StatEcoNet: Statistical Ecology Neural Networks for Species Distribution Modeling

Eugene Seo, [1] **Rebecca A. Hutchinson,** [1,2] **Xiao Fu,** [1] **Chelsea Li,** [1]
Tyler A. Hallman, [4] **John Kilbride,** [3] **W. Douglas Robinson** [2]

[1]School of Electrical Engineering and Computer Science, [2]Department of Fisheries and Wildlife,
[3]College of Earth, Ocean, and Atmospheric Sciences
Oregon State University, Corvallis, OR 97331
[4]Monitoring Department, Swiss Ornithological Institute, Sempach, Switzerland
{seoe,rah,xiao.fu,lichel}@oregonstate.edu, tyler.hallman@vogelwarte.ch, {kilbridj,douglas.robinson}@oregonstate.edu

Abstract

This paper focuses on a core task in computational sustainability and statistical ecology: species distribution modeling (SDM). In SDM, the occurrence pattern of a species on a landscape is predicted by environmental features based on observations at a set of locations. At first, SDM may appear to be a binary classification problem, and one might be inclined to employ classic tools (e.g., logistic regression, support vector machines, neural networks) to tackle it. However, wildlife surveys introduce structured noise (especially under-counting) in the species observations. If unaccounted for, these observation errors systematically bias SDMs. To address the unique challenges of SDM, this paper proposes a framework called `StatEcoNet`. Specifically, this work employs a graphical generative model in statistical ecology to serve as the skeleton of the proposed computational framework and carefully integrates neural networks under the framework. The advantages of `StatEcoNet` over related approaches are demonstrated on simulated datasets as well as bird species data. Since SDMs are critical tools for ecological science and natural resource management, `StatEcoNet` may offer boosted computational and analytical powers to a wide range of applications that have significant social impacts, e.g., the study and conservation of threatened species.

1 Introduction

Estimating species distributions across a landscape is a fundamental problem in ecology. Species distribution models (SDMs) learn the relationship between the species of interest and a set of environmental features (e.g., elevation, land cover) from data collected at points on the landscape (Elith and Leathwick 2009; Franklin and Miller 2010). The species data may come from historical records (Elith et al. 2006), professional surveys (Betts et al. 2008), or volunteers in community science projects [1] (Fink et al. 2010). The environmental data may be collected *in situ* or linked to the observation points post hoc (e.g., via remote sensing (Shirley et al. 2013)). SDMs are critical tools for both scientific inquiry and natural resource management, as they are employed to investigate how environmental features define species habitat and predict where species can persist successfully (Araújo and Peterson 2012).

At first glance, species distribution modeling may appear to be a straightforward machine learning (ML) problem, but the complex nature of ecological systems and the noise-prone data acquisition process entail unique challenges that are not addressed in conventional ML frameworks. First, data on species distributions are persistently plagued by *imperfect detection*, in which some individuals of the species are missing from the data because of poor observation conditions, species behavioral traits, and/or limited survey efforts. Second, species respond to their environment in complex ways, so models of this process must handle many input variables and represent nonlinear relationships. Third, models must be as interpretable as possible in order to translate their conclusions to meaningful scientific insights and effective management policies. Finally, SDMs are often built from smaller datasets than some other ML domains, with hundreds rather than thousands or millions of examples.

Classic approaches like regression models fail to capture systematic imperfect detection (Guillera-Arroita et al. 2014; Lahoz-Monfort, Guillera-Arroita, and Wintle 2014). Instead, a family of latent variable models has been developed in statistical ecology to account for error in the observation process (Royle and Dorazio 2008; MacKenzie et al. 2018). This family originated with *occupancy models*, in which the species *occupancy* (occurrence) at a set of sites is represented with binary latent variables, and the species observations depend on occupancy status as well as a *detection probability* (MacKenzie et al. 2002). In these models, the latent variables are of great scientific interest. Understanding how the environment determines occupancy may not only advance ecological research, but also assist policy decisions—e.g., making conservation policies for threatened species. Various extensions to this latent variable modeling framework have been introduced (e.g., with count-valued latent variables (Royle 2004)), but this paper focuses on the occupancy model as a representative example. These models are often used within a classic statistical paradigm, where the probabilities of occupancy and detection are linked to features with regression functions, and models are selected with criteria like AIC. This framework provides an effective approach to imperfect detection. However, it has lim-

[1]Also referred to as "citizen science" projects/data.

ited modeling capacity due to the use of the linear regression model and thus struggles to model complex (i.e., highly nonlinear) relationships in high-dimensional feature spaces.

To handle the challenge of complexity in species' environmental responses, many ecologists have turned to machine learning (Elith et al. 2006). In particular, boosted regression trees (BRT) and random forests (RF) are popular for their flexibility and predictive power (Elith, Leathwick, and Hastie 2008; Cutler et al. 2015); neural networks (NN) are an obvious alternative but have been explored less in this domain. Tree-based methods incorporate mechanisms for interpreting the model, such as feature importance metrics and partial dependence plots. However, these models treat species distribution modeling as a standard supervised classification problem without regard to the effects of imperfect detection; ignoring imperfect detection can cause systematic underestimation of species distributions. Furthermore, the effects of the features cannot be clearly separated into occupancy and detection components.

Contributions. This work puts forth a statistical ecology-inspired neural network model to address the above challenges. Our specific contributions are as follows. First, we propose a *statistical ecology-based neural network model* (`StatEcoNet`). The framework combines the statistical occupancy modeling approach that captures imperfect detection with neural networks that capture nonlinear relationships between the environment and species. We also introduce an easy-to-implement regularization strategy for selecting relevant features for the occupancy and detection submodels, instead of requiring the user to specify these assignments. Specifically, we propose to use a group-sparsity regularization in the first layers of the NNs in `StatEcoNet`, thereby clearly indicating importance of the features to the two sub-models of the occupancy framework. Note that group-sparse predictors are often considered in linear regression and compressive sensing (Jenatton, Audibert, and Bach 2011), but have not been considered in interpretable ecological system neural modeling. We show advantages of `StatEcoNet` over alternative approaches on simulated data as well as a case study modeling five bird species.

Prior Work. Two pieces of prior work have attempted to address combinations of these challenges. First, nonlinear models have been incorporated into occupancy models using boosted regression trees (called `OD-BRT`) (Hutchinson, Liu, and Dietterich 2011). That approach jointly fits two tree ensembles which are linked through an objective function that corresponds to the occupancy model likelihood. This addresses imperfect detection while automatically representing complex relationships to the features, but our experiments with this method indicate that it is difficult to tune properly and that it does not scale well to large datasets. Other recent work has also found that algorithms for learning BRT models are computationally intensive and can experience numerical instability (Ke et al. 2017). Second, recent work incorporates nonlinear models into occupancy models with neural networks instead of BRTs (Joseph 2020). However, it combines the features into a single network to model occupancy and detection, which limits interpretability.

2 Problem Statement

Consider a typical SDM setting where we are given binary observations (i.e., species detection or non-detection) made by observers at different sites. More formally, we define the following notation. The tth (where $t \in [T]$) observation at site i (where $i \in [M]$) is denoted by y_{it}. Note that $y_{it} \in \{0, 1\}$, where $y_{it} = 1$ means that the target species was observed at site i in the tth observation made, and $y_{it} = 0$ otherwise. For every observation, survey-specific features (e.g., temperature, time of day of the observation) are recorded and collected in $\mathbf{w}_{it} \in \mathbb{R}^K$. Every site is characterized by a number of site-specific features (e.g., elevation, forest type), which are collected in $\mathbf{x}_i \in \mathbb{R}^J$. The objective is to determine the occurrence pattern of the species from the observations and the site and survey features. After the relationship is learned, the model can be used to predict species observations for new sites. In many studies, it is also critical to interpret how site features affect the species—i.e., to identify the environmental drivers of its distribution.

Conventional Machine Learning Solution. From an ML viewpoint, it is tempting to treat the y_{it} as binary labels and concatenate the features to form $\mathbf{u}_{it} = [\mathbf{w}_{it}^\top, \mathbf{x}_i^\top]^\top \in \mathbb{R}^{J+K}$. Then, an *empirical risk minimization* (ERM)-type formulation could be employed:

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{M} \sum_{t=1}^{T} \mathcal{L}(y_{it} || f_{\boldsymbol{\theta}}(\mathbf{u}_{it})), \tag{1}$$

where $f_{\boldsymbol{\theta}}(\cdot) : \mathbb{R}^{K+J} \to \mathbb{R}$ is any established model in ML (e.g., logistic regression, neural networks), $\boldsymbol{\theta}$ collects the model parameters (e.g., neural network weights), and $\mathcal{L}(x||y)$ is a loss function (e.g., least squares, cross entropy).

Challenges. The ML solutions summarized in (1) seem reasonable, but the unique challenges of SDM may hinder performance. First, imperfect detection implies that some reports do not reflect the true status of the species at the site (e.g., when they are silent, hiding, or camouflaged), so these data contain structured noise. The probability of detecting a species varies across sites and surveys and is affected by numerous factors when conducting field surveys. Second, unlike classic applications of ML to 'big data,' many ecological datasets are collected under substantial resource constraints. It is common to analyze hundreds of sites, in contrast to millions of images. Hence, exclusively data-driven ML models, e.g., deep neural networks, may not be applicable.

To summarize, a completely data-driven complex ML model like deep neural networks may not be a viable solution for SDM. Nonetheless, neural networks offer appealing learning capacity in the presence of complex nonlinear transformations in the data generation process—and their companion algorithms balance modeling complexity, computational efficiency/stability, and generalization performance. These nice properties should be capitalized upon in SDM (e.g., for modeling the complex relations between site features and species distributions as well as the survey features and species detectability) with special attention paid to the ubiquity of missed detections and data scarcity challenges—this is the starting point of our work.

3 Proposed Framework

To address the challenges of applying advanced neural network-based learning techniques in SDM, we propose to integrate neural network-based nonlinear modeling with classic graphical generative models in statistical ecology. In a nutshell, the statistical model captures the effect of imperfect detection. The neural networks overcome model discrepancies that are often over-simplified in classic ecology models. This way, the neural networks are only responsible for handling the most challenging parts in the statistical model, while leaving the 'well-understood' part to the classic model based approach. This reduces the complexity of the network and makes the learning process more efficient.

3.1 Preliminaries: The Occupancy Model

The backbone of our proposed `StatEcoNet` is a widely accepted model in statistical ecology called the occupancy model (MacKenzie et al. 2002, 2018). The graphical representation of the latent variable model is shown in Fig. 1. For each site $i = 1, ..., M$, the biological model connects the true species occupancy status, $z_i \in \{0, 1\}$ to site features $\mathbf{x}_i$ through an occupancy probability o_i. The key advance of the occupancy model over the approach of (1) is the introduction of the latent variable z_i to capture the *true occupancy status* of the species at site i. The acquired data y_{it}'s are treated as noisy observations of z_i, since they are influenced by imperfect detection. Letting each site contain $t = 1, ..., T_i$ replicate surveys, the observation model links survey features $\mathbf{w}_{it}$ to a detection probability p_{it}. Note that introducing a detection probability that is associated with each observation is critical for SDM, since it explicitly models systematic under-counting. This model is intuitively and scientifically appealing, since it separates the causes for occupancy and detection; interpreting these effects separately is valuable in many ecological studies.

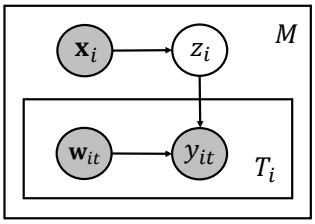

Figure 1: Graphical occupancy model. $z_i \in \{0, 1\}$ denotes latent species occupancy at site i (of M total) and $y_{it} \in \{0, 1\}$ denotes the tth observation (of T_i total). $\mathbf{x}_i$ and $\mathbf{w}_{it}$ are site and survey features, respectively.

In the generative model, the observed data y_{it} are produced by drawing from the occupancy Bernoulli and multiplying the result by the detection probability, i.e.,

$$y_{it} \sim \text{Bernoulli}(z_i d_{it}).$$

This encodes the assumption that unoccupied sites are always observed to be unoccupied, but that occupied sites might also be observed to be unoccupied. However, while it is clear that each observation y_{it} is affected by both the true occupancy z_i and the detection probability d_{it}, it is less clear how the site features $\mathbf{x}_i$ (resp. survey features $\mathbf{w}_{it}$) affect z_i (resp. d_{it}). In classical applications of occupancy models, linear models map $\mathbf{x}_i$ and $\mathbf{w}_{it}$ to occupancy probability o_i and detection probability d_{it}, respectively, through a linear logit modeling strategy (MacKenzie et al. 2018); i.e.,

$$o_i = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\alpha})}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\alpha})}, \quad d_{it} = \frac{\exp(\mathbf{w}_{it}^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{w}_{it}^\top \boldsymbol{\beta})}, \tag{2}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^K$ and $\boldsymbol{\beta} \in \mathbb{R}^J$ are model parameters to be estimated. The true occupancy has probability o_i, i.e.,

$$z_i \sim \text{Bernoulli}(o_i).$$

This framework makes sense, but the linear models are over-simplified for complex ecological systems.

3.2 Integrating Neural Networks into the Framework

In this work, we propose to use *two* neural networks to model the relations between $\mathbf{x}_i$ and o_i as well as $\mathbf{w}_{it}$ and d_{it} in (2). Our motivation is not to replace the well established graphical model in Fig. 1 by a completely data-driven neural network (as in (1)), but to leverage the power of neural networks to fill the 'modeling gap' of the graphical model. For statistical ecologists, this is perhaps the most natural way of integrating neural networks into SDM.

Specifically, we introduce two neural networks

$$F(\cdot) : \mathbb{R}^K \to \mathbb{R}, \; G(\cdot) : \mathbb{R}^J \to \mathbb{R}$$

as shown in Fig. 2. The first neural network $F(\mathbf{x}_i)$ predicts the occupancy probability from the given site features $\mathbf{x}_i$. The second neural network predicts the detection probability from given survey features $\mathbf{w}_{it}$.

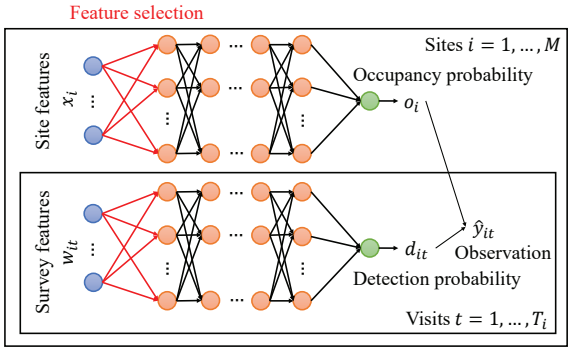

Figure 2: Proposed model (`StatEcoNet`) framework.

We employ fully connected networks to express F and G:

$$F(\mathbf{x}_i; \boldsymbol{\theta}_F) = \mathbf{u}_L^\top \boldsymbol{\sigma}(\mathbf{U}_{L-1} \boldsymbol{\sigma}(\cdots \boldsymbol{\sigma}(\mathbf{U}_1 \mathbf{x}_i))), \tag{3a}$$

$$G(\mathbf{w}_{it}; \boldsymbol{\theta}_G) = \mathbf{v}_L^\top \boldsymbol{\sigma}(\mathbf{V}_{L-1} \boldsymbol{\sigma}(\cdots \boldsymbol{\sigma}(\mathbf{V}_1 \mathbf{w}_{it}))). \tag{3b}$$

In (3a), $\mathbf{U}_\ell \in \mathbb{R}^{K_\ell \times K_{\ell-1}}$ is the network weight in the ℓth layer where K_ℓ is the number of neurons of the ℓth layer (and we define $K_0 = K$). The output layer has a combining vector $\mathbf{u}_L \in \mathbb{R}^{K_{L-1}}$ that maps the output to a scalar. The

weights in (3b) are defined in the same way. We also define $\boldsymbol{\theta}_F$ and $\boldsymbol{\theta}_G$ as the collections of the network parameters of F and G, respectively. The function $\boldsymbol{\sigma}(\cdot)$ applies onto every element of its input individually. We employ the popular rectified linear unit (ReLU) function as our activation function. With the neural networks defined, the occupancy and detection probabilities can be re-expressed as follows:

$$o_i = \frac{\exp(F(\mathbf{x}_i; \boldsymbol{\theta}_F))}{1 + \exp(F(\mathbf{x}_i; \boldsymbol{\theta}_F))}, \tag{4a}$$

$$d_{it} = \frac{\exp(G(\mathbf{w}_{it}; \boldsymbol{\theta}_G))}{1 + \exp(G(\mathbf{w}_{it}; \boldsymbol{\theta}_G))}. \tag{4b}$$

With the above construction and the overall graphical model, we define a maximum likelihood estimation problem whose log-likelihood function can be expressed as:

$$\begin{aligned} \log \mathcal{L} &= \sum_{i=1}^{M} \log \mathcal{L}_i \\ &= \sum_{i=1}^{M} \log \left(o_i \prod_{t=1}^{T_i} \left[d_{it}^{y_{it}} (1 - d_{it})^{1 - y_{it}} \right] + (1 - o_i) \kappa_i \right), \end{aligned} \tag{5}$$

where κ_i is an indicator function defined as $\mathbb{1}\left(\sum_{t=1}^{T_i} y_{it} = 0\right)$, in which $\mathbb{1}(\cdot)$ is 1 if the observations at a site were all zero and 0 otherwise. In the above, we have followed the derivation of (Hutchinson, Liu, and Dietterich 2011) to reach the expression of $\mathcal{L}_i$ from $\mathcal{L}_i = \sum_{z \in \{0,1\}} \mathsf{Pr}(z_i = z) \prod_{t=1}^{T_i} \mathsf{Pr}(y_{it} | z_i = z) = \sum_{z \in \{0,1\}} o_i^z (1 - o_i)^{1-z} \prod_{t=1}^{T_i} (z d_{it})^{y_{it}} (1 - z d_{it})^{1 - y_{it}}$.

3.3 Feature Selection via the $\ell_{2,1}$-Norm

On top of this structure, we incorporate regularization terms into our model in order to identify features that significantly impact each of the model probabilities. There has been little work on incorporating the feature selection process into neural network models. Instead, most prior work selects relevant features as a preprocessing before learning the neural network model (Cheng et al. 2020). Here, we add the $\ell_{2,1}$-norm into our `StatEcoNet` to reveal which features impact the occupancy and detection probabilities.

The mixed $\ell_{2,1}$-norm (also denoted as ℓ_2/ℓ_1-norm) is a matrix norm used for robust optimization problems that promotes sparsity of the matrix columns. It thus has widely been used in signal and image processing to handle noise and outliers (Steffens, Pesavento, and Pfetsch 2018). Accordingly, the $\ell_{2,1}$-norm has been considered an approach for feature selection (Jenatton, Audibert, and Bach 2011). The $\ell_{2,1}$-norm of $\mathbf{U}_\ell \in \mathbb{R}^{K_\ell \times K_{\ell-1}}$ is defined as

$$\|\mathbf{U}_\ell\|_{2,1} = \sum_{j=1}^{K_{\ell-1}} \Big(\sum_{i=1}^{K_\ell} |u_{ij}|^2\Big)^{1/2} = \sum_{j=1}^{K_{\ell-1}} \|\mathbf{U}_\ell(:, j)\|_2. \tag{6}$$

The $\ell_{2,1}$-norm behaves like an ℓ_1-norm on a vector for providing a sparse solution to the columns of a matrix. That is, the parameter matrix is regularized with the $\ell_{2,1}$-norm minimization in order to discover important features. We introduce this mixed $\ell_{2,1}$-norm into the first input layer of both neural networks, where the parameter matrix is connected to the input features as shown in Fig. 2.

Our regularized loss function is given by

$$-\sum_{i=1}^{M} \log \mathcal{L}_i + \lambda_F \|\mathbf{U}_1\|_{2,1} + \lambda_G \|\mathbf{V}_1\|_{2,1}, \tag{7}$$

where λ_F and λ_G are regularization weights for the occupancy and detection features, respectively. Thus, the goal of the learning algorithm is to minimize the negative log-likelihood of our occupancy model as well as the $\ell_{2,1}$-norms.

3.4 Training via Subgradient

A benefit of using neural network based modeling is that the computational tools for neural network-related optimization problems are well-developed. In particular, using a subgradient-based framework and leveraging effective backpropagation-based subgradient computation, the per-iteration complexity of the algorithm is appealing. The maximum likelihood estimation problem is unconstrained, and thus a simple subgradient descent algorithm can be naturally employed. Since the three terms in (7) are all non-differentiable (since the neural networks use the ReLU activation function), subgradient should be used, instead of gradient. More algorithmic details are in the supplement.

4 Experiment Design

We evaluated our model with both simulated and avian point count data. We compared our models with three other approaches, each tuned individually for a peak-to-peak comparison. The code and supplementary material are available at https://github.com/Hutchinson-Lab/StatEcoNet-AAAI21.

4.1 Synthetic Data

We simulated data to evaluate the models' ability to predict probabilities and observations as well as discover important features under the assumed model. We constructed ten features each for the occupancy and detection components, but only the first five features had non-zero coefficients (i.e., each sub-model had five irrelevant features). This setting is for testing the effectiveness of the feature selection layer in `StatEcoNet`. We generated data with both linear and nonlinear effects of the features on the occupancy and detection probabilities. In total, we simulated training and validation sets from the eight combinations of $M \in \{100, 1000\}$, $T \in \{3, 10\}$, and feature-occupancy/detection model $\in$ {linear, nonlinear}. Testing sets always had $M = 1000$ for more robust performance estimates. More detailed simulation settings can be seen in the supplemental material.

4.2 Avian Point Count Data

We also analyzed data on bird distributions to evaluate the proposed method on real-world datasets. We analyzed 10,845 5-minute point count bird surveys extracted from the Oregon 2020 dataset collected in Oregon, United States (Robinson et al. 2020). Surveys were conducted during the bird breeding season (May 15-July 10) by trained field ornithologists from 2011 to 2019. We selected five common

Oregon species for this analysis. Common Yellowthroat (*Geothlypis trichas*), Eurasian Collared-Dove (*Streptopelia decaocto*), Pacific Wren (*Troglodytes pacificus*), Song Sparrow (*Melospiza melodia*), and Western Meadowlark (*Sturnella neglecta*) vocalize frequently during the breeding season and have conspicuous, easily identifiable vocalizations. These species have very different habitat preferences (see supplement for more details). Tab. 1 shows statistics of our datasets (i.e., the percentage of the sites and surveys with positive observations of the species).

Species	Percent observed	
	sites	surveys
Common Yellowthroat (COYE)	19.5%	10.7%
Eurasian Collared-Dove (EUCD)	14.0%	8.2%
Pacific Wren (PAWR)	24.3%	14.5%
Song Sparrow (SOSP)	45.8%	27.8%
Western Meadowlark (WEME)	15.5%	12.2%

Table 1: Species analyzed and the percent of sites (of 942 total) and surveys (of 942 sites $\times$ 3 visits per site $=$ 2,826 total) with positive observations of the species.

Before fitting models, we acquired site and survey features, grouped observations into sites, and divided the data for cross-validation. We constructed 28 environmental features describing the sites from Landsat satellite image composites (details in supplement). The observation-related features were year, day, and time of observation, to capture time-varying detectability. For bird datasets, we consider both environmental and observation-related features as detection features because the site-specific information can affect species detectability. Though Oregon 2020 did not explicitly pre-define sites with multiple visits, its clustered sampling design simplified survey-to-site-assignment. We pre-processed the data by excluding sites that were only surveyed one or two times, and for sites visited more than three times, we randomly selected three surveys. This resulted in a total of 942 sites. We divided these data into three spatially distinct cross-validation folds (Valavi et al. 2018). The site distribution and fold assignments are shown in Fig. 3.

4.3 Performance Metrics

We evaluated model quality along several dimensions. We measured the Pearson correlation coefficient between the true and estimated model probabilities for the simulated datasets. For predicting held-out observations, we measured both the area under the Receiver Operating Characteristic Curve (AUROC) and the area under the Precision-Recall Curve (AUPRC). Note that the probability of a positive observation is the product of the occupancy and detection probabilities. In this case, AUPRC may be preferred over AUROC since AUPRC is better suited to class-imbalanced data (Davis and Goadrich 2006) (Tab. 1). With synthetic datasets, we compared the features that each model selects based on its own relative influence scores against the truly relevant features in the data-generation procedure. When applying `StatEcoNet` to the avian datasets, we present the ℓ_2-norms of $\mathbf{U}_1(:, j)$ and $\mathbf{V}_1(:, k)$ as the indicators of the

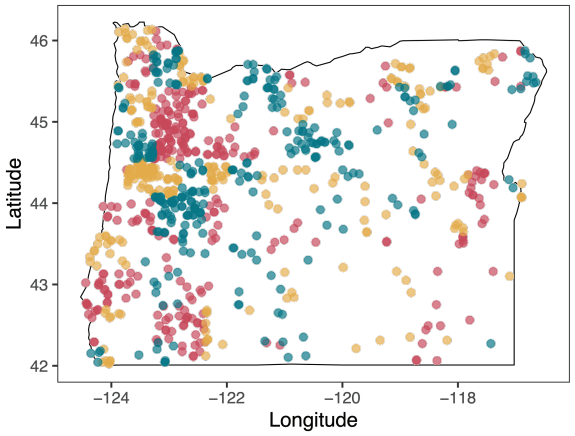

Figure 3: Map of the survey sites over Oregon, United States. Each site had at least three surveys. Colors indicate assignment of sites into three folds (training, validation, test) for Western Meadowlark.

importance of the features $[\mathbf{x}_i]_j$ and $[\mathbf{w}_{it}]_k$, respectively. In `OD-BRT`, we present the number of times that each feature was selected as a split variable as the indicator of feature importance. Finally, we compare the performance of the models by measuring training time. We repeated experiments 5 times for synthetic datasets and 10 times for avian point count datasets, and summarized the performance evaluation metrics with mean and standard deviation values.

4.4 Baselines and Parameter Tuning

We compared the proposed method to three baselines: `OD-LR` (MacKenzie et al. 2002), `OD-BRT` (Hutchinson, Liu, and Dietterich 2011), and `OD-1NN` (Joseph 2020). Note that the superior performance of latent variable models compared to standard ML methods (e.g., logistic regression and ensembles of trees) has been demonstrated in the prior work (Hutchinson, Liu, and Dietterich 2011). In addition, the latent variable structure of the occupancy model is critical for scientific inference in ecology. Hence, we focus on comparing our proposed method against two non-NN latent variable models (`OD-LR` for linear model and `OD-BRT` for tree-based nonlinear model) and one alternative of NN-based latent variable model (`OD-1NN`). The parameter tuning strategies for the methods under comparison are as follows. For `StatEcoNet`, we selected the key parameters, i.e., learning rate, batch size, number of neurons per layer, and number of layers, from $\{0.01, 0.001, 0.0001\}$, $\{32, all\}$, $\{8, 16, 32, 64\}$, and $\{1, 3\}$, respectively, to maximize the AUPRC performance on the validation set. Similarly, we tuned all parameters for the baselines. For `OD-BRT`, we used Bayesian optimization (Snoek, Larochelle, and Adams 2012; Yan 2016) to tune the shrinkage, bag fraction, tree depth, and number of trees since this method was computationally intensive. The input features of bird species data were normalized for all methods except `OD-BRT`, as trees based methods do not require this procedure. More details are in the supplemental material.

Method	Training Time	Occ.Prob.Corr.	Det.Prob.Corr.	AUPRC	AUROC
OD-LR	**3.66** ± 3.11 s	0.05 ± 0.001	0.01 ± 0.001	0.32 ± 0.002	0.51 ± 0.001
OD-1NN	30.3 ± 5.15 s	**0.84** ± 0.01	0.004 ± 0.003	0.39 ± 0.004	0.61 ± 0.01
OD-BRT	320 ± 60.6 s	0.83 ± 0.01	**0.97** ± 0.002	**0.53** ± 0.003	0.72 ± 0.002
StatEcoNet	94.2 ± 17.5 s	**0.84** ± 0.01	**0.97** ± 0.003	**0.53** ± 0.001	**0.73** ± 0.003

Table 2: Performance metrics (mean ± st. dev.) on simulated data with $M = 1000$, $T = 10$, and nonlinear relationships.

	COYE		EUCD		PAWR		SOSP		WEME	
Method	mean	st.dev	mean	st.dev	mean	st.dev	mean	st.dev	mean	st.dev
OD-LR	0.375	0.0614	0.208	0.0462	0.474	0.0382	0.563	0.0230	0.559	0.1320
OD-1NN	0.376	0.0495	0.272	0.0462	0.461	0.0311	0.567	0.0311	0.545	0.1269
OD-BRT	0.369	0.0458	0.183	0.0453	0.473	0.0348	0.558	0.0322	**0.634**	0.0665
StatEcoNet	**0.383**	0.0519	**0.283**	0.0610	**0.496**	0.0314	**0.571**	0.0210	0.593	0.1049

Table 3: AUPRC for the five species on predicting held-out observations. This quantity is what we can measure on these data, since we do not have ground truth for occupancy, but it is not as scientifically interesting. Performance differences are minor.

5 Results

5.1 Simulation Study

Overall, StatEcoNet was more effective than the baseline methods on simulated data. The estimated occupancy and detection probabilities from StatEcoNet were more correlated with the true probabilities than estimates from the other methods. Tab. 2 shows results for a case where the relationships between features and the occupancy/detection probability are nonlinear, $M = 1000$, and $T = 10$; results for a variety of other settings are in the supplemental material. OD-LR's performance suffers since it does not fit nonlinear relationships. OD-1NN estimated detection probabilities poorly, since the occupancy and detection sub-models were confounded in the single network, which may have made the network size unnecessarily large and the model hard to learn. OD-BRT estimated the target probabilities well on nonlinear data, but its training time was more than three times of that used for StatEcoNet. In addition, a perhaps unexpected observation is that OD-BRT struggled to learn the models when the feature-occupancy/detection probability models were linear (see details in the supplemental material). This may reflect difficulties with approximating lines by a 'staircase' of axis-parallel splits.

Fig. 4 shows the parameters learned by StatEcoNet: $\|\mathbf{U}_1(:,j)\|_2$ and $\|\mathbf{V}_1(:,k)\|_2$. StatEcoNet successfully identified most of the truly relevant features, as evidenced by the larger norms of the $\mathbf{U}_1(:,j)$ and $\mathbf{V}_1(:,k)$ corresponding to the relevant features (see more in supplement). This indicates the efficacy of the $\ell_{2,1}$-norm based regularization.

5.2 Avian Point Count Study

Performance evaluation in this study is challenging because ground truth for the model probabilities and feature importances are unknown. We can compare the methods' abilities to predict held-out observations (y_{it}), but it is important to note that *occupancy*, not *observation*, is of primary scientific interest in the model—precisely what we cannot evaluate directly. StatEcoNet outperforms the baseline methods on four of the five species tested (Tab. 3).

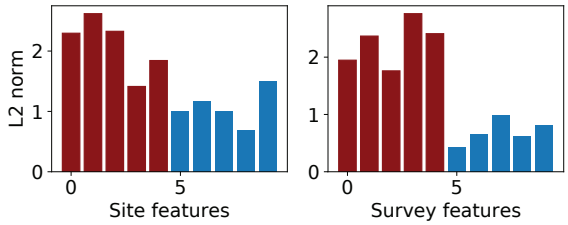

Figure 4: Selected features by StatEcoNet for the synthetic dataset with M=1000, T=10, and nonlinear relationships. The dark red bars correspond to relevant features, and the blue bars irrelevant features.

While impossible to validate exactly, it is illustrative to examine the occupancy and detection probabilities estimated by the different methods on these data. Recall that the predictions are a product of these probabilities (i.e., $\widehat{y}_{it} = \widehat{o}_i\widehat{d}_{it}$). Intuitively, if $\widehat{o}_i$ and $\widehat{d}_{it}$ are estimated correctly, the product $\widehat{o}_i\widehat{d}_{it}$ should be close to the observed events $y_{it} = 1$ (detection) and $y_{it} = 0$ (non-detection) on the test set. To use this intuition for evaluation, consider the Pacific Wren as an example. Fig. 5 shows two-dimensional histograms of the learned occupancy probabilities $\widehat{o}_i$ and detection probabilities $\widehat{d}_{it}$ for each method, separated for the cases of positive and negative observations. The histogram is color coded, where brighter grids mean the corresponding events happen with higher frequencies. Ideally, a good model and learning algorithm would 'light up' the upper right corner of the histogram for $y_{it} = 1$ (first row in Fig. 5), which means that the estimated occupancy probability $\widehat{o}_i$ and detection probability $\widehat{d}_{it}$ can reproduce the held-out detected events. Similarly, for the $y_{it} = 0$ events, an ideal method will make the bottom left corner 'brighter' (second row in Fig. 5).

In Fig. 5, many of the OD-BRT model probability estimates are highly clustered around 0.5. This seems to indicate underfitting and is biologically unrealistic. The OD-LR and OD-1NN histograms did exhibit high frequencies at the

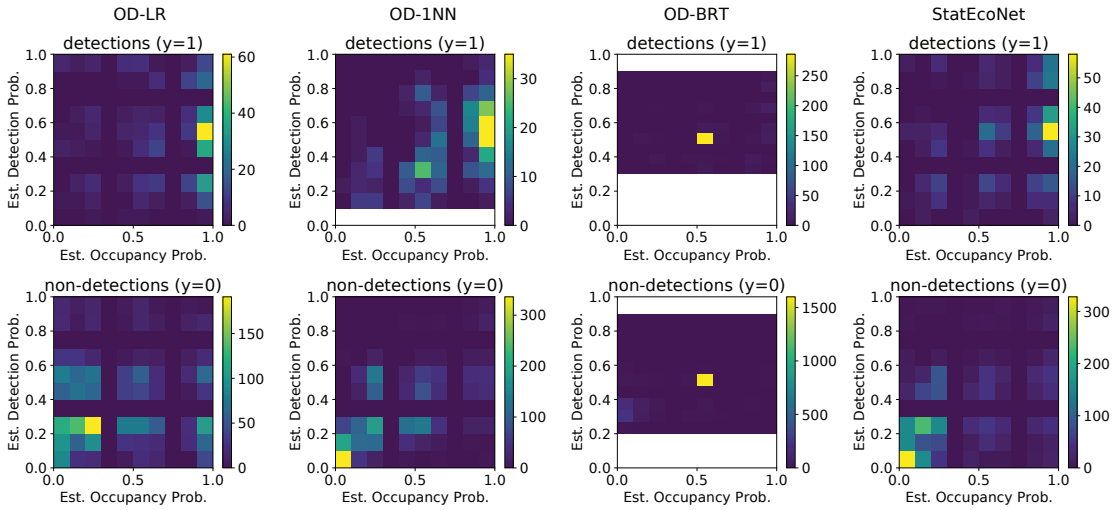

Figure 5: Histograms for Pacific Wren. OD−BRT shows excessive clumping near 0.5. Ground truth is unknown, but StatEcoNet shows more realistic variability than OD−LR and OD−1NN.

upper right and lower left corners for the detection and non-detection events, respectively. However, the events and the learned models are concentrated in a relatively small number of grid cells, making the histograms spiky. This may be pathological since it models the observations with a small number of $\widehat{o}_i$ and $\widehat{d}_{it}$—but different sites and surveys may admit a large variety of $\widehat{o}_i$ and $\widehat{d}_{it}$ in reality. Hence, although these models could have good estimates for the product $\widehat{o}_i\widehat{d}_{it}$ (and thus similar AUPRCs to StatEcoNet), the individual estimates $\widehat{o}_i$ and $\widehat{d}_{it}$ may not be insightful for ecologists. Encouragingly, the histograms from StatEcoNet show more variability—the probabilities concentrate in the desired regions but also gracefully spread out.

Finally, we examined feature importances on the bird datasets. Continuing with the Pacific Wren, Fig. 6 shows the top five site and survey features selected by OD−BRT and StatEcoNet. Interestingly, OD−BRT emphasizes time almost exclusively in the detection model, while StatEcoNet blends the influence of the time-varying features with site-specific environmental features. For both methods, the most important feature was the mean of the land cover index, Tasseled Cap Angle (TCA) at the 75 meter scale. Since this species is found in wet forests with rich undergrowth on the forest floor, this feature may make intuitive sense because TCA is the land cover index that captures the information of both brightness and greenness of land cover, and thus it can indicate dense vegetation (White et al. 2011). Even more promisingly, StatEcoNet selected another land cover index, Tasseled Cap Wetness (TCW) which represents wetness of area. The results for the other four species can be found in the supplemental material.

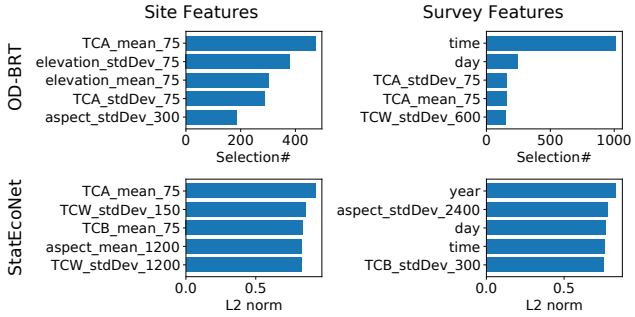

Figure 6: Comparison of feature importances from OD−BRT and StatEcoNet for Pacific Wren (fold 1). The plots on the left (right) show the important site (survey) features selected by each method.

6 Conclusion

This paper contributes StatEcoNet, an interpretable computational framework to integrate the power of neural networks into statistical ecology models that account for the critical challenge of imperfect detection in species distribution modeling. Experiments on simulated datasets showed that StatEcoNet outperforms alternative approaches under various metrics for SDM. In particular, the examination of the learned probabilities and the selected features using real community science data on bird species exhibits intuitively pleasing and encouraging results. Since species distribution models are critical for science and conservation, and imperfect detection and model complexity are pervasive challenges for building these models, StatEcoNet's ability to meet both of these challenges simultaneously has the potential for broad application and impact. In future work, we will aim to maximize this impact by analyzing more species datasets in collaboration with ecologists, improving the optimization procedure for sites with variable numbers of observations, and extending this framework beyond binary characterizations of species distributions.

7 Acknowledgements

We thank Laurel Hopkins, Jing Wang, and Mark Roth for helpful discussions and the anonymous reviewers for their valuable comments. This work was supported in part by the National Science Foundation under projects NSF IIS-1910118 and NSF ECCS 1808159.

8 Ethics Statement

This work has broad positive societal implications. Species distribution models are widely used to develop conservation and management policies for threatened species. In the midst of the sixth global mass extinction, effective actions for slowing biodiversity loss are critical for preserving our fellow inhabitants of Earth, as well as the ecosystem services they may provide to humans. The method contributed in this paper, `StatEcoNet`, offers new capacity in this area by simultaneously fitting nonlinear models for species occupancy and detection probabilities and identifying the features most important to each of those components. Our contributions capitalize on recent advances in neural networks to build more powerful SDMs. Our effort may lead to enhanced understanding of highly complex ecosystems and facilitate more effective conservation policies. This may be particularly critical in an age of drastic climate change, devastating hurricanes, and raging wildfire—the effects of which compound to threaten species persistence globally.

References

Araújo, M. B.; and Peterson, A. T. 2012. Uses and Misuses of Bioclimatic Envelope Modeling. *Ecology* 93(7): 1527–1539.

Betts, M. G.; Rodenhouse, N. L.; Scott Sillett, T.; Doran, P. J.; and Holmes, R. T. 2008. Dynamic Occupancy Models Reveal Within-Breeding Season Movement Up a Habitat Quality Gradient by a Migratory Songbird. *Ecography* 31(5): 592–600.

Cheng, D.; Xiang, S.; Shang, C.; Zhang, Y.; Yang, F.; and Zhang, L. 2020. Spatio-Temporal Attention-Based Neural Network for Credit Card Fraud Detection. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 362–369. New York, New York: AAAI Press.

Cutler, D. R.; Edwards, T. C.; Beard, K. H.; Cutler, A.; Hess, K. T.; Gibson, J.; and Lawler, J. J. 2015. Random Forests for Classification in Ecology. *Ecology* 88(11): 2783–2792.

Davis, J.; and Goadrich, M. 2006. The Relationship between Precision-Recall and ROC Curves. In *Proceedings of the Twenty-Third International Conference on Machine Learning*, 233–240. Pittsburgh, Pennsylvania: ACM Press.

Elith, J.; Graham, C. H.; Anderson, R. P.; Dudik, M.; Ferrier, S.; Guisan, A.; J. Hijmans, R.; Huettmann, F.; Leathwick, J. R.; Lehmann, A.; Li, J.; Lohmann, L. G.; Loiselle, B. A.; Manion, G.; Moritz, C.; Nakamura, M.; Nakazawa, Y.; Overton, J. M. M.; Peterson, A. T.; Phillips, S. J.; Richardson, K.; Scachetti-Pereira, R.; Schapire, R. E.; Soberón, J.; Williams, S.; Wisz, M. S.; and Zimmermann, N. E. 2006. Novel Methods Improve Prediction of Species' Distributions from Occurrence Data. *Ecography* 29(2): 129–151.

Elith, J.; and Leathwick, J. R. 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics* 40(1): 677–697.

Elith, J.; Leathwick, J. R.; and Hastie, T. 2008. A Working Guide to Boosted Regression Trees. *Journal of Animal Ecology* 77(4): 802–813.

Fink, D.; Hochachka, W. M.; Zuckerberg, B.; Winkler, D. W.; Shaby, B.; Munson, M. A.; Hooker, G.; Riedewald, M.; Sheldon, D.; and Kelling, S. 2010. Spatiotemporal Exploratory Models for Broad-Scale Survey Data. *Ecological Applications* 20(8): 2131–2147.

Franklin, J.; and Miller, J. A. 2010. *Mapping Species Distributions: Spatial Inference and Prediction*. Leiden: Cambridge University Press.

Guillera-Arroita, G.; Lahoz-Monfort, J. J.; MacKenzie, D. I.; Wintle, B. a.; and McCarthy, M. a. 2014. Ignoring Imperfect Detection in Biological Surveys Is Dangerous: A Response to 'Fitting and Interpreting Occupancy Models'. *PLoS ONE* 9(7): e99571.

Hutchinson, R. A.; Liu, L.-P.; and Dietterich, T. G. 2011. Incorporating Boosted Regression Trees into Ecological Latent Variable Models. In *Proceedings of the Twenty-fifth AAAI Conference on Artificial Intelligence*, 1343–1348. San Francisco, California: AAAI Press.

Jenatton, R.; Audibert, J.-Y.; and Bach, F. 2011. Structured Variable Selection with Sparsity-Inducing Norms. *Journal of Machine Learning Research* 12: 2777–2824.

Joseph, M. B. 2020. Neural Hierarchical Models of Ecological Populations. *Ecology Letters* 23(4): 734–747.

Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; and Liu, T.-Y. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems 30*, 3146–3154. Long Beach, California: Curran Associates, Inc.

Lahoz-Monfort, J. J.; Guillera-Arroita, G.; and Wintle, B. A. 2014. Imperfect Detection Impacts the Performance of Species Distribution Models. *Global Ecology and Biogeography* 23(4): 504–515.

MacKenzie, D. I.; Nichols, J. D.; Lachman, G. B.; Droege, S.; Andrew Royle, J.; and Langtimm, C. A. 2002. Estimating Site Occupancy Rates When Detection Probabilities Are Less Than One. *Ecology* 83(8): 2248–2255.

MacKenzie, D. I.; Nichols, J. D.; Royle, J. A.; Pollock, K. H.; Bailey, L. L.; and Hines, J. E. 2018. *Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence*. San Diego: Elsevier Science & Technology.

Robinson, W. D.; Hallman, T.; Curtis, J.; and Moore, R. 2020. Oregon 2020 Bird Survey Project. http://oregon2020.com/.

Royle, J. A. 2004. N-mixture Models for Estimating Population Size from Spatially Replicated Counts. *Biometrics* 60(1): 108–115.

Royle, J. A.; and Dorazio, R. 2008. *Hierarchical Modeling and Inference in Ecology: The Analysis of Data from Populations, Metapopulations and Communities*. San Diego: Elsevier Science & Technology.

Shirley, S. M.; Yang, Z.; Hutchinson, R. A.; Alexander, J. D.; McGarigal, K.; and Betts, M. G. 2013. Species Distribution Modelling for the People: Unclassified Landsat TM Imagery Predicts Bird Occurrence at Fine Resolutions. *Diversity and Distributions* 19(7): 855–866.

Snoek, J.; Larochelle, H.; and Adams, R. P. 2012. Practical Bayesian Optimization of Machine Learning Algorithms. In *Advances in Neural Information Processing Systems 25*, 2951–2959. Tahoe City, California: Curran Associates, Inc.

Steffens, C.; Pesavento, M.; and Pfetsch, M. E. 2018. A Compact Formulation for the $\ell_{2,1}$ Mixed-Norm Minimization Problem. *IEEE Transactions on Signal Processing* 66(6): 1483–1497.

Valavi, R.; Elith, J.; Lahoz-Monfort, J. J.; and Guillera-Arroita, G. 2018. blockCV: An R Package for Generating Spatially or Environmentally Separated Folds for k-Fold Cross-Validation of Species Distribution Models. *Methods in Ecology and Evolution* 10(2): 225–232.

White, J. C.; Wulder, M. A.; Gómez, C.; and Stenhouse, G. 2011. A History of Habitat Dynamics: Characterizing 35 Years of Stand Replacing Disturbance. *Canadian Journal of Remote Sensing* 37(2): 234–251.

Yan, Y. 2016. *rBayesianOptimization: Bayesian Optimization of Hyperparameters*. URL https://CRAN.R-project.org/package=rBayesianOptimization. R package version 1.1.0.