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

A.D. Maldonado[1], P.A. Aguilera[2], A. Salmerón[1]

[1] Department of Mathematics, University of Almería

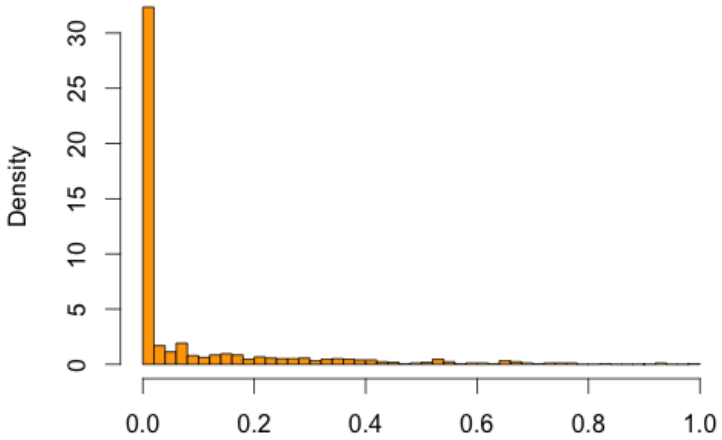
[2] Informatics and Environment Laboratory, Department of Biology and Geology, University of Almería

Probabilistic Graphical Models for Scalable Data Analytics
Granada, 4 February 2016

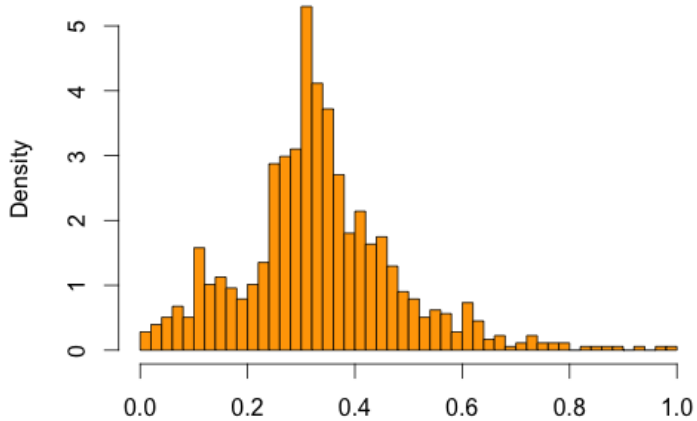
Justification

Depending on the scale and type of variable, environmental datasets may contain a large proportion of zeros.

Zero-inflated variable, e.g. *eutric regosols*



Variable with standard distribution, e.g. *rainfall*



The application of standard analysis techniques may yield inaccurate parameter estimates and misleading inferences.

Background

In the literature, the problem is typically focused on the **dependent variable**, which is modeled depending on its nature and the kind of zero.

Approach	**Type of zero**	**Model**
Mixture models	True and false zeros	Zero-inflated models
Conditional models	True zeros	Hurdle models

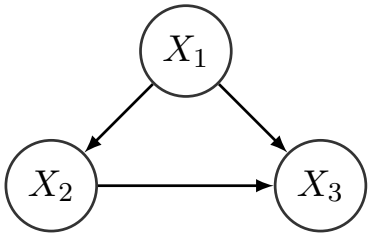
True zero: the species does NOT saturate its entire habitat or the habitat is NOT suitable for the species.
False zero: the species DO occupy the habitat, but it is NOT there at the sampling moment or we failed to detect it.

Objective

Our goal is to improve the predictive power of hybrid Bayesian network classifiers applied to **Species Distribution Models** by explicitly modeling the zero values in **explanatory variables**.
The new model, *zero-inflated tree augmented naive bayes* (ZiTAN), extends the already known *tree augmented naive bayes* (TAN).

Bayesian network classifiers - Brief Introduction

Qualitative component



Nodes = random variables
Edges = direct dependence

Quantitative component

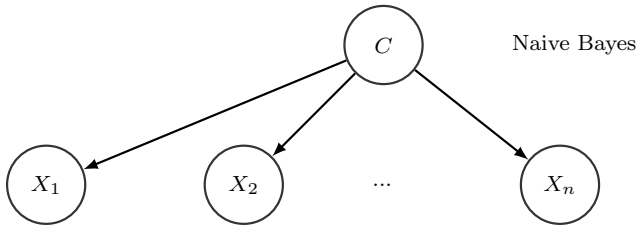
Joint probability distribution

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i | pa(x_i))$$

$$p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)$$

BNs for classification

Discrete variable of interest C and a set of either continuous or discrete explanatory variables $(X_1, \ldots, X_n)$.



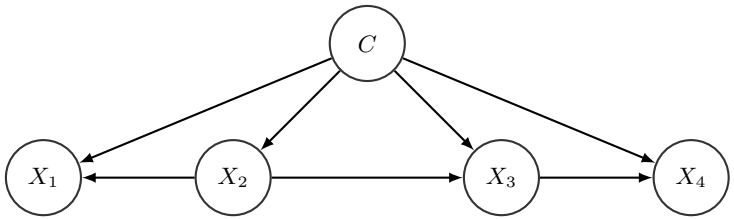
Each observation in the dataset will be classified as belonging to class c^* as

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

The tree augmented naive bayes classifier

Structural learning

Tree augmented naive Bayes



Each feature has one more parent besides C.

The conditional mutual information between the feature variables given C is used to obtain the TAN structure.

Parametric learning

Mixtures of Truncated Exponentials

$$f(z_1, \ldots, z_c) = a_0 + \sum_{i=1}^{m} a_i \exp\left\{\sum_{j=1}^{c} b_i^{(j)} z_j\right\}$$

MTE densities are used to model the distributions in the network. A mixture of potentials is fit in each partition of the variable.

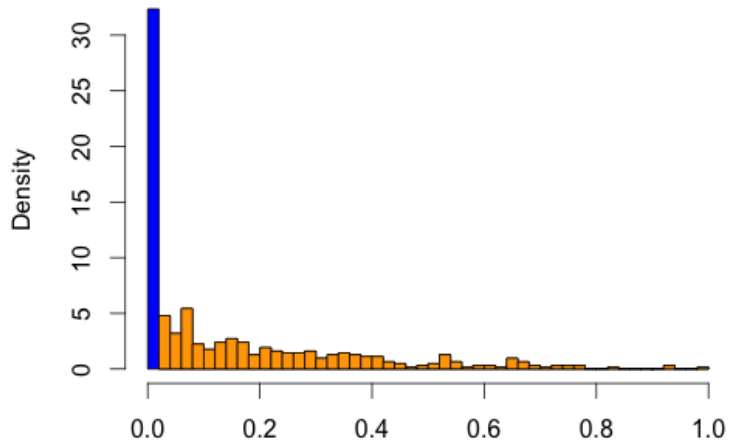
3 intervals were used.

The Zero-inflated variables, X

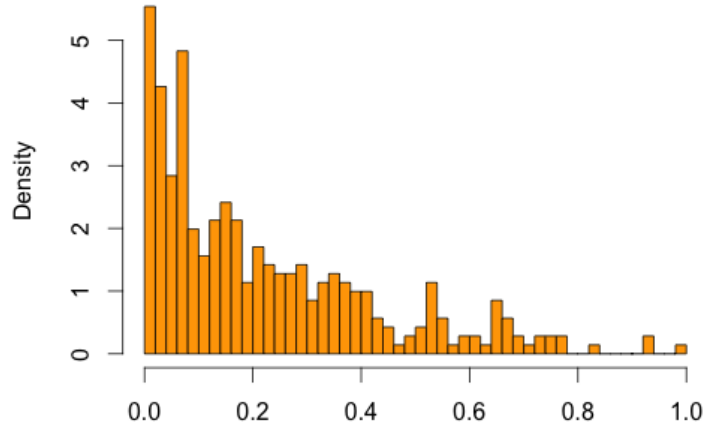
A random variable, X, is a zero-inflated random variable if

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

where $0 < p < 1$,
$g(x) \geq 0$ for $0 < x \leq 1$
and $\int_0^1 g(x)dx = 1 - p$.



p represents the proportion of 0



$g(x)$ corresponds to the positive values of the variable

The artificial variables, X^*

DATASET

X_1	X_2	X_3	X_4	C
0	0.97	0.26	0.51	0
0	0.36	0.83	0.06	1
0	0.12	0.81	0	1
0	0.76	0.07	0	0
0	0.34	0.65	0.2	0
0	0.54	0.21	0	1
0.12	0.34	0.27	0	0
0.05	0.08	0.74	0	1
0	0.13	0.1	0	1
0.43	0.19	0.27	0	0

ZERO-INFLATED DATASET

X_1	X_1^*	X_2	X_3	X_4	X_4^*	C
0	0	0.97	0.26	0.51	1	0
0	0	0.36	0.83	0.06	1	1
0	0	0.12	0.81	0	0	1
0	0	0.76	0.07	0	0	0
0	0	0.34	0.65	0.2	1	0
0	0	0.54	0.21	0	0	1
0.12	1	0.34	0.27	0	0	0
0.05	1	0.08	0.74	0	0	1
0	0	0.13	0.1	0	0	1
0.43	1	0.19	0.27	0	0	0

An artificial binary random variable is defined as

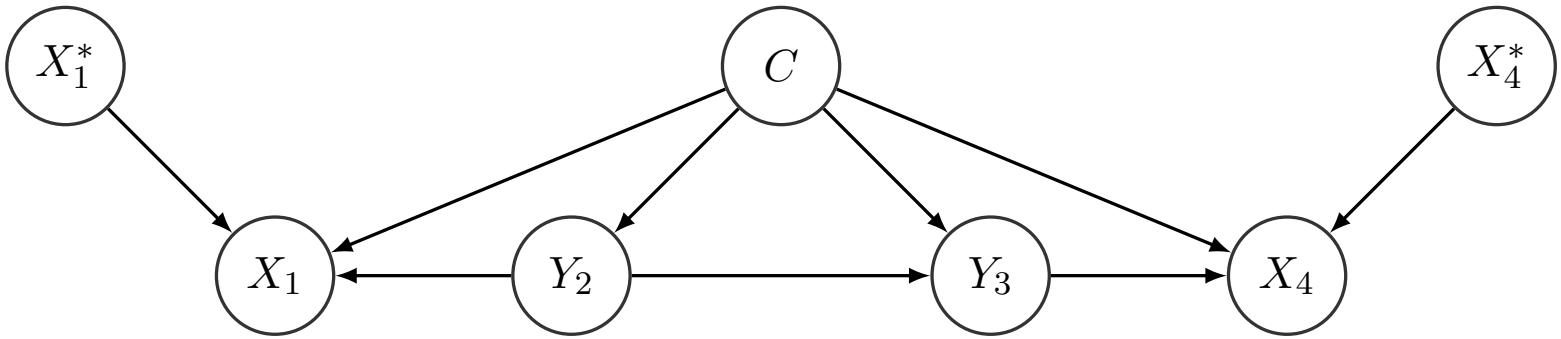
$$X^* = \begin{cases} 0 & \text{if } X = 0 \\ 1 & \text{otherwise} \end{cases},$$

and its probability function is

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

The Zero-inflated TAN classifier

For each zero-inflated random variable, X, in the TAN model, we incorporated an artificial variable, X^*.



For each node X, we attached a new conditional distribution

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

Case study

The aforementioned methodology was applied to **Species Distribution Models** through a case study involving *Salamandra salamandra* and *Aquila adalberti*.

Study area

The study area is Andalusia.

Each cell is a sampling unit.

There are 887 cells.

Data description

Variable	Description
Salamander /Eagle	Presence/absence of the given species in each cell
T (°C)	Average of annual mean temperature for the 30 year period 1971-2000 in each cell
Rainfall (mm)	Average of annual rainfall for the 30 year period 1971-2000 in each cell
PET (mm)	Average of the annual potential evapotranspiration for the 30 year period 1971-2000 in each cell
Humidity index	Average of annual humidity index for the 30 year period 1971-2000 in each cell
Land uses (%)	Percentage of occupation of each land-use (#44) within each cell
Soil (%)	Percentage of occupation of each soil type (#63) within each cell
Lithology (%)	Percentage of occupation of each lithological unit (#41) within each cell
Z (m a.s.l.)	Average elevation of each cell
Slope (%)	Average slope of each cell
Aspect (°)	Average aspect of each cell

number of variables

Variable selection

The variables were selected by experts.

Variable	**Zeros**/887	**Variable**	**Zeros**/887
Salamander	587	**Eagle**	832
Rainfall	-	Rainfall	-
Humidity index	-	Temperature	-
Dense oak	485	Evapotranspiration	-
Oak with shrub	301	Oak with shrub	301
Oak with herbaceous crops	463	Oak with herbaceous crops	463
Woodlands with herb. crops	394	Rainfed herbaceous crops	224
Grassland	220	Marshes	857
Olive groves	268	Albic arenosols	864
Eutric regosols	684	Eutric regosols	684
Calcaric regosols	553	Solonchaks	834
Eutric cambisols	702	Eutric cambisols	702
Sand-silt-clay-gravel	399	Slate-shale-greywacke-quarzite	810
Slate-greywacke-sandstone	726	Slate-greywacke-sandstone	726
Volcano-sedimentary complex	795	Sand	862
-	-	Silt-clay	832

Learning Species Distribution Models

DATASET

Variable 1	Variable 2	Variable 3	Class
97	76	34	0
36	2	6	1
12	43	25	1
76	34	79	0
34	65	62	0
54	87	72	1
34	12	47	0
8	5	45	1
13	81	75	1
19	43	34	0

80%

TRAINING SET

Variable 1	Variable 2	Variable 3	Class
97	76	34	0
36	2	6	1
12	43	25	1
76	34	79	0
34	65	62	0
54	87	72	1
34	12	47	0
8	5	45	1

MODEL LEARNING

PROBABIITY OF PRESENCE OF A GIVEN SPECIES

VALIDATION

20%

TESTING SET

Variable 1	Variable 2	Variable 3	Class
13	81	75	1
19	43	34	0

Model validation

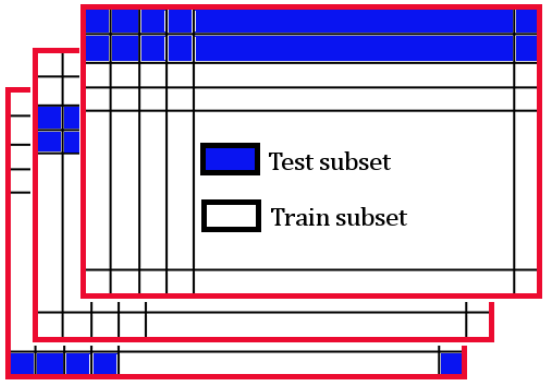
1 Confusion matrix and performance statistics

0: absences and 1: presences.
TP: true presence; TN: true absence; FP: false presence; FN: false absence

		Predicted	
		TAN	
		0	1
Real	0	TN	FP
	1	FN	TP

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

2 10-fold Cross-validation

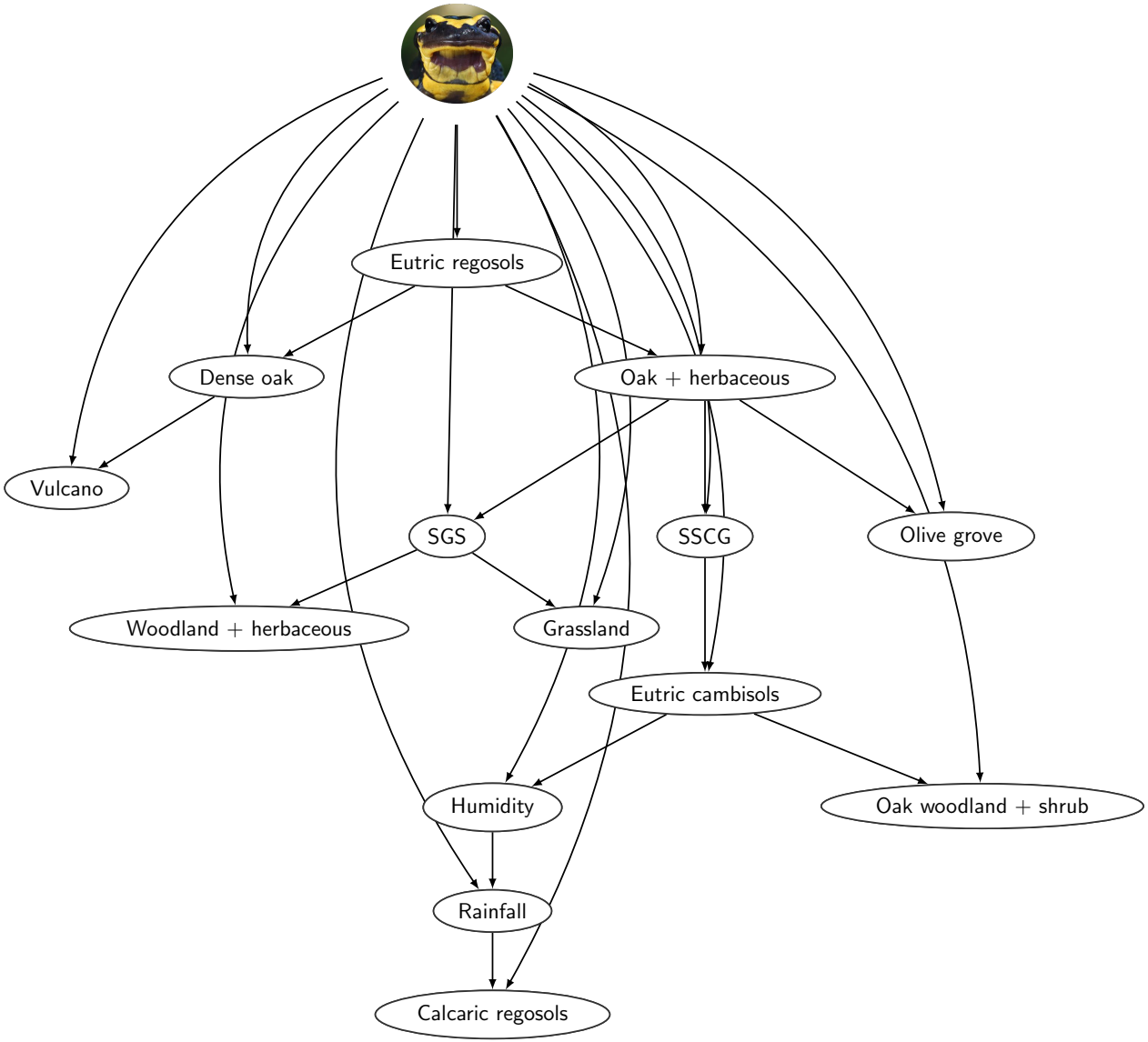


The 10 measures of overall accuracy are compared using Wilcoxon's signed rank Test.

3 Cochran's Q test

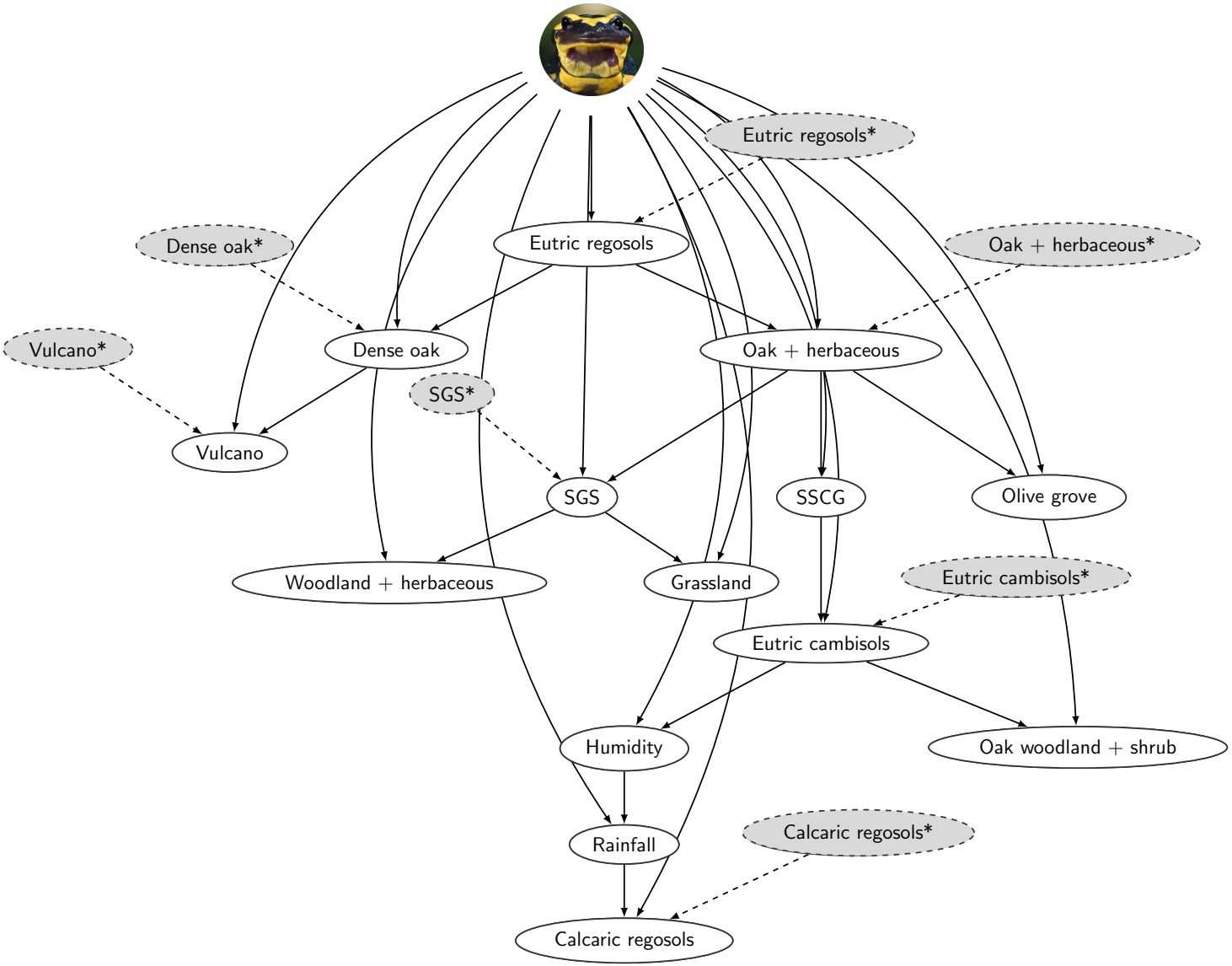
The proportion of presences in the 3 groups are compared. If significant differences are found, pairwise exact McNemar's test is used.

Results - *Salamandra salamandra*



Structure of TAN

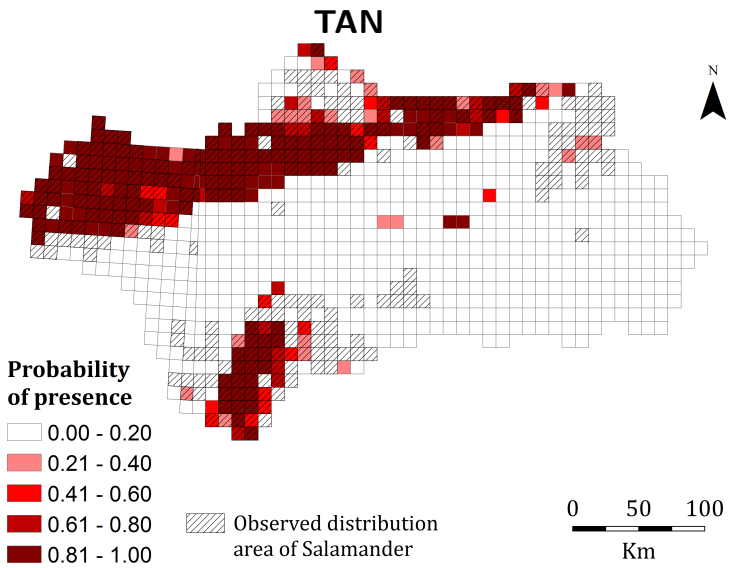
Results - *Salamandra salamandra*

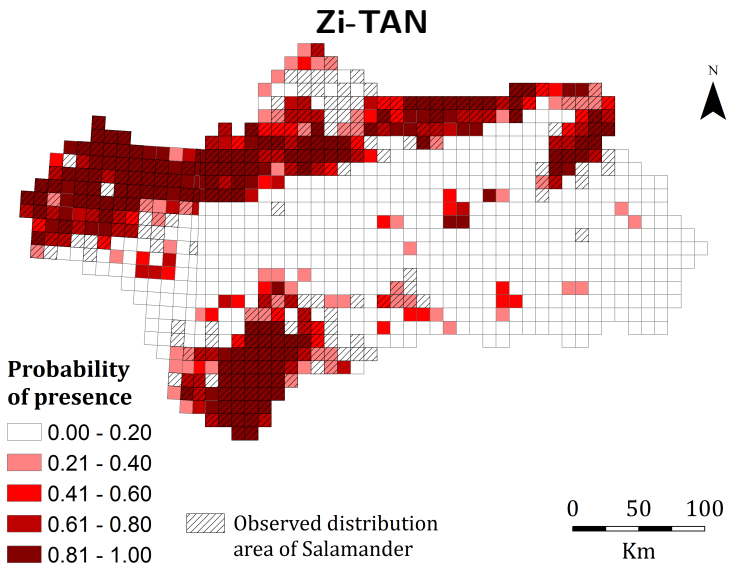


Structure of Zi-TAN

Results - *Salamandra salamandra*

Potential distribution area of Salamander





		Predicted			
		TAN		Zi-TAN	
		0	1	0	1
Real	0	104	5	103	6
	1	32	37	19	50
Accuracy		0.792		0.860	
Recall		0.536		0.725	
f-score		0.667		0.800	
AUC		0.745		0.835	

Wilcoxon Test: $p - value > 0.01$

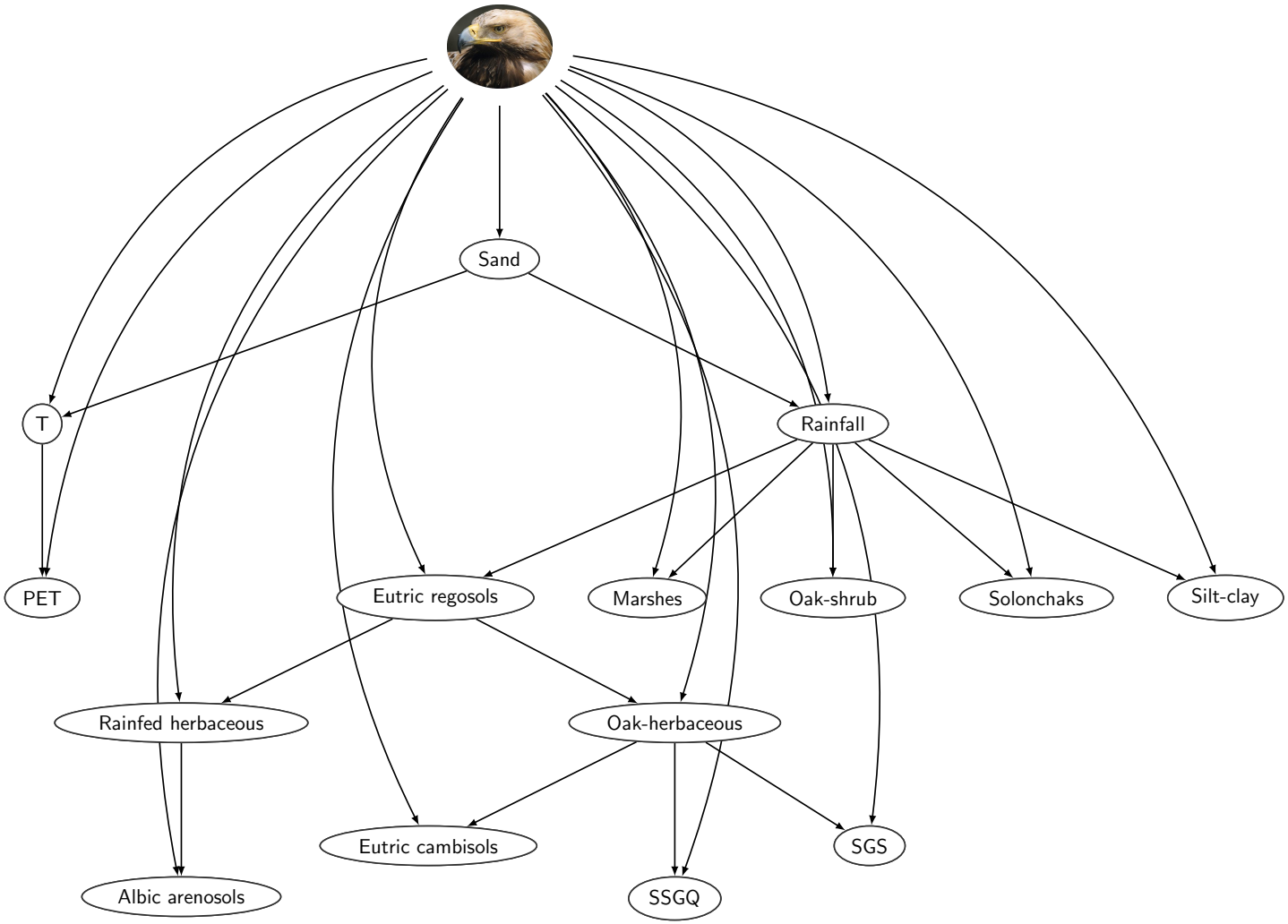
Cochran Test: $p - value < 0.01$

$\downarrow$ Pairwise McNemar Test

Observed - TAN: $p - value$ <0.01

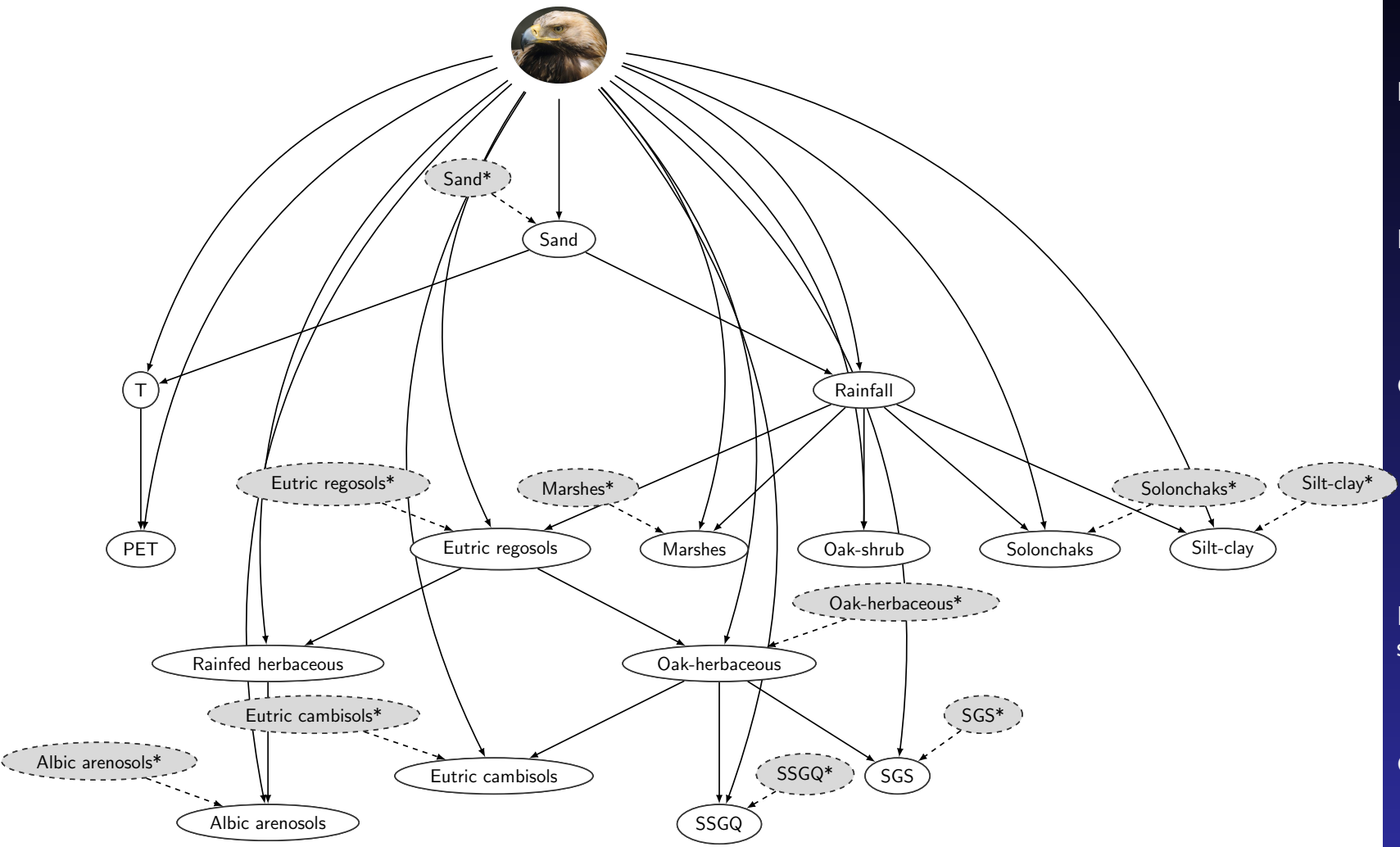
Observed - Zi-TAN: $p - value$ >0.01

Results - *Aquila adalberti*



Structure of TAN

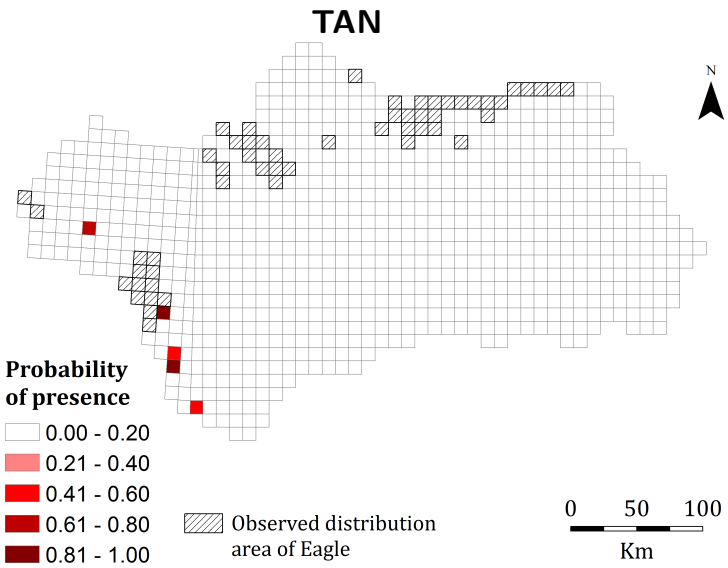
Results - *Aquila adalberti*

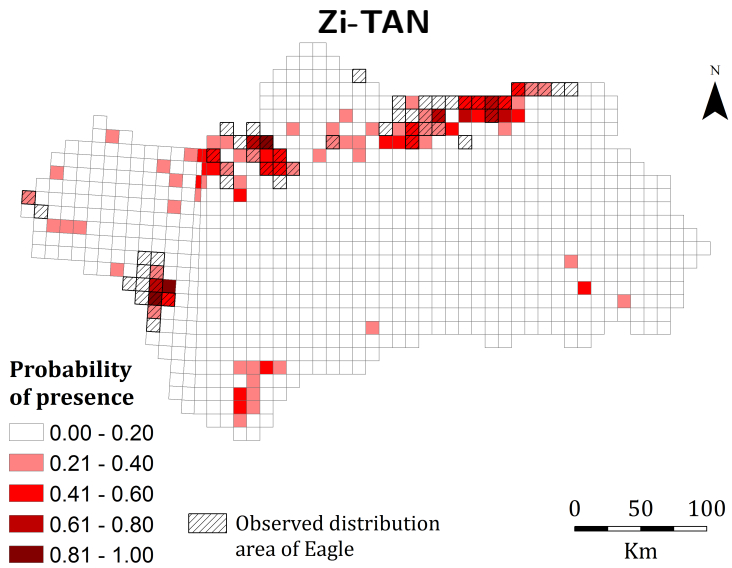


Structure of Zi-TAN

Results - *Aquila adalberti*

Potential distribution area of Eagle





		Predicted			
		TAN		Zi-TAN	
		0	1	0	1
Real	0	165	2	167	0
	1	11	0	8	3
Accuracy		0.927		0.955	
Recall		0		0.273	
f-score		-		0.429	
AUC		0.494		0.636	

Wilcoxon Test: $p - value > 0.01$
Cochran's Test: $p - value < 0.01$

↓ Pairwise McNemar Test

Observed - TAN: $p - value$ <0.01
Observed - Zi-TAN: $p - value$ <0.01

Conclusions

- Modeling zero-inflated feature variables improves the performance of the classifier.
- For **salamander**, a frequent species in the study area, the results given by both classifiers were reasonable.
- For **eagle**, a scarce species in the study area, the Zi-TAN model substantially improved the distribution area predicted by the TAN model.