability of presence of the species", P_S, was transformed into absence and presence records, coded as "1" if $P_S \geq 0.5$ or "0" if $P_S < 0.5$ respectively. In this case study, a k value of 10 was applied and therefore 10 confusion matrices of each model were obtained.

A number of statistics were computed from the confusion matrices. Their definition is given in Table 2. Of special interest is the area under the ROC curve (AUC), which measures the predictive power of the model considering both the true positive (Recall) and the true negative (Specificity) rates. When AUC is defined by only 1 run, it is known as *balanced accuracy* (Sokolova et al., 2006).

The models were validated in terms of Accuracy, Recall, $\mathrm{f-score}$ and AUC. Precision and Specificity are also defined in Table 2 because they are used in the definition of the four selected statistics. All of them measure some aspect of the performance of the model and range from 0 to 1, with models scoring close to 1 being better. In the case of AUC, a score of 0.5 represents a predictive power no better than predicting the class at random. Wilcoxon's signed rank Test was performed to detect differences between the performance statistics of TAN and Zi-TAN models.

4.6. Results and discussion of the case study

4.6.1. Salamander

Figure 6 represents the qualitative component of the TAN (a) and Zi-TAN (b) models for salamander. Figure 7 shows the box plots corresponding to the 10 measures of the 4 performance statistics for TAN and Zi-TAN, along with the $p - values$ obtained in the pairwise comparison.

The Wilcoxon Test showed no significant differences ($p > 0.05$) between both models regarding Accuracy. However, this measure is not recommendable when the distribution of the classes is unequal, as in this case where there are 300 presences and 587 absences. On the other hand, the test showed significant differences ($p < 0.05$) between TAN and Zi-TAN for the remaining statistics: Recall, f-score and AUC. According to the statistical test applied, Zi-TAN performs better than TAN with reference to the true presence rate (Recall), i.e. the proportion of observed presences correctly classified was higher in the Zi-TAN model. Regarding f-score and AUC measures, Zi-TAN model scores higher than TAN, suggesting that the former classifies better than the latter in the case of salamander.

Figure 8 shows the probability of presence of the salamander according to the TAN (a) and Zi-TAN (b) models, along with its observed distribution area. Both models recognize the general distribution pattern of the species, identifying 3 main populations: 1 sub-population located in the north, along the Sierra Morena mountain range; another sub-population located in the northeast, in the Prebaetic System; and another sub-population located in the south-southwest, in the Penibaetic System. Note that Zi-TAN misclassified more observed absences than TAN, i.e. the type I error is lower in the TAN model. From the ecological point of view, the type I error or, in this case, false presence may be understood as the potential distribution area for the species, i.e., where salamander has never been seen but the environmental conditions meet its necessities. According to Martin et al. (2005) and Lecomte et al. (2013), observed true absences may occur if the species does not saturate its entire suitable area. The potential distribution area, based on the selected variables, predicted by the Zi-TAN model nearly corresponds to the observed distribution of the species.

4.6.2. Eagle

Figure 9 represents the qualitative component of the TAN (a) and Zi-TAN (b) models for eagle. Figure 10 shows the box plots corresponding to the 10 measures of the 4 performance statistics for TAN and Zi-TAN, along with the $p-values$ obtained in the pairwise comparison.

The Wilcoxon Test showed no significant differences ($p > 0.05$) between both models regarding Accuracy. As mentioned above, Accuracy may not be appropriate, especially for imbalanced datasets (Chawla, 2005). On the other hand, the test showed significant differences ($p < 0.05$) between TAN and Zi-TAN for Recall and AUC. The f-score statistic could not be calculated for TAN since every fold of the cross validation yielded a Recall of 0 and an undefined Precision (0 divided by 0) due to the fact that the model only predicted absences in this case. According to the statistical test applied, Zi-TAN performs better than TAN with reference to the true presence rate (Recall), i.e. the proportion of observed presences correctly classified was higher in the Zi-TAN model. Regarding AUC, Zi-TAN model scores higher than TAN, suggesting that the former classifies better than the latter in the case of eagle.

Figure 11 shows the potential distribution area of eagle, based on the selected variables, given by TAN (a) and Zi-TAN (b) along with its observed distribution area. Examining these maps, it is noticeable that the TAN classifier is more conservative than Zi-TAN, since the former relies more on the probability of the dominant class, absences, and barely classifies observations as presences. The TAN model obtained just 5 cells with probability of presence higher than zero. The poor model's performance may be due to the combination of a great number of zero-inflated explanatory variables (10 out of 15) and the imbalanced class variable (55 observed presences out of 887 observations). In contrast, the Zi-TAN model fairly detected the 3 main populations of eagle in Andalusia: Doñana, Eastern Sierrra Morena and Central Sierra Morena. The model also marked, with low probability, the Campo de Gibraltar county, in the south, as a potential distribution area of eagle, which was occupied by the eagle in the past (Gonzalez et al., 1989; González et al., 2008).

5. Conclusions

We have developed Zi-TAN, a new model for dealing with *zero-inflated* feature variables, using hybrid Bayesian networks. Our experimental results showed strong evidence that our proposed methodology for modeling explanatory variables with zero excess improves the performance of the classifier. In the case of salamander, an abundant species in the study area, the TAN model recognized the general pattern of the species and had a fair performance whereas the distribution area predicted by the Zi-TAN model corresponds almost exactly to the observed distribution area. In the case of eagle, a scarce species in the study area, the TAN model had a poor performance while Zi-TAN substantially improved the distribution area predicted by the former. The technique explained in this paper can be applied to species distribution models where the explanatory variables have an excessive number of zeros. Further research needs to be done in order to argue its application to other disciplines.

Acknowledgements

This work has been supported by the Spanish Ministry of Economy and Competitiveness, through project TIN2013-46638-C3-1-P , by Junta de Andalucía through project P11-TIC-7821 and by ERDF-FEDER funds. A.D. Maldonado is being supported by the Spanish Ministry of Education, Culture and Sport through an FPU research grant, FPU2013/00547.

References

Aguilera, P. A., Fernández, A., Fernández, R., Rumí, R., Salmerón, A., 2011. Bayesian networks in environmental modelling. Environmental Modelling & Software 26, 1376–1388.

Aguilera, P. A., Fernández, A., Reche, F., Rumí, R., 2010. Hybrid Bayesian network classifiers: Application to species distribution models. Environmental Modelling & Software 25 (12), 1630–1639.

Aguilera, P. A., Fernández, A., Ropero, R. F., Molina, L., 2013. Groundwater quality assessment using data clustering based on hybrid Bayesian networks. Stochastic Environmental Research & Risk Assessment 27 (2), 435–447.

Ancelet, S., Etienne, M.-P., Benoît, H., Parent, E., 2010. Modelling spatial zero-inflated continuous data with an exponentially compound Poisson process. Environmental and Ecological Statistics 17, 347–376.

Böhning, D., Dietz, E., Schlattmann, P., 1999. The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. Journal of the Royal Statistical Society A 162(2), 195–209.

Calama, R., Mutke, S., Tom'e, J., Gordo, J., Monterio, G., Tomé, M., 2011. Modelling spatial and temporal variability in a zero-inflated variable: The case of stone pine (*Pinus pinea* L.) cone production. Ecological Modelling 222, 606–618.

Chawla, N., 2005. Data mining for imbalanced datasets: An overview. In: Maimon, O., Rokach, L. (Eds.), Data Mining and Knowledge Discovery Handbook. Springer US, pp. 853–867.

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

Cragg, J. G., 1971. Some statistical models for limited dependent variables with application to the demand for durable goods. Econometrica 39(5), 829–844.

Damgaard, C., 2008. Modelling pin-point plant cover data along an environmental gradient. Ecological Modelling 214, 404–410.

Davis, J., 1986. Statistical and Data Analysis in Geology. J. Wiley.

Degani, G., Warburg, M., 1978. Population structure and seasonal activity of the adult *Salamandra salamandra* (L.) (Amphibia, Urodela, Salamandridae) in Israel. Journal of Herpetology 12, 437–444.

Dorevitch, S., Doi, M., Hsu, F.-C., Lin, K.-T., Roberts, J. D., Liu, L. C., Gladding, R., Vannoy, E., Li, H., Javor, M., Scheff, P. A., 2011. A comparison of rapid and conventional measures of indicator bacteria as predictors of waterborne protozoan pathogen presence and density. Journal of Environmental Monitoring 13, 2427–2435.

Edmeades, S., Smale, M., 2006. A trait-based model of the potential demand for a genetically engineered food crop in a developing economy. Agricultural Economics 35, 351–361.

Elvira Consortium, 2002. Elvira: An Environment for Creating and Using Probabilistic Graphical Models. In: Proceedings of the First European Workshop on Probabilistic Graphical Models. pp. 222–230.
URL `http://leo.ugr.es/elvira`

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

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

Fytilis, N., Rizzo, D. M., 2013. Coupling self-organizing maps with a Naïve Bayesian classifier: Stream classification studies using multiple assessment data. Water Resources Reseach 49, 7747–7762.

Gonzalez, L. M., Hiraldo, F., Delibes, M., Calderon, J., 1989. Reduction in the range of the Spanish Imperial Eagle (*Aquila adalberti* Brem, 1861) since AD 1850. Journal of Biogeography 16, 305–315.

González, L. M., Oria, J., Sánchez, R., Margalida, A., Aranda, A., Prada, L., Caldera, J., Molina, J. I., 2008. Status and habitat changes in the endangered Spanish Imperial Eagle *Aquila adalberti* population during 1974-2004: implications for its recovery. Bird Conservation International 18, 242–259.

Greene, W. H., 1994. Accounting for excess zeros and sample selection in Poisson and Negative Binomial regression models. Tech. rep., Department of Economics, Stern School of Business, New York University.

Hall, D. B., 2000. Zero-Inflated Poisson and Binomial regression with random effects: A case study. Biometrics 56, 1030–1039.

Kamarianakis, Y., Feidas, H., Kokolatos, G., Chrysoulakis, N., Karatzias, V., 2008. Evaluating remotely sensed rainfall estimates using nonlinear mixed models and geographically weighted regression. Environmental Modelling & Software 23, 1438–1447.

Lambert, D., 1992. Zero-Inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34, 1–14.

Langseth, H., Nielsen, T., Pérez-Bernabé, I., Salmerón, A., 2014. Learning mixtures of truncated basis functions from data. International Journal of Approximate Reasoning 55, 940–956.

Langseth, H., Nielsen, T. D., Rumí, R., Salmerón, A., 2009. Maximum likelihood learning of conditional MTE distributions. ECSQARU'09. Lecture Notes in Artificial Intelligence 5590, 240–251.

Langseth, H., Nielsen, T. D., Rumí, R., Salmerón, A., 2012. Mixtures of Truncated Basis Functions. International Journal of Approximate Reasoning 53 (2), 212–227.

Lecomte, J., Benoît, H., Etienne, M., Bel, L., Parent, E., 2013. Modeling the habitat associations and spatial distribution of benthic macroinvertebrates: A hierarchical Bayesian model for zero-inflated biomass data. Ecological Modelling 265, 74–84.

Maldonado, A. D., Aguilera, P. A., Salmerón, A., 2015. Continuous Bayesian networks for probabilistic environmental risk mapping. Stochastic Environmental Research & Risk Assessment In press.

Markus, M., Hejazi, M. I., Bajcsy, P., Giustolisi, O., Savic, D. A., 2010. Prediction of weekly nitrate-N fluctuations in a small agricultural watershed in Illinois. Journal of Hydroinformatics 12.3, 251–261.

Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., Tyre, A. J., Possingham, H. P., 2005. Zero torelance ecology: improving ecological inference by modelling the source of zero observations. Ecology Letters 8, 1235–1246.

McDavid, A., Finak, G., Chattopadyay, P. K., Dominguez, M., Ma, L. L. S. S., Roederer, M., Gottardo, R., 2013. Data explotation, quiality contorl and testing in single-cell qPCR-based gene expression experiments. Bioinformatics 29(4), 461–467.

Moral, S., Rumí, R., Salmerón, A., 2001. Mixtures of Truncated Exponentials in Hybrid Bayesian Networks. In: Benferhat, S., Besnard, P. (Eds.), Symbolic and Quantitative Approaches to Reasoning with Uncertainty. Vol. 2143 of Lecture Notes in Artificial Intelligence. Springer, pp. 156–167.

Mullahy, J., 1986. Specification and testing of some modified count data models. Journal of Econometrics 33, 341–365.

Neil, M., Tailor, M., Marquez, D., 2007. Inference in hybrid Bayesian networks using dynamic discretization. Statistics and Computing 17, 219–233.

Ngatchou-Wandji, J., Paris, C., 2011. On the zero-inflated count models with application to modelling annual trends in incidences of some occupational allergic diseases in France. Journal of Data Science 9, 639–659.

Nie, L., Wu, G., Brockman, F., Zhang, W., 2006. Integrated analysis of transcriptomic and proteomic data of *D*esulfovibrio vulgaris: zero-inflated Poisson regression models to predict abundance of undetected proteins. Bioinformatics 22(13), 1641–1647.

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

Potts, J. M., Elith, J., 2006. Comparing species abundance models. Ecological Modelling 199, 153–163.

Ropero, R. F., Aguilera, P. A., Fernández, A., Rumí, R., 2014. Regression using hybrid Bayesian networks: Modelling landscape-socioeconomy relationships. Environmental Modelling & Software 54, 127–137.

Ropero, R. F., Aguilera, P. A., Rumí, R., 2015. Analysis of the socioecological structure and dynamics of the territory using a hybrid Bayesian network classifier. Ecological Modelling 311, 73–87.

Rumí, R., Salmerón, A., 2007. Approximate probability propagation with mixtures of truncated exponentials. International Journal of Approximate Reasoning 45, 191–210.

Shenoy, P. P., West, J. C., 2011. Inference in hybrid Bayesian networks using mixtures of polynomials. International Journal of Approximate Reasoning 52 (5), 641–657.

Sokolova, M., Japkowicz, N., Szpakowicz, S., 2006. Beyond accuracy, f-score and roc: A family of discriminant measures for performance evaluation. In: Sattar, A., Kang, B.-h. (Eds.), AI 2006: Advances in Artificial Intelligence. Vol. 4304 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 1015–1021.

Solé-Auró, A., Guillén, M., Crimmins, E. M., 2012. Health care usage among immigrants and native-born elderly populations in eleven European countries: results from SHARE. European Journal of Health Economics 13, 741–754.

Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society. Series B (Methodological) 36 (2), 111–147.

Varona, L., Sorensen, D., 2010. A genetic analysis of mortality in pigs. Genetics Society of America 184, 277–284.

Wenger, S. J., Freeman, M. C., 2008. Estimating species occurrence, abundance, and detection probability using zero-inflated distributions. Ecology 89(10), 2953–2959.

Table 1: Summary of variables. Note that not all these variables were included in the models but a variable selection process was carried out.

Variable	**Description**	**Source**
Salamander /Eagle	Presence/absence of the given species in each cell	Spanish inventory of terrestrial species [a]
T (°C)	Average of annual mean temperature for the 30 year period 1971-2000 in each cell	Annual Mean Temperature Dataset of Andalusia. TIFF raster format with 100 m spatial resolution [b]
Rainfall (mm)	Average of annual rainfall for the 30 year period 1971-2000 in each cell	Annual Precipitation Dataset of Andalusia. TIFF raster format with 100 m spatial resolution [b]
PET (mm)	Average of the annual potential evapotranspiration for the 30 year period 1971-2000 in each cell	Annual Mean Evapotranspiration Dataset of Andalusia. TIFF raster format with 100 m spatial resolution [b]
Humidity index	Average of annual humidity index for the 30 year period 1971-2000 in each cell	Annual Mean Humidity Index Dataset of Andalusia. Shapefile format [b]
Land uses (%)	Percentage of occupation of each land-use (#44) within each cell	Andalusian Land Use and Land Cover Map (1:25,000) [b]
Soil (%)	Percentage of occupation of each soil type (#63) within each cell	Andalusian Soil Map (1:400,000) [b]
Lithology (%)	Percentage of occupation of each lithological unit (#41) within each cell	Andalusian Lithological Map (1:400,000) [b]
Z (m a.s.l.)	Average elevation of each cell	Andalusian Digital Terrain Model. Grid width 200 m spatial resolution [c]
Slope (%)	Average slope of each cell	
Aspect (°)	Average aspect of each cell	

number of variables

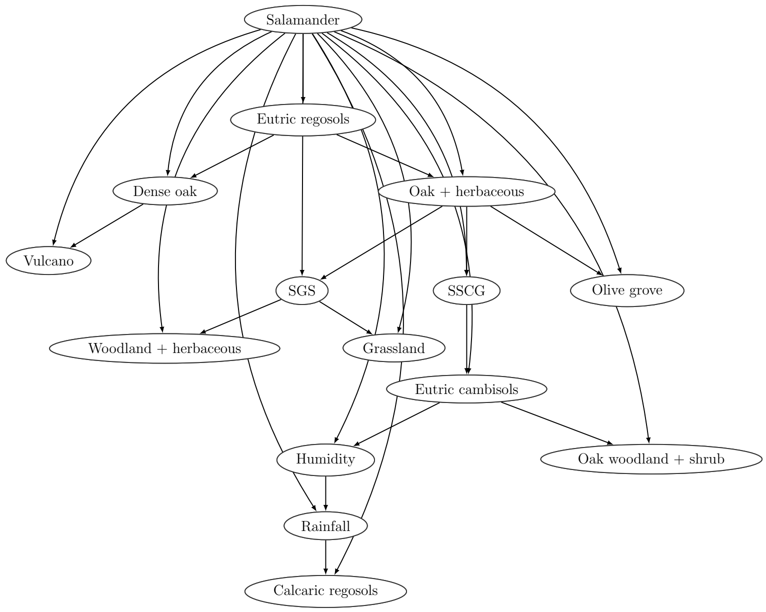
[a]Provided by the Ministry of Agriculture, Food and Environment

[b]Provided by the Andalusian Environmental Information Network

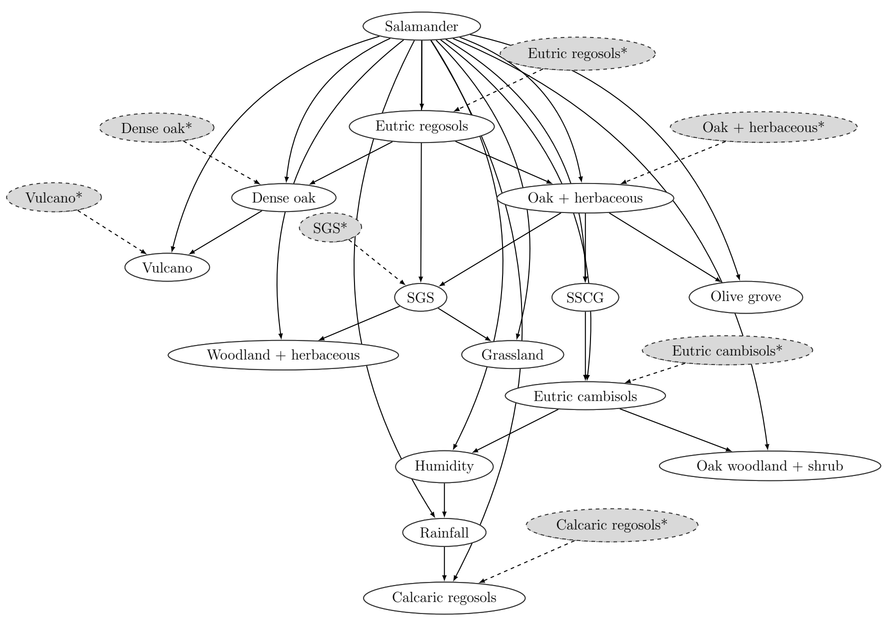
[c]Provided by the Spanish National Geographic Institute

Table 2: Statistics used to validate the classification models. TP, TN, FP and FN are, respectively, the true positive, true negative, false positive and false negative rates. Parameter β is the relative importance of Precision vs Recall, and was set to 1 in the case study.

Statistic	Definition	Statistic	Definition
Accuracy	$\frac{TP+TN}{TP+FN+FP+TN}$	$\mathrm{f-score}$	$\frac{(1+\beta)\times\mathrm{Recall}\times\mathrm{Precision}}{\beta^2\times\mathrm{Recall}+\mathrm{Precision}}$
Recall	$\frac{TP}{TP+FN}$	Specificity	$\frac{TN}{TP+FN}$
Precision	$\frac{TP}{TP+FP}$	AUC	$\frac{1}{2}(\mathrm{Recall}+\mathrm{Specificity})$



(a) TAN structure



(b) Zi-TAN structure

Figure 6: Structure of TAN (a) and Zi-TAN (b) models for predicting presence of salamander given the explanatory variables. SGS: slate-greywacke-sandstone; SSCG: sand-silt-clay-gravel. Shaded nodes with dashed lines represent the artificial binary variables used to model the zero-inflated explanatory variables. Note that the dependence relationships existing between the feature variables just allow the models to perform better.

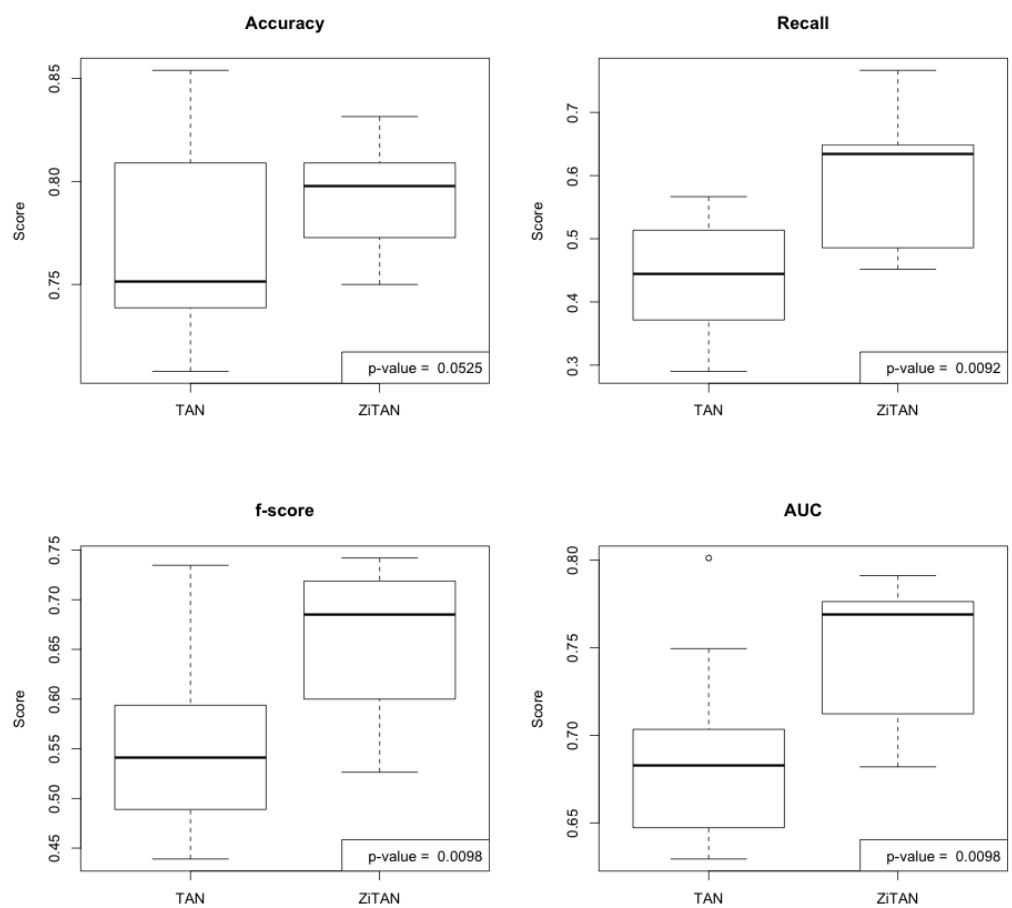
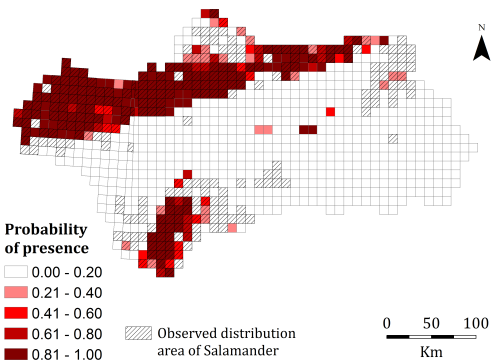
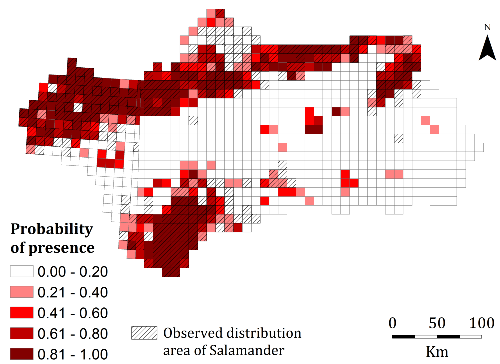


Figure 7: Box plots comparing TAN and Zi-TAN in terms of their performance statistics in the *Salamandra salamandra* case. The $p - values$ obtained in the Wilcoxon Test are shown.

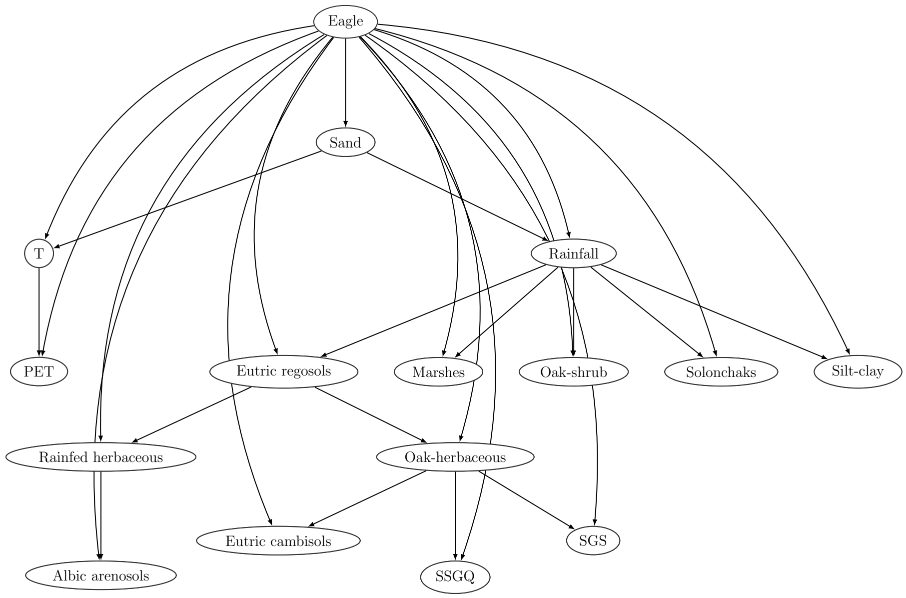


(a) TAN model

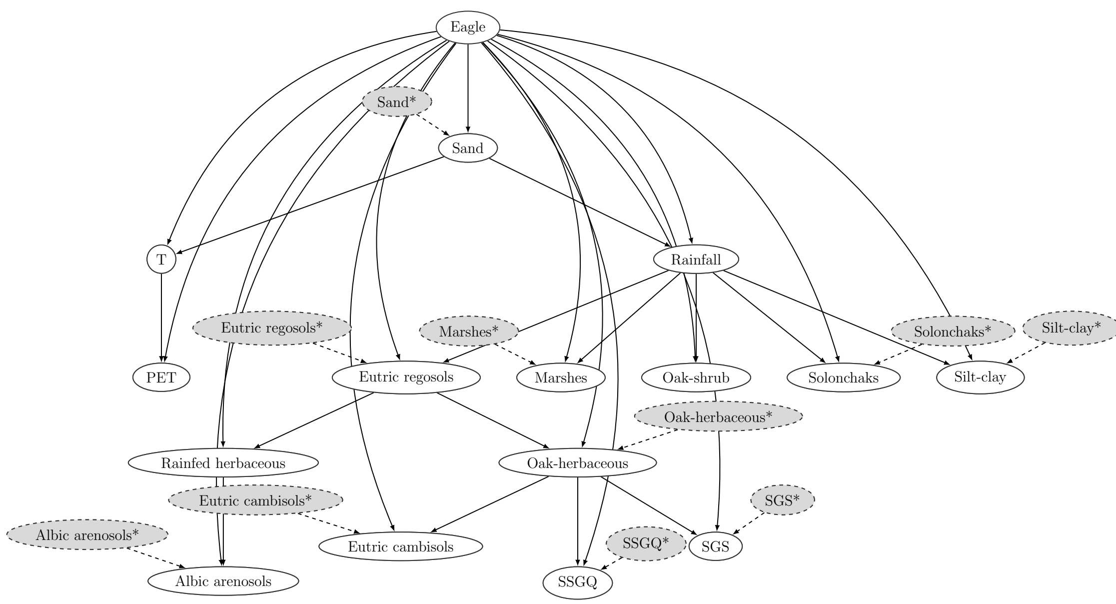


(b) Zi-TAN model

Figure 8: Potential distribution area of salamander, expressed as probability of presence, predicted by TAN (a) and Zi-TAN (b) models. Cells filled in with straight lines represent the observed distribution area of the species.



(a) TAN structure



(b) Zi-TAN structure

Figure 9: Structure of TAN (a) and Zi-TAN (b) models for predicting presence of eagle given the explanatory variables. SSGQ: slate-shale-greywacke-quartzite; SGS: slate-greywacke-sandstone. Shaded nodes with dashed lines represent the artificial binary variables used to model the zero-inflated explanatory variables. Note that the dependence relationships existing between the feature variables just allow the models to perform better.

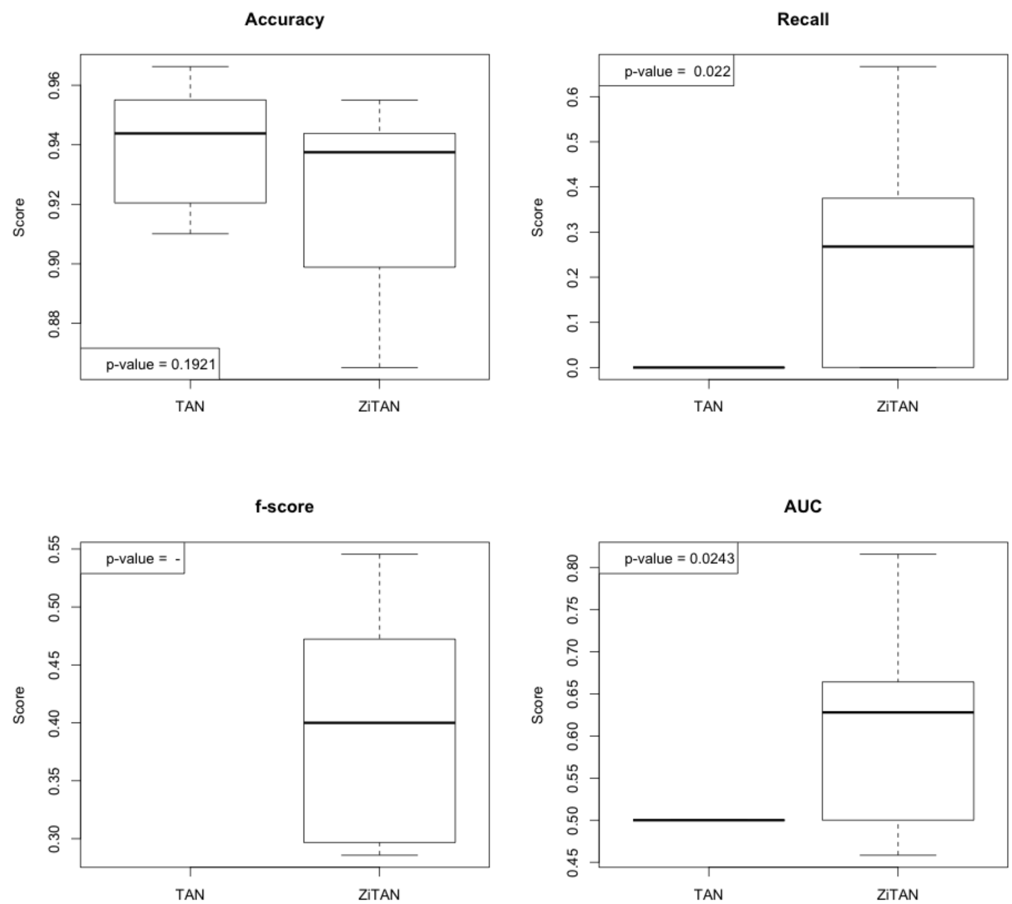
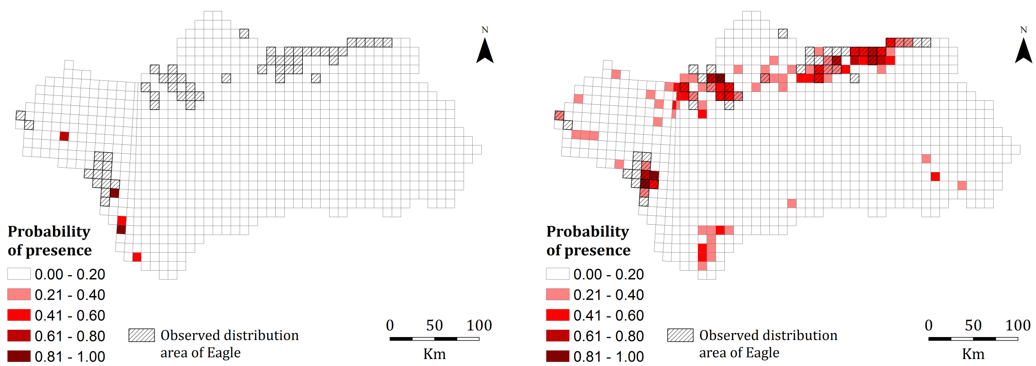


Figure 10: Box plots comparing TAN and Zi-TAN in terms of their performance statistics in the *Aquila adalberti* case. The $p - values$ obtained in the Wilcoxon Test are shown.



(a) TAN model

(b) Zi-TAN model

Figure 11: Potential distribution area of eagle, expressed as probability of presence, predicted by TAN (a) and Zi-TAN (b) models. Cells filled in with straight lines represent the observed distribution area of the species.