LETTERS IN BIOMATHEMATICS
An International Journal

Intercollegiate Biomathematics Alliance

RESEARCH ARTICLE

OPEN ACCESS

Expectation of Occupancy for Habitat Quality Research for Threatened and Endangered Avian and Bat Species

Dinesh B. Ekanayake[a], Indrajith Wasala Mudiyanselage[b], Nisansala Wickramasinghe[b]

[a] Department of Mathematics and Philosophy, Western Illinois University, 1 University Circle, Macomb 61455, USA; [b] The Department of Mathematical Sciences, The University of Texas at Dallas, 800 W Campbell Rd, Richardson, TX 75080, USA

ABSTRACT

We introduce a model for the expectation of occupancy as a function of spatial scale variables and variables related to characteristics of sites for avian and bat habitat quality assessment. We utilized a unique functional representation for the joint density based on the estimated modes of the univariate distributions to model the expectation of occupancy. Unlike binary classification methods, the proposed construction does not require a clear distinction of used versus unused habitats. It also allows extending variable ranges for spatial generalization. We demonstrate how one could utilize the expectation of occupancy to reliably predict habitats a particular avian species may choose for nesting or roosting, to compare various tree species and sites, and to identify prominent characteristics that govern the selection of a site by avian and bat species. The methods can be utilized in the management of suitable sites and the conservation of endangered species.

ARTICLE HISTORY
Received April 14, 2023
Accepted October 24, 2023

KEYWORDS
habitat classification, mode-based density estimation, restricted data ranges, feature selection

1 Introduction

Management of suitable habitats that a species can occupy is important for the conservation of threatened and endangered species. Evaluating a habitat for suitability requires an understanding of the expectation of occupancy that quantifies the association between habitat conditions and the selection of the habitat by the focal species. In this paper, we introduce a unique model for the expectation of occupancy of a habitat to compare potential avian nesting or roosting sites or qualitative habitat properties for threatened species. While all the applications of the methods here are related to avian and bat nesting and roosting habitats, one could easily use the proposed methods for any threatened species population.

Various empirical and statistical methods have been utilized in evaluating avian and bat habitat selection (Carter and Feldhamer, 2005; DeBoer and Diamond, 2006; Doherty et al., 2010; Emrick et al., 2010; Keating and Cherry, 2004; Manly et al., 2002; Pauli et al., 2015; Phillips et al., 2006). The common objective of these studies is either to identify the primary characteristics of the nesting or roosting habitats or to model the probability of use of a tree or a forest by the focal species. Habitat selection studies often utilize location information and resource availability of the sites that are used versus available (or not used) to assess which habitat characteristics are important (Boyce et al., 2002; Emrick et al., 2010; Kroll, 1980). Such analyses are essential in modeling species habitats and the likelihood for occupancy. The most important conservation and management implications of modeling the probability of use is to reliably predict the habitats that a particular avian species may choose for nesting or roosting. However, there are various challenges that impact the reliability of models that characterize the habitats and the likelihood of occupancy for threatened and endangered species:

1. Location and tree characteristics can create a relatively large feature space that requires a large number of observations to generate a reliable probability function. Yet data for available habitats can be limited when the species population is minimal (Aldridge et al., 2012; Emrick et al., 2010; Schroder et al., 2017).

2. Many factors may govern avian and bat habitat selection, including food abundance (Burke and Nol, 1998; Kusch et al., 2004; Livingston et al., 1990; Verner and Willson, 1966), proximity to water sources (Andrew and Mosher, 1982; Mundahl

CONTACT Dinesh B. Ekanayake db-ekanayake@wiu.edu

et al., 2013), protection from predators (Kelly, 1993; Lima, 2009; Martin, 1993; Quinn et al., 2003), ease of defense (Bernstein and McLean, 1980; Sonerud, 1985), and offspring survival (Gibson et al., 2016; Newlon and Saab, 2011; Willson, 1966). Many spatial scale variables and habitat characteristics may support several of these factors, making it difficult to distinguish the variables that are most important for predicting site selection.

3. Data sets can be burdened with correlations among variables (Battin and Lawler, 2006; Lichstein et al., 2002; Schroder et al., 2017). Spatial patterns, stand age, and fragmentation in the landscape create correlated variables. For example, an area measure can be highly correlated with distance to the forest and water boundaries; tree size can be correlated with stand age; canopy cover can be correlated with surrounding trees' status as live versus dead; and distance from the habitat to water bodies can be correlated with distance between water bodies.

4. Most of the data ranges are restricted by region-dependent geographical, biophysical, or ecological constraints, limiting prediction reliability over spatial distributions. Examples include: distance measures, restricted by landscape features; area measures, restricted by fragmentation; and tree characteristic measures, restricted by the tree species. Spatial generalization of the model may be essential when the prevalence of the species is minimal.

5. Habitat classifications are usually based on estimating a resource selection function (RSF) using data from locations that are used by the species versus those that are either unused or available (Manly et al., 2002). Effective classification by RSF models derived from used versus available habitats requires a clear distinction of the characteristics of each of the two categories so that one can estimate the probability of use given the data. While used locations are always confirmed, many sites may be unused only because of the scant population size of a threatened or endangered species, and the possibility that many available sites may be rather classified as used in the presence of a larger population may impact the accuracy of the classification.

We introduce a model for the expectation of use, using mode-based density estimations, to address the challenges that result from the various aforementioned limitations. The proposed methods reduce redundancy and overfitting, expand model usefulness for spatial generalization and avoid the need of clear distinction between used versus unused or available data sets.

In avian habitat studies, variables are chosen either because they have already been utilized in previous literature or based on hypotheses relating species abundance and the ecology of the species. However, because of the existence of correlated and irrelevant features, not all data sets can support a meaningful integration of the chosen variables in a model for expectation of occupancy. The difficulty of distinguishing unsuitable habitats and the infrequent use of sites by the focal species limit the usefulness of many existing subset selection methods that are based solely on relevance to the response. We model the expectation of occupancy by utilizing a variable set that reduces redundancy. In the example section, we demonstrate that the minimum redundancy maximum relevance (MRMR) algorithm (Ding and Peng, 2005) is a good fit to identify model variables. In some cases, only the locations that were used by the threatened or endangered species population are available (presence-only data points; see Pauli et al., 2015; Hammond et al., 2016; Schroder et al., 2017 for some examples), for which case the MRMR algorithm is not applicable. Here we introduce a variable filter based on presence-only data points alone. The proposed filter algorithm follows the minimum description length principle (Rissanen, 1978), in which the best model for a data set is the one that provides the best compression of the data set. Subsequently, the expectation of occupancy can be effectively modeled without explicit variable selection analysis. Using both simulated and real data sets, we demonstrate the fitting performance of the algorithm for the task.

We combine presence-only data points with data from random locations to model the expectation of occupancy (given the variables). The conditional expectation is the regression function for Bayes classifiers. When the features are conditionally independent, Bayes classifiers minimize the probability of incorrect binary classifications (Devroye et al., 1996). As we demonstrate in Section 3, the proposed method is capable of performing better than logistic regression models when comparing used versus random habitats. The most direct and functional application of the proposed method is for micro-scale investigations. For example, the expectation of occupancy can be utilized for statistical inferences in identifying preferred tree species for nesting habitats, evaluating the impact of forest fragmentation or human developments, and evaluating preference for specific tree characteristics for maternity roosts.

If data ranges are restricted by geographical or ecological constraints, data may not be available for a large portion of the domain. Such data sets do not support the identification of the population mean or median. We propose a unique representation of the joint density of the site characteristic and spatial scale variables, based on the estimated modes of the univariate distributions. The underlying assumption is that the value at which the probability density function has a maximum (the mode of the distribution) is within the restricted data range and can be obtained from the empirical density function. The proposed approximations are robust for data sets with limited ranges and allows extending the variable ranges for predictions. Here we present the construction, related theory, and applications that can be useful in the management of suitable sites and the conservation of endangered species.

2 Model

In this section we discuss the construction of the model for the expectation of occupancy. We denote the variables associated with the habitats of interest by X_i, $i = 1,\dots,n$, and the low dimensional subset of characteristics containing the variables of greatest importance by the random vector $\mathbf{X} = [X_1, X_2, \dots, X_k]^T$, where superscript T denotes the transpose and $k \leq n$. For a given region, suppose $S \in \{0,1\}$ represents the use (no = 0, yes = 1) of a site by the focal species. We write $\mathbb{E}[S \mid \mathbf{X} = \mathbf{x}]$ to represent the expectation of occupancy given $\mathbf{X} = \mathbf{x}$ for some $\mathbf{x} \in \mathbb{R}^k$. We have the following result.

Lemma 2.1. *Suppose $S \sim$ Bernoulli(p) and $\mathbf{X} = [X_1, X_2, \dots, X_k]^T$. Then*

$$\mathbb{E}[S \mid \mathbf{X} = \mathbf{x}] = p\frac{f_{\mathbf{X}\mid S}(\mathbf{x} \mid 1)}{f_{\mathbf{X}}(\mathbf{x})}, \tag{1}$$

where $\mathbf{x} = [x_1, \dots, x_k]^T$, p is a constant, $f_{\mathbf{X}}(\mathbf{x})$ is the probability distribution function of $\mathbf{X}$, and $f_{\mathbf{X}\mid S}(\mathbf{x}\mid 1)$ is the conditional joint probability distribution of $\mathbf{x}$ given $S = 1$.

Proof. Let $\mathcal{R}_{\curlywedge}(\mathbf{x}) = \prod_{i=1}^{k}[0, x_i] \subset \mathbb{R}_+^k$ be the nonnegative k-orthotope generated by the Cartesian product of k intervals $[0, x_i]$, $i = 1,\dots,k$. We may describe the joint distribution of $\mathbf{X}$ and S by $\mathbb{P}(S = s, \mathbf{X} \leq \mathbf{x}) = \int_{\mathcal{R}_{\curlywedge}(\mathbf{x})} f_{S,\mathbf{X}}(s,\mathbf{t})\,dt$, where $\mathbf{x} \geq 0$, $\mathbf{t} = [t_1, \dots, t_k]^T$, $dt = dt_1 dt_2 \dots dt_k$ is a hypervolume element in $\mathbb{R}^k$, and $f_{S,\mathbf{X}}$ is the density of the mixed distribution. Then $p_S(1) = \mathbb{P}(S = 1) = \int_{\mathbb{R}_+^k} f_{S,\mathbf{X}}(1,\mathbf{t})\,dt$ and the probability density of $\mathbf{X}$ is $f_{\mathbf{X}}(\mathbf{x}) = f_{S,\mathbf{X}}(0,\mathbf{x}) + f_{S,\mathbf{X}}(1,\mathbf{x})$. If $p_s(1) > 0$, $f_{\mathbf{X}|S}(\mathbf{x}|1) = f_{S,\mathbf{X}}(1,\mathbf{x})/p_S(1)$, and if $f_{\mathbf{X}}(\mathbf{x}) > 0$, $p_{S|\mathbf{X}}(1,\mathbf{x}) = f_{S,\mathbf{X}}(1,\mathbf{x})/f_{\mathbf{X}}(\mathbf{x})$. Subsequently, $p_{S|\mathbf{X}}(1|\mathbf{x}) = f_{\mathbf{X}|S}(\mathbf{x}|1)p_S(1)/f_{\mathbf{X}}(\mathbf{x})$ and hence, $\mathbb{E}[S|\mathbf{X} = \mathbf{x}] = p_{S|\mathbf{X}}(1|\mathbf{x}) = p\dfrac{f_{\mathbf{X}|S}(\mathbf{x}|1)}{f_{\mathbf{X}}(\mathbf{x})}$. □

Next we model $\mathbb{E}[S \mid \mathbf{X} = \mathbf{x}]$ based on unimodal smoothing approximations for the univariate probability density function of each variable X_i using

$$\phi(x) \triangleq \frac{\zeta e^{-4\alpha(x-\gamma)}}{\left(1 + e^{-\beta(x-\gamma)}\right)^4}, \quad \alpha/\beta \in (0,1). \tag{2}$$

Lemma 2.2. *Let $\{x_j\}_{j=1}^{m}$, $m \geq 7$, be a distinct observation set of the random variable X arranged in ascending order, $\boldsymbol{x} = \{x_j\}_{j=4}^{m-3}$ and $\boldsymbol{y} = \{y_j\}_{j=4}^{m-3}$, $y_j = 28/n(3x_{j+3}+2x_{j+2}+x_{j+1}-x_{j-1}-2x_{j-2}-3x_{j-3})$. If $(\zeta,\alpha,\beta,\gamma) = \arg\min \left\|\phi(\boldsymbol{x}) - \boldsymbol{y}\right\|_2$ subject to $\alpha/\beta \in (0,1)$, then ϕ is a unimodel approximation for the empirical probability density function of X.*

Proof. Define $\eta \triangleq \alpha/\beta$. Since $\phi(x) \geq 0$ and $0 < \int_0^\infty \phi(\tau)\,d\tau \leq \int_{-\infty}^{\infty} \dfrac{\zeta e^{-4\alpha(\tau-\gamma)}}{1 + e^{-4\beta(\tau-\gamma)}}\,d\tau = \dfrac{\zeta}{4\beta}\Gamma(\eta)\,\Gamma(1-\eta) < \infty$, by choosing ζ appropriately, the function $\phi(x)$ is well suited as a density function for $\eta \in (0,1)$, where Γ represents the gamma function. For $\eta \in (0,1)$, $\phi'(x)$ has one zero, $\max \phi(x) = \zeta(1-\eta)^{4(1-\eta)}\eta^{4\eta}$ at $x = \gamma + \ln\left(\frac{1-\eta}{\eta}\right)^{1/\beta}$ and $\phi(x)$ is unimodel. Suppose $f \in C^1(\Omega)$ is a strictly increasing smoothing approximation for the empirical distribution on S_j such that $f(x_j) = j/n$, where Ω is an open set containing $[x_{j-3}, x_{j+3}]$. Then we have $f'(x) > 0$ for all $x \in \Omega$. Let J be the image of Ω under f. From the inverse function theorem, function f has an inverse $g\colon J \to \Omega$ such that $g \in C^1(J)$ with derivative $g'(s) = \dfrac{1}{f'(g(s))}$, $s \in J$. Here we let $f(x)$ be linear and, using the slope of the simple linear regression equation $g(s) = ms + b$, we find

$$y_j = f'(x_j) = \frac{1}{m} = \frac{\sum_{p=-3}^{3}(p/n)^2}{\sum_{p=-3}^{3} j\left(x_{j+p} - \overline{x}\right)/n}$$

where $\overline{x}$ is the mean of S_j. □

With an appropriate scaling, function $\phi(x)$ approximates well many distributions within the exponential family (see Figure 1).

Suppose $X_1, X_2, \dots, X_k$ are independent. Let $\phi_{X_i\mid S}(x_i \mid 1)$ and $\phi_{X_i}(x_i)$ be the probability density approximations of X_i using data for the locations selected by the focal species (presence-only data) and for available locations, respectively. Then $f_{\mathbf{X}\mid S}(\mathbf{x} \mid 1) \approx \prod_{i=1}^{k}\phi_{X_i\mid S}(x_i \mid 1)$ and $f_{\mathbf{X}}(\mathbf{x}) \approx \prod_{i=1}^{k}\phi_{X_i}(x_i)$. Suppose μ_{X_i} and $\mu_{X_i\mid S}$ are the sample means for X_i and $X_i \mid S$, respectively, and $\mu_{X_i} < \mu_{X_i\mid S}$. Let

$$\mathbb{E}_i = a_i\frac{\phi_{X_i\mid S}(x_i \mid 1)}{\phi_{X_i}(x_i)}\mathbf{1}_{[0,m_i)}(x_i) + \mathbf{1}_{[m_i,\infty)}(x_i),$$

where m_i is the mode of the unimodal approximation $\phi_{X_i|S}$, $\mathbf{1}$ is the indicator function, and a_i is chosen such that $\mathbb{E}_i$ is continuous. If $\mu_{X_i} \geq \mu_{X_i \mid S}$, we set

$$\mathbb{E}_i = \mathbf{1}_{[0,m_i)}(x_i) + a_i \frac{\phi_{X_i \mid S}(x_i \mid 1)}{\phi_{X_i}(x_i)} \mathbf{1}_{[m_i,\infty)}(x_i).$$

Figure 2 represents two examples.

We obtain a model for the expectation of occupancy by $\mathbb{E}[S \mid \mathbf{X} = \mathbf{x}] = \prod_{i=1}^{k} \mathbb{E}_i$. For each X_i, suppose the value at which the probability density function has a maximum (the mode of the distribution) is within the data range. Then any range restriction beyond the mode does not change the expectation of occupancy. If we were to include features that do not have much relevance to site selection, then the corresponding ratio $\phi_{X_i \mid S}(x_i \mid 1)/\phi_{X_i}(x_i)$ would be closer to one, limiting their impact on the expectation of occupancy.

Consider the binary classification problem of identifying suitable habitats for conservation. The primary goal is to minimize the error of incorrectly failing to identify the suitable habitats (false negative). We identify the decision boundary for $\mathbb{E}[S \mid \mathbf{X} = \mathbf{x}]$ by finding a value q for which at least $100\pi\%$ of the presence-only data points have the label $S = 1$. Suppose for the presence-only data set, $\arg\min_{q \in [0,1]} \mathbb{P}\left(\mathbb{E}[S \mid \mathbf{X}] \geq q\right) \geq \pi$, where $\mathbb{P}$ represents the probability. The corresponding classifier $\mathcal{CL}$ is the function given by

$$\mathcal{CL}(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbb{E}[S \mid \mathbf{X} = \mathbf{x}] > q, \\ 0 & \text{if } \mathbb{E}[S \mid \mathbf{X} = \mathbf{x}] \leq q. \end{cases} \tag{3}$$

With the proposed classifier $\mathcal{CL}$, we set the false negative rate to at most $1 - \pi$ for a presence-only data set.

2.1 Variable filter

The model for $\mathbb{E}[S \mid \mathbf{X} = \mathbf{x}]$ uses an uncorrelated variable set $\mathbf{X} = [X_1, X_2, \ldots, X_k]^T$ from the variables X_i, $i = 1, \ldots, n$, associated with the habitats of interest. Unknown class labels for the available data points make it difficult to employ a variable selection algorithm. When the proportion of $S = 1$ labels is expected to be small, we may utilize a robust variable selection method such as lasso regularization (Tibshirani, 1996). Otherwise, we may use an algorithm to reduce redundancy among the variables, such as the MRMR algorithm (Ding and Peng, 2005). However, for the proposed model, the inclusion of irrelevant features has minimal impact, circumventing the need for a variable selection algorithm. Here we propose a variable filter based on reducing redundancy, which is particularly useful for presence-only data sets (see Section 3). The filter construction is similar to a rank-revealing QR factorization, which is useful in determining the rank of a matrix (Chan, 1987; Gu and Eisenstat, 1996). Suppose $m \times n$ matrix $\mathbf{A}$ represents a standardized (z-scores calculated along each column) presence-only data set with n variables and m observations, where $n < m$. We find a column permutation matrix $\mathbf{\Pi}$ such that $\mathbf{A\Pi} = [\mathbf{A}_1, \mathbf{A}_2]$, where $\mathbf{A}_1$ represents important and linearly independent features and $\mathbf{A}_2$ represents features to be discarded. We seek to find an appropriate subset of variables by minimizing the distance between the subset data matrix $\mathbf{A}_1 \in \mathbb{R}^{m \times k}$ and the standardized data matrix $\mathbf{A}$ (defined as $\min_{\mathbf{v}} \|\mathbf{A}_1\mathbf{v} - \mathbf{A}\|_F$, where $\mathbf{v}$ is any $k \times n$ real matrix, and F represents the Frobenius norm) so that $\mathbf{A}_1$ captures a significant amount of information from $\mathbf{A}$. Suppose the QR decomposition of $\mathbf{A}$ is given by

$$\mathbf{A\Pi} = [\mathbf{A}_1, \mathbf{A}_2] = \mathbf{QR} = [\mathbf{Q}_1, \mathbf{Q}_2] \begin{bmatrix} \mathbf{A_k} & \mathbf{B_k} \\ 0 & \mathbf{C_k} \end{bmatrix}, \tag{4}$$

where $\mathbf{Q}_1 \in \mathbb{R}^{m \times k}$, $\mathbf{Q}_2 \in \mathbb{R}^{m \times (m-k)}$, matrix $\mathbf{A}_k \in \mathbb{R}^{k \times k}$ is upper triangular with nonnegative diagonal elements, $\mathbf{B}_k \in \mathbb{R}^{k \times (n-k)}$, and $\mathbf{C}_k \in \mathbb{R}^{(m-k) \times (n-k)}$.

Lemma 2.3. *The minimizer for* $\min_{\mathbf{v}} \|\mathbf{A}_1\mathbf{v} - \mathbf{A}\|_F$ *is given by* $\mathit{Trace}(\mathbf{C}_k^T\mathbf{C}_k) = \|\mathbf{C}_k\|_F^2$.

Proof. Suppose $\widehat{\mathbf{v}} = \mathbf{A}_1^+\mathbf{A}$, where $\mathbf{A}_1^+$ is the Moore-Penrose inverse of A_1. Then

$$\begin{aligned} \|\mathbf{A}_1\mathbf{v} - \mathbf{A}\|_F^2 &= \|\mathbf{A}_1\mathbf{v} - \mathbf{A}_1\widehat{\mathbf{v}} + \mathbf{A}_1\widehat{\mathbf{v}} - \mathbf{A}\|_F^2 \\ &= \|\mathbf{A}_1\widehat{\mathbf{v}} - \mathbf{A}\|_F^2 + \|\mathbf{A}_1(\mathbf{v} - \widehat{\mathbf{v}})\|_F^2 + \mathrm{Trace}\left((\mathbf{A}_1(\mathbf{v} - \widehat{\mathbf{v}}))^T (\mathbf{A}_1\widehat{\mathbf{v}} - \mathbf{A})\right) \\ &= \|\mathbf{A}_1\widehat{\mathbf{v}} - \mathbf{A}\|_F^2 + \|\mathbf{A}_1(\mathbf{v} - \widehat{\mathbf{v}})\|_F^2 + \mathrm{Trace}\left((\mathbf{v} - \widehat{\mathbf{v}})^T \left((\mathbf{A}_1\mathbf{A}_1^+\mathbf{A}_1)^T \mathbf{A} - \mathbf{A}_1^T\mathbf{A}\right)\right) \\ &= \|\mathbf{A}_1\widehat{\mathbf{v}} - \mathbf{A}\|_F^2 + \|\mathbf{A}_1(\mathbf{v} - \widehat{\mathbf{v}})\|_F^2. \end{aligned}$$

We have $\|\mathbf{A}_1\mathbf{v} - \mathbf{A}\|_F \geq \|\mathbf{A}_1\widehat{\mathbf{v}} - \mathbf{A}\|_F$ and the equality holds if and only if $\mathbf{v} = \widehat{\mathbf{v}} = \mathbf{A}_1^+\mathbf{A}$. Therefore, the minimizer for $\min_{\mathbf{v}} \|\mathbf{A}_1\mathbf{v} - \mathbf{A}\|_F$ is given by $\mathbf{v} = \mathbf{A}_1^+\mathbf{A}$. From Equation 4, $\mathbf{A}_1 = \mathbf{Q}_1\mathbf{A}_k$ and $\mathbf{A}_2 = \mathbf{Q}_1\mathbf{B}_k + \mathbf{Q}_2\mathbf{C}_k$. Since $\mathbf{Q}$ is orthogonal,

$$\mathbf{I}_m = \mathbf{Q}^T\mathbf{Q} = [\mathbf{Q}_1, \mathbf{Q}_2]^T [\mathbf{Q}_1, \mathbf{Q}_2] = \begin{bmatrix} \mathbf{Q}_1^T\mathbf{Q}_1 & \mathbf{Q}_1^T\mathbf{Q}_2 \\ \mathbf{Q}_2^T\mathbf{Q}_1 & \mathbf{Q}_2^T\mathbf{Q}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{m-k} \end{bmatrix},$$

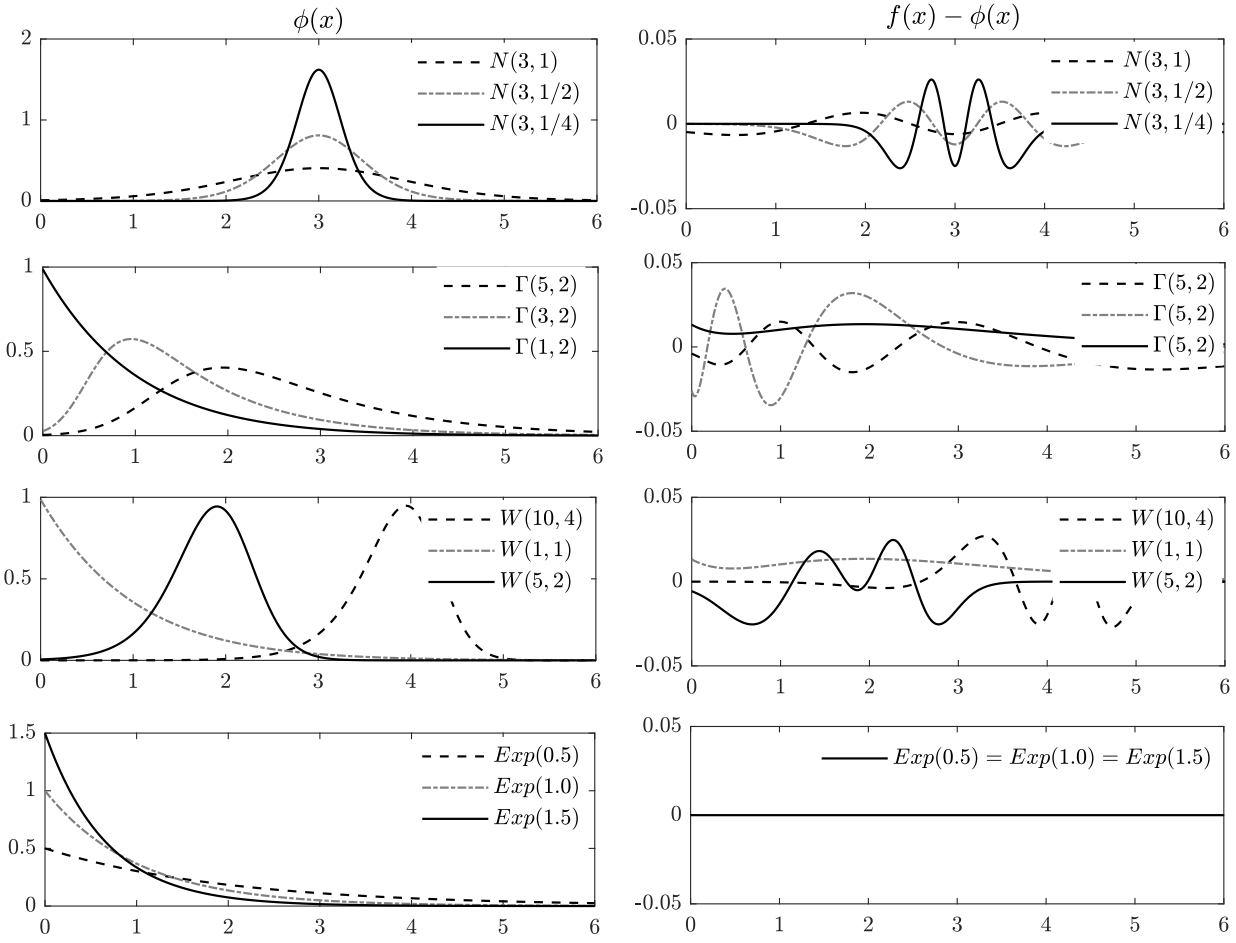


Figure 1: Approximations of distributions in the exponential family by $\phi(x)$. In the left column, $N(\mu,\sigma)$, $\Gamma(k,\theta)$, $W(k,\lambda)$, and $Exp(\lambda)$ represent the approximations $\phi(x)$ for the normal, gamma, Weibull and exponential distributions, respectively; in the right column, they represent the error $f(x) - \phi(x)$, where $f(x)$ denotes the true distribution. The absolute error in the approximations is less than 0.05 for all the distributions.

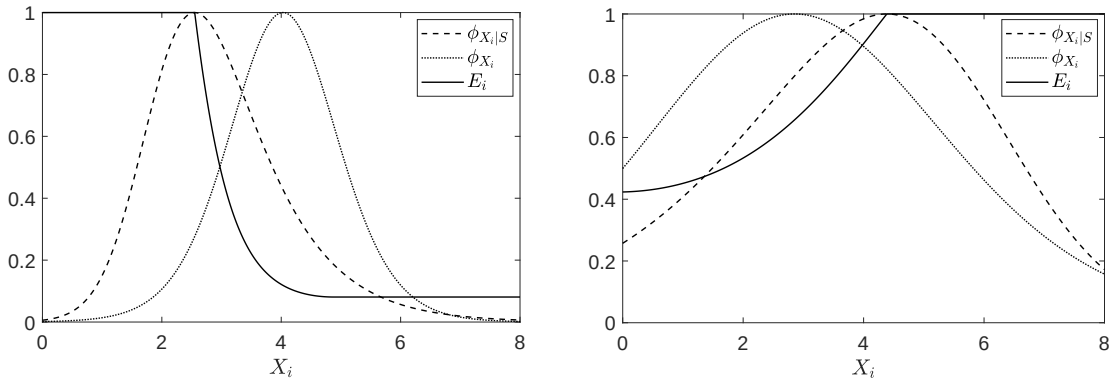


Figure 2: Two examples of $\mathbb{E}_i$. Here $\phi_{X_i \mid S}(x_i \mid 1)$ and $\phi_{X_i}(x_i)$ are scaled to be one at the respective modes.

where $\mathbf{I}_m$ and $\mathbf{I}_{m-k}$ are $m \times m$ and $(m-k) \times (m-k)$ identity matrices, respectively. Assuming that $\mathbf{A}_1$ has full column rank, $\mathbf{A}_k$ is invertible, and hence, $\mathbf{A}_1^+ = \left(\mathbf{A}_1^T\mathbf{A}_1\right)^{-1}\mathbf{A}_1^T = \left(\mathbf{A}_k^T\mathbf{Q}_1{}^T\mathbf{Q}_1\mathbf{A}_k\right)^{-1}\mathbf{A}_k^T\mathbf{Q}_1{}^T = \left(\mathbf{A}_k^T\mathbf{A}_k\right)^{-1}\mathbf{A}_k^T\mathbf{Q}_1{}^T = \mathbf{A}_k^{-1}\mathbf{Q}_1{}^T$. Using the fact that $\|\mathbf{A}\|_F^2 = \sum_j \sum_i |a_{ij}|^2$,

$$\begin{aligned}
\min_{\mathbf{v}} \|\mathbf{A}_1\mathbf{v} - \mathbf{A}\|_F^2 &= \|\mathbf{Q}_1\mathbf{A}_k\mathbf{A}_k^{-1}\mathbf{Q}_1{}^T\mathbf{A} - \mathbf{A}\|_F^2 \\
&= \left\|(\mathbf{Q}_1\mathbf{Q}_1{}^T - \mathbf{I}_m)\left[\mathbf{Q}_1\mathbf{A}_k,\ \mathbf{Q}_1\mathbf{B}_k + \mathbf{Q}_2\mathbf{C}_k\right]\right\|_F^2 \\
&= \left\|\left[\mathbf{Q}_1\mathbf{Q}_1{}^T\mathbf{Q}_1\mathbf{A}_k - \mathbf{Q}_1\mathbf{A}_k,\ \mathbf{Q}_1\mathbf{Q}_1{}^T\mathbf{Q}_1\mathbf{B}_k - \mathbf{Q}_1\mathbf{B}_k + \mathbf{Q}_1\mathbf{Q}_1{}^T\mathbf{Q}_2\mathbf{C}_k - \mathbf{Q}_2\mathbf{C}_k\right]\right\|_F^2 \\
&= \|\mathbf{Q}_2\mathbf{C}_k\|_F^2 = \operatorname{trace}(\mathbf{C}_k^T\mathbf{Q}_2{}^T\mathbf{Q}_2\mathbf{C}_k) = \operatorname{trace}(\mathbf{C}_k^T\mathbf{C}_k) = \|\mathbf{C}_k\|_F^2.
\end{aligned}$$

□

To identify $\mathbf{A}_1$, we find the minimum k that satisfies the inequality

$$\text{Trace}\left(\mathbf{C}_\mathbf{k}^\mathbf{T}\mathbf{C}_\mathbf{k}\right) \le k(m-1)\frac{1-t^2}{t^2}, \quad t \in [0.75, 1]. \tag{5}$$

We then select as $\mathbf{A}_1$ the first k columns of $\mathbf{A\Pi}$. Since $\mathbf{A}$ is standardized along each column, we have $\|\mathbf{A}_1\|_F^2 = k(m-1)$ and $\|\mathbf{A}\|_F^2 = n(m-1)$. From inequality (5), we obtain $\|\mathbf{C}_k\|_F \le \sqrt{k\tilde{t}/n}\|\mathbf{A}\|_F, \tilde{t} = (1-t^2)/t^2$. Then $\dfrac{\|\mathbf{A}_1\|_F}{\|\mathbf{R}\|_F} = \dfrac{\|\mathbf{A}_k\|_F}{\|\mathbf{A}_k\|_F + \|\mathbf{C}_k\|_F} \ge t$. We select the minimum number of variables that, when combined, account for 100t% or more of the Frobenius norm of the upper triangular matrix $\mathbf{R}$. If we utilize both presence-only data points and data from random locations, then we may choose a value of t closer to one. If the data set is limited to presence-only points, we choose a t closer to 0.75 to remove as many redundant features as possible.

3 Examples and Results

In this section, we demonstrate various applications of the proposed method for the expectation of occupancy. We start with a set of simulation examples to show the superior performance of the proposed classifier, regardless of the level of correlation between variables and the proportion of $S = 1$ labels in the unclassified data set. Then we show the applicability of the MRMR algorithm for variable selection for the model. We also demonstrate that the minimum redundancy algorithm proposed here performs comparably to the MRMR algorithm. Next we use a bald eagle nesting habitat data set to identify suitable nesting habitats, to obtain an RSF model, and to show that cottonwood trees have the highest expectation of occupancy as nesting habitats in the Upper Mississippi River National Wildlife and Fish Refuge. Finally we use an Indiana bat maternity roost (presence-only) data set to demonstrate variable selection using the proposed filter algorithm and to show that its bark structure gives shagbark hickory the highest expectation of occupancy as maternity roots. All computations were carried out on a 64-bit laptop with a Core i7 - 4 core processor at 1.8 GHz and 8 GB memory. All examples are completed using MATLAB version R2020a.

For the simulated data examples, we generate a random (unclassified) data set $\mathbf{M}_0$ using a multivariate normal distribution. Then we construct a binary response vector using a logistic function and obtain a subset associated with the label 1 to use as the presence-only data set, $\mathbf{M}_1$. Next, we create a random nonstratified partition, $\mathbf{M}_3$, using 30% of the rest of the data points for holdout validation. Finally, we strip the labels from the remaining data points to create a second data subset, $\mathbf{M}_2$, to be used for comparison with binary classification methods. We use $\mathbf{M}_0$ and $\mathbf{M}_1$ as the training data sets to obtain the classification model, $\mathcal{CL}$. We evaluate the predictive performance on the held-out validation set using classification error $\mathbb{P}\left(\mathcal{CL}(\mathbf{M}_3) \ne \mathbf{y}_3\right)$ as the performance evaluation metric for comparisons, where $\mathbf{y}_3$ is the responses associated with $\mathbf{M}_3$.

For micro-scale investigations, we use $\mathbb{E}\left[S \mid \mathbf{X} = \mathbf{x}\right]$ to compare the suitability of each member in a collection (tree species or potential sites for conservation) as nesting or roosting locations for the focal species. Suppose that a presence-only data set is available for a given region and let $\{A_t : t = 1, \ldots, N\}$ be a partition of the data points. Suppose that each partition $A_t = \{\mathbf{x}^{(t1)}, \mathbf{x}^{(t2)}, \ldots, \mathbf{x}^{(tm)}\} \subset \mathbb{R}^k$ contains distinct data points. Then

$$\mathbb{E}[S \mid A_t] = \sum_{j=1}^{m} \mathbb{P}(S = 1 \mid \mathbf{x}^{(tj)})\mathbb{P}(\mathbf{x}^{(tj)} \mid A_t) \approx C \sum_{j=1}^{m} \frac{\prod_{i=1}^{k} \phi_{X_i \mid S}(x_i^{(tj)} \mid 1)}{\prod_{i=1}^{k} \phi_{X_i}(x_i^{(tj)})}$$

for some constant $C > 0$. Hence, the expectation of occupancy can be compared between partitions by simply comparing the averages of $\left(\prod_{i=1}^{k} \phi_{X_i \mid S}(x_i \mid 1)\right) / \left(\prod_{i=1}^{k} \phi_{X_i}(x_i)\right)$ for each partition. We use the Kruskal-Wallis H-test to compare the partitions based on tree species for the real data sets. For the corresponding examples, we assume that the sample data can be partitioned using the given categorical variable and there is no relationship between the observations within each partition, nor between the partitions. We further assume that the shapes of the distributions of $\mathbb{E}\left[S \mid \mathbf{X} = \mathbf{x}\right]$ for each partition are unknown and thus the null hypothesis for each test is that the distribution of $\mathbb{E}\left[S \mid \mathbf{X} = \mathbf{x}\right]$ does not differ between the partitions.

Example 1

In this simulation example, we compare classifications using the expectation of occupancy and logistic regression. We generate the random data matrix $\mathbf{M}_0$ using 500 independent realizations from a 50-dimensional multivariate normal distribution. Consider the correlation matrix $\mathbf{\Sigma}$ with entries $\mathbf{\Sigma}_{ij} = \rho^{|\sqrt{i}-\sqrt{j}|}$ for $i, j \in \{1, \ldots, 50\}$, where $\rho \in (0, 1)$ controls the level of correlation. By choosing matrix Σ as the correlation matrix (with Pearson product-moment correlation coefficients), we model high correlation for observations which are close together in the correlation matrix, decreasing correlation for observations which are increasingly far away, and slightly increasing correlation among sets of variables as one moves to the right within the matrix. We draw the rows of matrix $\mathbf{M}_0 \in \mathbb{R}^{500\times 50}$ from a Normal$(\mu, \mathbf{\Sigma})$ distribution, where $\mu = [3, 3, \ldots, 3]^T$ is a constant mean vector chosen such that almost all the data points are nonnegative.

Let $\beta = [\beta_1, \beta_2, \ldots, \beta_{50}]^T$ with $\beta_i = 1$ for 10 indices i from set $\{1, 2, \ldots, 50\}$, as explained in each simulation experiment below, and otherwise $\beta_i = 0$. Let $\mathbf{z} = \dfrac{1}{1 + e^{-\mathbf{M}_0\beta+\beta_0}}$, where the choice of β_0 controls the class sizes. We define the categorical response $\mathbf{y}$ by

$$y_i = \begin{cases} 1 & \text{if } z_i > 0.5, \\ 0 & \text{if } z_i \le 0.5, \end{cases} \qquad \text{and} \qquad \mathbf{y} = [y_1, y_2, \ldots, y_{500}]^T \qquad \text{and} \qquad \mathbf{z} = [z_1, z_2, \ldots, z_{500}]^T.$$

With this model, we generate a best-suited data set for logistic regression to compare to the performance of the classifier $\mathcal{CL}$. To produce the presence-only data set $\mathbf{M}_1$, we take the first 50 data points with the label "1". As explained earlier, we randomly partition the remaining data points to obtain the second data sample $\mathbf{M}_2$ and the nonstratified held-out validation set $\mathbf{M}_3$. We obtain the decision boundary for $\mathcal{CL}$ by setting $\pi = 0.975$. We use $\mathbf{M}_1$ and $\mathbf{M}_2$ to obtain the logistic regression model assuming that the second data set represents the class with "0" labels. We obtain one logistic classifier (Logistic 1) by setting the false negative rate for the presence-only data set to 2.5% (the same boundary criterion as used for $\mathcal{CL}$). For reference, we also include the logistic classifier that minimizes the classification error for the training set (Logistic 2). We use the validation set $\mathbf{M}_3$ to evaluate the performance by comparing the classification error between the actual responses $\mathbf{y}_3$ and the predicted responses $\mathcal{CL}(\mathbf{M}_3)$. We repeat the simulation for 25 randomly generated data matrices $\mathbf{M}_0, \mathbf{M}_1, \mathbf{M}_2$ and distinct pairs of validation sets $(\mathbf{M}_3, \mathbf{y}_3)$, and then average the results to obtain the predictive performance.

Simulation experiment 1. In this experiment, we compare the classification performance by assuming that the subset of variables that generates the output is known. We choose $\beta_i = 1$ for 10 equality spaced variables X_i, $i = 5, 10, 15, \ldots, 50$. By choosing nonzero coefficients spaced as far apart as possible, we force the least possible correlation among the model variables. We set increasing correlation levels, $\rho = 0.5, 0.7, 0.9$, and use $\beta_0 = 28, 30, 32, 34$, so that the proportion of "1" labels in the validation set, $\mathbf{M}_3$, decreases from about 60% to about 20% as β_0 increases. The classification errors and the false negative rates are given in Figure 3 and in Table 1 of Appendix A. The classification error for $\mathcal{CL}$ is always less than 0.09, and for the majority of the cases, the method accurately classifies more than 94% of the responses. The errors for the logistic classifiers are always larger than 0.09. The large classification errors associated with Logistic 2 show the inapplicability of classifiers that assume the two data sets belong to two distinct classes. Also, the false negative rate for the proposed method is always better than the other two methods (see Table 1). While the accuracy of $\mathcal{CL}$ increases with an increase in correlation, the accuracy does not show a dependent relationship with the percentage of "l" labels in the validation set. The logistic classifiers, however, show a significant increase in classification error with an increase in the proportion of "1" labels.

Next we evaluate the performance of the classifier when the percentage of "1" labels in the data sets is very small. We set $\rho = 0.5, 0.7, 0.9$ and, for each value of ρ, we find two appropriate β_0 values such that the proportion of "1" labels in the validation set is about 5% and 10%, respectively. In this case, a 2.5% false negative rate results in a larger classification error. We obtain the decision boundary for $\mathcal{CL}$ by setting $\pi = 0.9$ and, for the first logistic regression model, by setting the false negative rate for the presence-only data set to 10%. The classification errors are given in Figure 4 and in Table 1 of Appendix A. The proposed method accurately classifies more than 96% of the responses and outperforms the logistic classifiers except when $\rho = 0.5$. Also, $\mathcal{CL}$ accurately classifies more than 99% of the responses at a higher correlation (using $\rho = 0.9$), with no more than 40% of the classification error of the other methods.

Simulation experiment 2. For the second simulation experiment, we set $\beta_i = 1$ for the first 10 coefficients, $i = 1, 2, 3, \ldots, 10$. By choosing consecutive nonzero coefficients, we increase the possible correlation among the model variables. We again set the correlation levels $\rho = 0.5, 0.7, 0.9$ and $\beta_0 = 34, 32, 30, 28$. The classification errors and the false negative rates are given in Figure 5 and in Table 3 of Appendix A. The proposed classifier accurately classifies at least 95% of the responses for every case. The classification error for the logistic classifiers is always higher than that of $\mathcal{CL}$: at least 4 times higher for all but one case. In fact, the classification error of $\mathcal{CL}$ is almost always smaller than the respective cases in simulation experiment 1, even though the derivation of the method assumes uncorrelated variables. If we set $\beta_i = 1$ for the last 10 coefficients, $i = 41, 42, 43, \ldots, 50$, where the correlation among model variables is maximized within this data matrix, $\mathcal{CL}$ does an even better job of outperforming the other methods by classifying more than 97% of the responses accurately for each value of ρ.

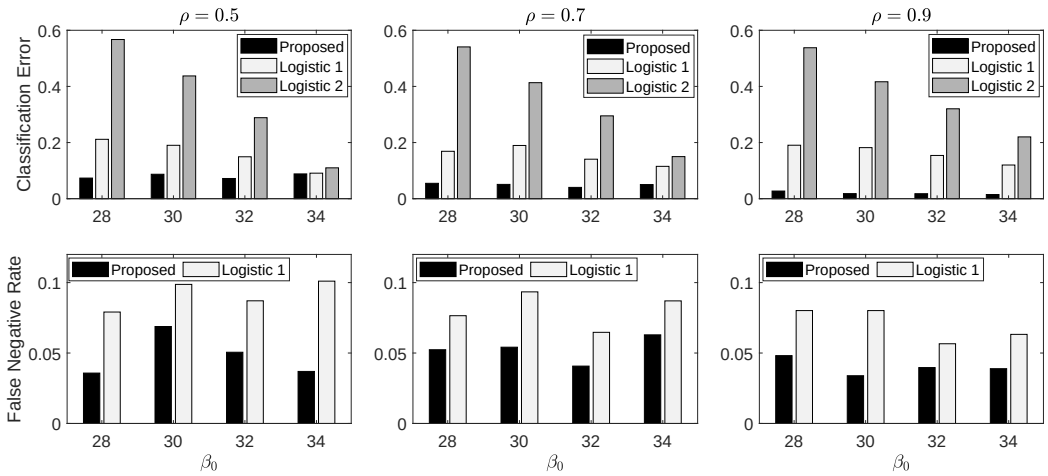


Figure 3: Classification performance for tests of the proposed classifier $\mathcal{CL}$ and the logistic classifier for simulation experiment 1. The decision boundary for Logistic 1 is obtained by setting the false negative rate for the presence-only data set to 2.5%. The decision boundary for Logistic 2 is obtained by minimizing the classification error for the training set assuming the two data sets belong to two distinct classes.

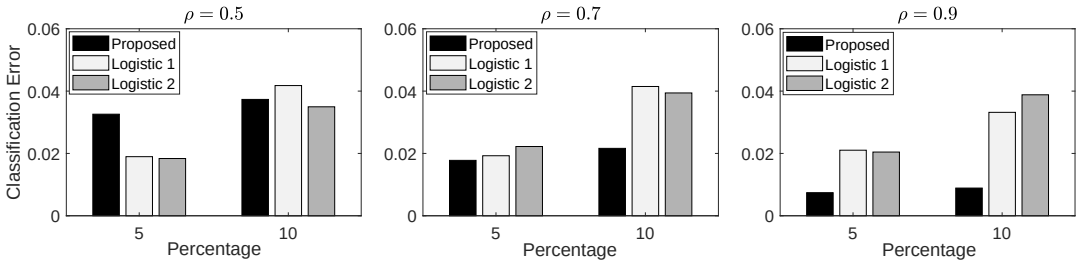


Figure 4: Classification performance for tests of $\mathcal{CL}$ and the logistic classifier using simulation experiment 1 when the proportion of "1" labels in the validation set is about 5% and 10%. The decision boundary for Logistic 1 is obtained by setting the false negative rate for the presence-only data set to 10%. The decision boundary for Logistic 2 is obtained by minimizing the classification error for the training set assuming the two data sets belong to two distinct classes.

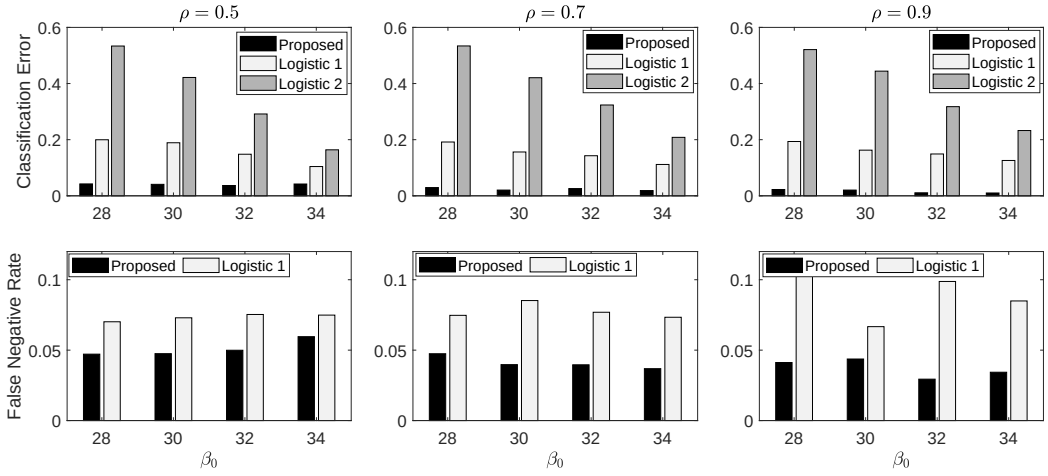


Figure 5: Classification performance for tests of $\mathcal{CL}$ and the logistic classifier for simulation experiment 2. The decision boundary for Logistic 1 is obtained by setting the false negative rate for the presence-only data set to 2.5%. The decision boundary for Logistic 2 is obtained by minimizing the classification error for the training set assuming the two data sets belong to two distinct classes.

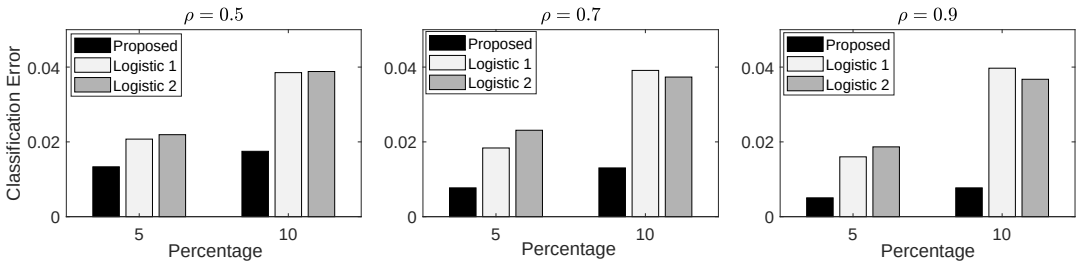


Figure 6: Classification performance for tests of $\mathcal{CL}$ and the logistic classifier for simulation experiment 2, for which the proportion of "1" labels in the validation set is about 5% and 10%. The decision boundary for Logistic 1 is obtained by setting the false negative rate for the presence-only data set to 10%. The decision boundary for Logistic 2 is obtained by minimizing the classification error for the training set assuming the two data sets belong to two distinct classes.

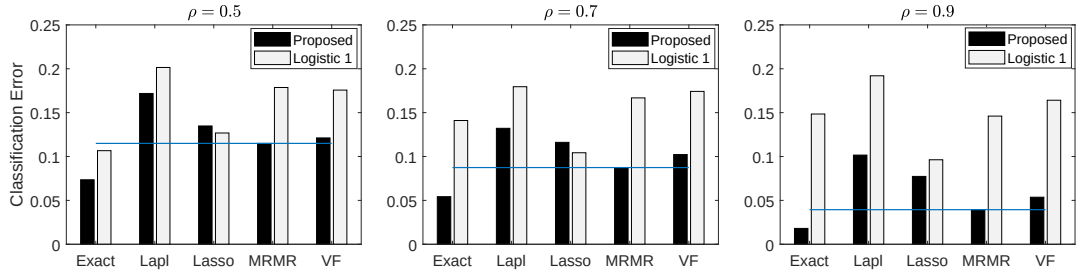


Figure 7: Comparison of the proportion of classification errors between four variable selection methods that have been applied to the data set described in the setup of simulation experiment 1. The classification error for the exact model variables is compared with the variable selection from Laplacian scores (Lapl), lasso regularization (Lasso), the MRMR algorithm (MRMR), and the proposed variable filter (VF).

As in simulation experiment 1, next we find two appropriate β_0 values such that the proportion of "1" labels in the validation set is about 5% and 10%, respectively. The classification errors are given in Figure 6 and in Table 3 of Appendix A. The proposed classifier $\mathcal{CL}$ outperforms the others by accurately classifying more than 98% of the responses. The classification error for the logistic classifiers is always higher than that of $\mathcal{CL}$: at least 2 times higher for all but one case. These simulations demonstrate the superior performance of the proposed method compared to logistic classifiers for the correlation spectrum considered here.

Simulation experiment 3. In this experiment, we evaluate the variable filter performance by applying it to the data sets described in the setup of simulation experiments 1 and 2. We compare the classification error produced by the exact model variables with the variables selected by the proposed filter, Laplacian scores (He et al., 2005), the MRMR algorithm, and lasso regularization. Laplacian scores produce a variable ranking with unsupervised learning, and are thus suitable for filtering variables using a presence-only data set alone. Since both the MRMR algorithm and lasso regularization need a response vector, we assign all entries in the the unclassified data set to the "0" class. For the proposed filter, we use $t = 0.9$ to account for 90% of the variation within the presence-only data matrix. Let n_0 be the number of variables from the filter. We select the first n_0 variables from the ranking generated by Laplacian scores and by the MRMR algorithm for comparison. We use lasso regularization to remove redundant predictors using 10-fold cross-validation to identify the model that corresponds to the minimum cross-validated mean squared error (MSE). First we choose β and ρ as in simulation experiment 1, but we set $\beta_0 = 32$ so that only 30-35% of the validation data set has "1" labels. Since we force the least possible correlation among the model variables, we have the most favorable setting for the proposed variable filter. The classification errors are given in Figure 7 and in Table 2 of Appendix A. The MRMR algorithm combined with $\mathcal{CL}$ produces the least classification error, with more than 88% accurate classifications. In comparison, the proposed variable filter produces only a marginally greater classification error, but by using only the data set with "1" labels, whereas both MRMR and lasso require both classes.

Next we choose β and ρ as in simulation experiment 2 and $\beta_0 = 32$. Since we force a higher correlation among the model variables, we have an unfavorable setting for proposed variable filter. Also, since we choose consecutive variables for the model, we reduce the likelihood of choosing a highly correlated adjacent variable in place of a model variable. Subsequently, we expect that methods producing lower classification errors will have chosen more of the true model variables and any method producing variables far away from the true model variables will produce much larger classification errors. The proportion of classification errors are given in Figure 8 and in Table 4 of Appendix A. As in the previous experiment, the MRMR algorithm combined

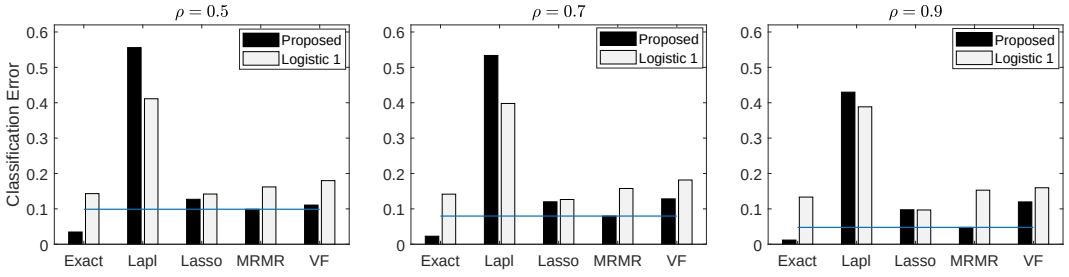


Figure 8: Comparison of the proportion of classification errors between four variable selection methods that have been applied to the data set described in the set up of simulation experiment 2. The classification error for the exact model variables is compared with the variable selection from Laplacian scores (Lapl), lasso regularization (Lasso), the MRMR algorithm (MRMR), and the proposed variable filter (VF).

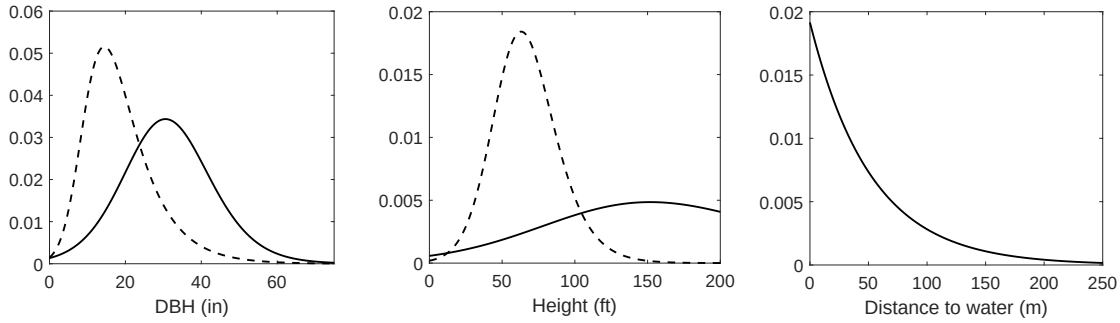


Figure 9: Density estimations of presence-only (continuous line) and random sample (dashed line) sites for the bald eagle nesting data from Mundahl et al. (2013) using $\phi(x)$ (Equation 2). The plots demonstrate a preference for larger and taller trees near water bodies.

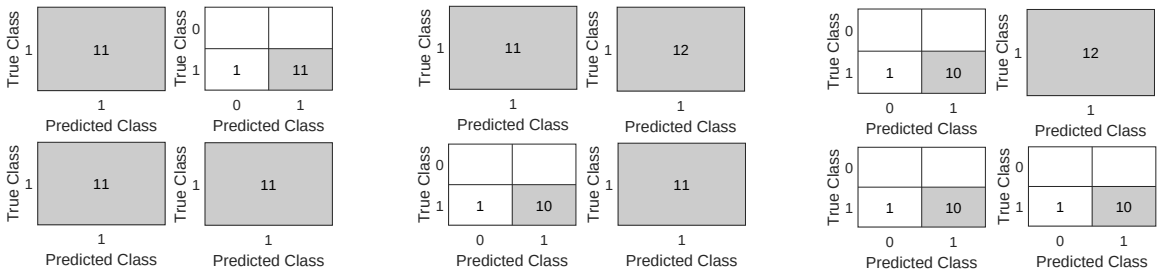
with the proposed classifier $\mathcal{CL}$ produces the least classification error for all the cases, with accuracy above 90%. Lasso produces the second best solutions and the proposed filter solutions closely follow. In contrast, variable ranking from Laplacian scores produces classifications with more than 40% errors. These simulations demonstrate that, while the MRMR algorithm may be the best to use in conjunction with the proposed classifier, the proposed filter is a reasonable alternative when the variables must be selected using presence-only data sets.

Example 2

In this example, we use $\mathbb{E}[S \mid \mathbf{X} = \mathbf{x}]$ to predict bald eagle nesting habitats in the Upper Mississippi River National Wildlife and Fish Refuge using data from Mundahl et al. (2013). The presence-only data set (45 data points) is comprised of four variables: tree diameter at breast height (DBH), tree height, nest height, and distance to water. While the placement of the nest is an important variable for habitat characterization, it cannot be utilized for assessing potential habitat selections. Therefore, we do not use it in this analysis. A random sample of unclassified data (380 data points) is comprised of DBH and tree height. Since the distance to the water is not included in the random sample, we cannot use variable selection methods. Using the presence-only data set, all three variables (DBH, height, and distance to water) are selected by the variable filter at $t = 0.75$. Figure 9 shows the probability density function approximations for each variable using $\phi(x)$. The plots demonstrate a preference for larger and taller trees near water bodies, supporting the significance of the variables selected by the filter.

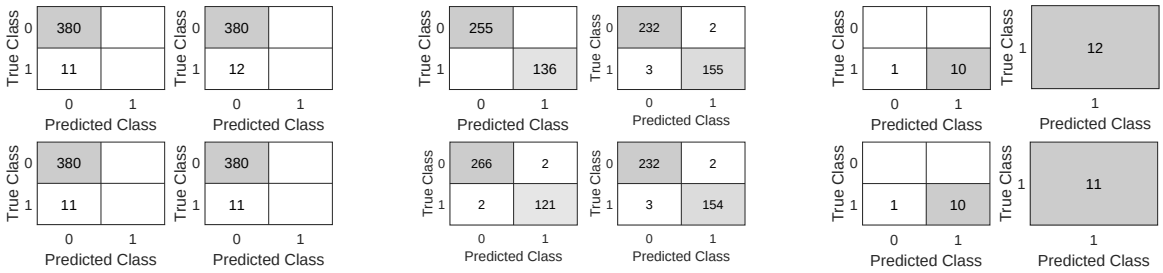
Next we use $\mathcal{CL}$ to identify suitable nesting habitats among the unclassified data set. We create a random nonstratified partition for 4-fold cross-validation on the presence-only data set. The folds are chosen randomly but with roughly equal size. We first find an appropriate false negative rate by comparing the classifications. The classification results for the test data for three false negative rates are depicted by means of confusion matrices in Figure 10. Each figure contains four confusion matrices, each representing the test data for the four iterations of 4-fold cross-validation. Both rates 1% and 2.5% produce only one instance of a false negative and we choose 2.5% for classifications.

To compare the classification results, we use the binary support vector machine (SVM) model. We compare the SVM results obtained by assigning the entire unclassified data set to the "0" class against the results obtained by assigning to the unclassified data the classifications predicted by the proposed classifier. We use the held-out presence-only partition and the unclassified



(a) 1% false negative rate (b) 2.5% false negative rate (c) 5% false negative rate

Figure 10: Confusion matrices for the three false negative rates for each test data set in the four iterations of 4-fold cross-validation. The random nonstratified partition for 4-fold cross-validation on the presence-only data set is created using the presence-only sites for the bald eagle nesting data from Mundahl et al. (2013).

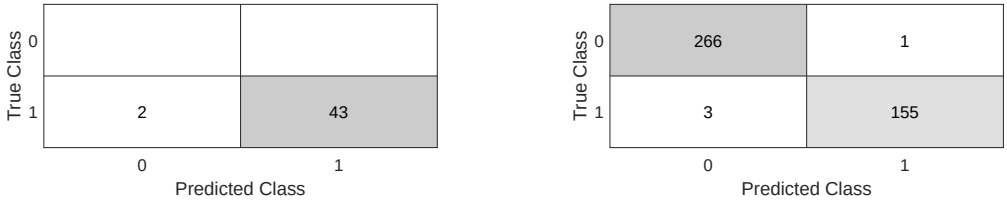


(a) With unclassified data (b) With $\mathcal{CL}$ classifications (c) Held-out presence-only

Figure 11: Confusion matrices for the SVM models for each test data set in the four iterations of 4-fold cross-validation in Example 2. Confusion matrices for the test data in (a) are from the SVM model when the entire unclassified data set is assigned to the "0" class. Confusion matrices for the test data in (b) and for the held-out partition in (c) are from the SVM model using $\mathcal{CL}$-assigned classifications for the unclassified data set.

data as the test data set. The classification results for the test data for each model are depicted by means of confusion matrices in Figure 11. Assigning label "0" to the entire unclassified data set results in a 100% false negative rate for every held-out presence-only partition (Figure 11a), whereas the SVM model that uses $\mathcal{CL}$-assigned classifications for the unclassified data set produces at most one false negative occurrence for each held-out presence-only partition (Figure 11c). In fact, the SVM model produces classifications that are almost identical to the proposed classifier (see Figure 11b). The trained SVM model correctly classifies at least 98% of the test data. This result demonstrates how the proposed classification method clearly distinguishes the two classes without explicit information about both classes.

Next we use the logistic regression model with the classifications assigned by $\mathcal{CL}$ to obtain an appropriate RSF model for the data set. The resulting model is given by $\text{RSF} = 1/\left(1 + e^{-2.24x_1 - 0.423x_2 + 82.99}\right)$, where x_1 and x_2 correspond to DBH and height, respectively. Using the decision boundary that produces the minimum error, the classification results for the presence-only and the complete data sets produced by the RSF function are given by the confusion matrices in Figure 12. Only 4 instances out of 425 classifications differ between the two classifiers, suggesting an excellent RSF function for the data set.



(a) Presence-only (b) All data

Figure 12: Confusion matrices for the RSF model for each test data set in the four iterations of 4-fold cross-validation.

We conclude this example by including a micro-scale analysis. It is noteworthy that, even though tree species was not included in the calculation of $\mathbb{E}[S \mid \mathbf{X} = \mathbf{x}]$, 65% of the random trees with label "1" assigned by $\mathcal{CL}$ are cottonwood trees. For comparison, only 30% of the sample of unclassified trees are cottonwood. This is not a coincidence. In the data set from Mundahl et al. (2013), bald eagle nesting trees were of four species: cottonwood (*Populus*), silver maple (*Acer saccharinum*), swamp white oak (*Quercus bicolor*), and red maple (*Acer rubrum*). We test whether there is a preference for cottonwood over the other species using tree height and DBH data alone. The Wilcoxon rank sum test indicates that there is enough evidence to reject the null hypothesis that the distribution of $\mathbb{E}[S \mid \mathbf{X} = \mathbf{x}]$ for cottonwood trees does not differ from the distribution of $\mathbb{E}[S \mid \mathbf{X} = \mathbf{x}]$ for other species (p-value = 0.00018 and ranksum = 886). Estimations for the average expectation of occupancy for cottonwood trees is more than 4 times higher than for the other species. Cottonwood has the highest expectation of occupancy given the tree height and DBH.

Example 3

We conclude the example section with a micro-scale example using a single class data set. We investigate tree bark structure preference for Indiana bat maternity roost selection using the presence-only data set (19 data points) from Schroder et al. (2017). The data set includes variables: tree species, height, DBH, distance to forest edge, distance to water, percentage of peeling bark, canopy opening, distance between maternity colonies, and potential maternity colony habitat area. Since there is no data for a random sample, we cannot use any variable selection algorithm, such as MRMR or lasso. If we use Laplacian scores, the distance between maternity colonies and potential maternity colony habitat area rank as the most important variables, and the distance to water and percentage of bark cover rank as the least important variables. Although distance to water and percentage of bark cover have been shown to be important (Carter and Feldhamer, 2005; Schroder et al., 2017; Kurta et al., 2002), there is no known correlation between maternity roosts and the variables distance between maternity colonies and colony habitat area. In contrast, five variables are selected by the proposed variable filter algorithm with $t = 0.8$, all of which have strong evidence of relevance in the literature: DBH, tree height, percentage of peeling bark, distance to forest edge, and distance to water. See Schroder et al. (2017) for a comprehensive discussion.

In this example, we do not have an unclassified data set. Therefore, we cannot directly calculate $\mathbb{E}[S \mid \mathbf{X} = \mathbf{x}]$. However, we may use the conditional joint probability distribution to compare partitions. We choose only tree DBH and percentage of peeling bark as the variables and analyze the difference between the joint distribution $\prod_{i=1}^{k} \phi_{X_i \mid S}(x_i \mid 1)$ for shagbark hickory (*Carya ovata*) versus the other three species: black locust (*Robinia pseudoacacia*), red oak (*Quercus rubra*), elm (*Ulmus spp.*), black oak (*Quercus velutina*), and walnut (*Juglans nigra*). We remove one outlier with respect to the chosen variables, which is located zero distance from both the water boundary and the forest edge, producing very favorable conditions for roosting. The Wilcoxon rank sum test indicates that there is enough evidence to reject the null hypothesis that the distribution of $\mathbb{E}[S \mid \mathbf{X} = \mathbf{x}]$ for shagbark hickory does not differ from the distribution of $\mathbb{E}[S \mid \mathbf{X} = \mathbf{x}]$ for the other species (p-value = 0.0142 and ranksum = 58). Estimation for the average expectation of occupancy for shagbark hickory is more than 5 times greater than for the other species. Shagbark hickory has the highest expectation of occupancy given only the bark structure and DBH.

4 Conclusion

In this paper, we introduced a model for the expectation of occupancy for the purpose of making comparisons based on variables that are most important in predicting habitat quality. We proposed the construction of a joint probability estimation, having taken into account the possibility that the ranges of some variables may be restricted by geographical or ecological constraints. We combined presence-only data with a sample of random unclassified data to model the expectation of occupancy. We included various examples of the expectation of occupancy to demonstrate the suitability of the proposed methods for avian nesting and roosting habitat quality assessments.

A Example 1 Simulation Results

Table 1: Simulation Experiment 1. Classification performance for tests of the proposed method $\mathcal{CL}$ and the logistic classifier using simulation experiment 1 in Example 1. Logistic 1 and Logistic 2 correspond to the classification performance of the logistic regression function. The decision boundary for Logistic 1 is obtained by setting the false negative rate for the presence-only data set to $100 \times (1 - \pi)\%$. The decision boundary for Logistic 2 is obtained by minimizing the classification error for the training set assuming the two data sets belong to two distinct classes. The solution with the least classification error for each case is in boldface. CE - Classification error, FN - Proportion of false negatives.

π	β_0	ρ	Percentage "1" labels	Proposed		Logistic 1		Logistic 2	
				CE	FN	CE	FN	CE	FN
0.975	34	0.5	18.3	**0.088**	0.037	0.091	0.101	0.110	0.596
	32	0.5	31.5	**0.072**	0.051	0.149	0.087	0.289	0.916
	30	0.5	45.2	**0.087**	0.069	0.190	0.099	0.437	0.966
	28	0.5	58.1	**0.074**	0.036	0.212	0.079	0.567	0.976
	34	0.7	22.2	**0.050**	0.063	0.115	0.087	0.150	0.663
	32	0.7	32.4	**0.040**	0.041	0.141	0.065	0.295	0.906
	30	0.7	43.1	**0.051**	0.054	0.190	0.093	0.413	0.960
	28	0.7	55.7	**0.055**	0.052	0.169	0.077	0.541	0.971
	34	0.9	25.5	**0.015**	0.039	0.120	0.063	0.220	0.858
	32	0.9	34.3	**0.018**	0.040	0.154	0.057	0.320	0.929
	30	0.9	43.5	**0.018**	0.034	0.182	0.080	0.417	0.956
	28	0.9	55.1	**0.027**	0.048	0.191	0.080	0.538	0.974
0.9	36.75	0.5	4.9	0.033	0.100	0.019	0.141	**0.018**	0.115
	35.75	0.5	9.4	0.037	0.123	0.042	0.188	**0.035**	0.303
	38.00	0.7	5.1	**0.018**	0.120	0.019	0.164	0.022	0.151
	36.50	0.7	9.7	**0.022**	0.084	0.042	0.116	0.039	0.316
	39.75	0.9	5.0	**0.007**	0.130	0.021	0.173	0.020	0.158
	38.25	0.9	9.9	**0.009**	0.092	0.033	0.166	0.039	0.274

Table 2: Variable selection for Simulation Experiment 1. Comparison of classification errors between four variable selection methods for the simulation experiment 1 setup. The classification error for the exact model variables is compared with the variable selection from Laplacian scores (Lapl), lasso regularization (Lasso), the MRMR algorithm (MRMR), and the proposed variable filter (VF). CE - Classification error, FN - Proportion of false negatives.

Method	ρ	% 1 labels in validation set	Proposed		Logistic 1		Logistic 2	
			CE	FN	CE	FN	CE	FN
Exact	0.5	31.0	0.073	0.065	0.107	0.073	0.287	0.926
Laplacian			0.172	0.059	0.201	0.108	0.267	0.863
Lasso			0.135	0.088	0.127	0.112	0.269	0.872
MRMR			0.115	0.084	0.179	0.077	0.263	0.848
VF			0.121	0.075	0.176	0.104	0.263	0.851
Exact	0.7	33.7	0.054	0.038	0.141	0.090	0.320	0.951
Laplacian			0.132	0.023	0.180	0.108	0.308	0.916
Lasso			0.116	0.066	0.104	0.098	0.314	0.935
MRMR			0.087	0.067	0.167	0.092	0.311	0.927
VF			0.102	0.051	0.174	0.103	0.313	0.929
Exact	0.9	33.8	0.018	0.040	0.148	0.073	0.326	0.964
Laplacian			0.102	0.024	0.192	0.089	0.324	0.960
Lasso			0.077	0.058	0.096	0.074	0.327	0.967
MRMR			0.039	0.066	0.146	0.083	0.325	0.963
VF			0.054	0.051	0.164	0.072	0.324	0.958

Table 3: Simulation Experiment 2. Classification performance for tests of the proposed method and the logistic classifier using simulation experiment 2. Logistic 1 and Logistic 2 correspond to the classification performance of the logistic regression function. The decision boundary for Logistic 1 is obtained by setting the false negative rate for the presence-only data set to $100 \times (1 - \pi)\%$. The decision boundary for Logistic 2 is obtained by minimizing the classification error for the training set assuming the two data sets belong to two distinct classes. The solution with the least classification error for each case is in boldface. CE - Classification error, FN - Proportion of false negatives.

π	β_0	ρ	Percentage "1" labels	Proposed		Logistic 1		Logistic 2	
				CE	FN	CE	FN	CE	FN
0.975	34	0.5	24.4	**0.042**	0.060	0.104	0.075	0.164	0.675
	32	0.5	32.1	**0.037**	0.050	0.148	0.075	0.292	0.911
	30	0.5	43.5	**0.041**	0.048	0.189	0.073	0.422	0.969
	28	0.5	54.6	**0.042**	0.047	0.200	0.070	0.534	0.978
	34	0.7	24.6	**0.019**	0.037	0.112	0.073	0.208	0.847
	32	0.7	34.4	**0.026**	0.040	0.143	0.077	0.324	0.941
	30	0.7	43.4	**0.020**	0.040	0.156	0.085	0.421	0.968
	28	0.7	54.6	**0.029**	0.047	0.192	0.075	0.534	0.977
	34	0.9	27.1	**0.010**	0.034	0.126	0.085	0.233	0.853
	32	0.9	34.8	**0.011**	0.029	0.149	0.099	0.318	0.912
	30	0.9	45.3	**0.020**	0.044	0.163	0.067	0.444	0.981
	28	0.9	53.2	**0.023**	0.041	0.193	0.104	0.521	0.979
0.9	38.50	0.5	4.9	**0.013**	0.093	0.021	0.127	0.022	0.123
	37.00	0.5	9.9	**0.017**	0.128	0.039	0.175	0.039	0.278
	39.25	0.7	5.2	**0.008**	0.083	0.018	0.131	0.023	0.134
	37.75	0.7	10.1	**0.013**	0.101	0.039	0.140	0.037	0.272
	40.00	0.9	5.0	**0.005**	0.082	0.016	0.175	0.019	0.185
	38.25	0.9	10.0	**0.008**	0.085	0.040	0.215	0.037	0.298

Table 4: Variable selection for Simulation Experiment 2. Comparison of classification errors between four variable selection methods for the simulation experiment 2 setup. The classification error for the exact model variables is compared with the variable selection from Laplacian scores (Lapl), lasso regularization (Lasso), the MRMR algorithm (MRMR), and the proposed variable filter (VF). CE - Classification error, FN - Proportion of false negatives.

Method	ρ	% 1 labels in validation set	Proposed		Logistic 1		Logistic 2	
			CE	FN	CE	FN	CE	FN
Exact	0.5	33.2	0.035	0.041	0.143	0.090	0.305	0.918
Laplacian			0.556	0.069	0.411	0.110	0.305	0.896
Lasso			0.127	0.071	0.142	0.123	0.298	0.889
MRMR			0.099	0.061	0.162	0.114	0.283	0.852
VF			0.111	0.061	0.180	0.108	0.289	0.868
Exact	0.7	35.6	0.023	0.039	0.142	0.077	0.333	0.936
Laplacian			0.534	0.043	0.398	0.079	0.334	0.925
Lasso			0.120	0.063	0.127	0.085	0.335	0.941
MRMR			0.080	0.060	0.158	0.093	0.326	0.912
VF			0.128	0.063	0.182	0.078	0.327	0.916
Exact	0.9	35.2	0.012	0.032	0.133	0.080	0.324	0.924
Laplacian			0.430	0.031	0.388	0.068	0.334	0.945
Lasso			0.097	0.043	0.097	0.077	0.327	0.925
MRMR			0.048	0.038	0.153	0.069	0.326	0.927
VF			0.120	0.030	0.160	0.075	0.320	0.906

References

Aldridge, C. L., D. J. Saher, T. M. Childers, K. E. Stahlnecker, and Z. H. Bowen (2012). Crucial nesting habitat for gunnison sage-grouse: A spatially explicit hierarchical approach. *The Journal of Wildlife Management 76*(2), 391–406. 117

Andrew, J. M. and J. A. Mosher (1982). Bald eagle nest site selection and nesting habitat in Maryland. *The Journal of Wildlife Management 46*(2), 382–390. 117

Battin, J. and J. J. Lawler (2006). Cross-scale correlations and the design and analysis of avian habitat selection studies. *The Condor 108*(1), 59–70. 118

Bernstein, N. P. and E. B. McLean (1980). Nesting of red-winged blackbirds in cattails and common reed grass in Mentor Marsh. *Ohio Journal of Science 80*(1), 14–19. 118

Boyce, M. S., P. R. Vernier, S. E. Nielsen, and F. K. Schmiegelow (2002). Evaluating resource selection functions. *Ecological modelling 157*(2–3), 281–300. 117

Burke, D. M. and E. Nol (1998). Influence of food abundance, nest-site habitat, and forest fragmentation on breeding ovenbirds. *The Auk 115*(1), 96–104. 117

Carter, T. C. and G. A. Feldhamer (2005). Roost tree use by maternity colonies of Indiana bats and northern long-eared bats in southern Illinois. *Forest Ecology and Management 219*(2–3), 259–268. 117, 128

Chan, T. F. (1987). Rank revealing QR factorizations. *Linear Algebra Appl 88/89*, 67–82. 120

DeBoer, T. S. and D. D. Diamond (2006). Predicting presence-absence of the endangered golden-cheeked warbler (Dendroica chrysoparia). *The Southwestern Naturalist 51*(2), 181–191. 117

Devroye, L., L. Gyorfi, and G. Lugosi (1996). *A probabilistic theory of pattern recognition*. New York, NY: Springer. 118

Ding, C. and H. Peng (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology 3*(02), 185–205. 118, 120

Doherty, K. E., D. E. Naugle, and B. L. Walker (2010). Greater sage-grouse nesting habitat: The importance of managing at multiple scales. *The Journal of Wildlife Management 74*(7), 1544–1553. 117

Emrick, V. R., S. Tweddale, and M. S. Germain (2010). Characterization of golden-cheeked warbler Dendroica chrysoparia habitat at Fort Hood, Texas, USA. *Endangered Species Research 11*(3), 215–220. 117

Gibson, D., E. J. Blomberg, M. T. Atamian, and J. S. Sedinger (2016). Nesting habitat selection influences nest and early offspring survival in greater sage-grouse. *The Condor 118*(4), 689–702. 118

Gu, M. and S. C. Eisenstat (1996). Efficient algorithms for computing a strong rank-revealing QR factorization. *SIAM J. Sci. Comput. 17*(4), 848—-869. 120

Hammond, K. R., J. M. O'Keefe, S. P. Aldrich, and S. C. Loeb (2016). A presence-only model of suitable roosting habitat for the endangered Indiana bat in the Southern Appalachians. *PLoS ONE 11*(4), e0154464. 118

He, X., D. Cai, and P. Niyogi (2005). Laplacian score for feature selection. *Advances in neural information processing systems 18*, 507–514. 125

Keating, K. A. and S. Cherry (2004). Use and interpretation of logistic regression in habitat-selection studies. *The Journal of Wildlife Management 68*(4), 774–789. 117

Kelly, J. P. (1993). The effect of nest predation on habitat selection by dusky flycatchers in limber pine-juniper woodland. *The Condor 95*(1), 83–93. 118

Kroll, J. C. (1980). Habitat requirements of the golden-cheeked warbler: management implications. *Journal of Range Management 33*(1), 60–65. 117

Kurta, A., S. W. Murray, and D. H. Miller (2002). Roost selection and movements across the summer landscape. *The Indiana bat: biology and management of an endangered species (A. Kurta and J. Kennedy, eds.). Bat Conservation International, Austin, Texas*, 118–129. 128

Kusch, J., C. Weber, S. Idelberger, and T. Koob (2004). Foraging habitat preferences of bats in relation to food supply and spatial vegetation structures in a western European low mountain range forest. *Folia Zoology 53*(2), 113–128. 117

Lichstein, J. W., T. R. Simons, and K. E. Franzreb (2002). Landscape effects on breeding songbird abundance in managed forests. *Ecological Applications 12*(3), 836–857. 118

Lima, S. L. (2009). Predators and the breeding bird: Behavioral and reproductive flexibility under the risk of predation. *Biological Reviews 84*(3), 485–513. 118

Livingston, S. A., C. S. Todd, W. B. Krohn, and R. B. Owen (1990). Habitat models for nesting bald eagles in Maine. *The Journal of Wildlife Management 54*(4), 644–653. 117

Manly, B. F. J., L. L. McDonald, D. L. Thomas, T. L. McDonald, and W. P. Erickson (2002). *Resource Selection by Animals: Statistical Design and Analysis for Field Studies*. Dordrecht, Netherlands: Kluwer Academic Publishers. 117, 118

Martin, T. E. (1993). Nest predation and nest sites. *BioScience 43*(8), 523–532. 118

Mundahl, N. D., A. G. Bilyeu, and L. Maas (2013). Bald eagle nesting habitats in the upper Mississippi river national wildlife and fish refuge. *Journal of Fish and Wildlife Management 4*(2), 362–376. 117, 126, 127, 128

Newlon, K. R. and V. A. Saab (2011). Nest-site selection and nest survival of lewis's woodpecker in aspen riparian woodlands. *The Condor 113*(1), 183–193. 118

Pauli, B. P., H. A. Badin, G. S. Haulton, P. A. Zollner, and T. C. Carter (2015). Landscape features associated with the roosting habitat of Indiana bats and northern long-eared bats. *Landscape Ecology 30*(10), 2015–2029. 117, 118

Phillips, S. J., R. P. Anderson, and R. E. Schapire. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling 190*(3–4), 231–259. 117

Quinn, J., J. Prop, Y. Kokorev, and J. M. Black (2003). Predator protection or similar habitat selection in red-breasted goose nesting associations: extremes along a continuum. *Animal Behaviour 65*(2), 297–307. 118

Rissanen, J. (1978). Modeling by shortest data description. *Automatica 14*(5), 465–471. 118

Schroder, E. S., D. B. Ekanayake, and S. P. Romano (2017). Indiana bat maternity roost habitat preference within Midwestern United States upland Oak-Hickory (Quercus-Carya) forests. *Forest ecology and management 404*, 65–74. 117, 118, 128

Sonerud, G. A. (1985). Nest hole shift in Tengmalm's owl Aegolius funereus as defence against nest predation involving long-term memory in the predator. *Journal of Animal Ecology 54*(1), 179–192. 118

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288. 120

Verner, J. and M. F. Willson (1966). The influence of habitats on mating systems of North American passerine birds. *Ecology 47*(1), 143–147. 117

Willson, M. F. (1966). Breeding ecology of the yellow-headed blackbird. *Ecological Monographs 36*(1), 51–77. 118