StatEcoNet: Statistical Ecology Neural Networks for Species Distribution Modeling

Supplemental Material - Technical Appendix

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

Contents

1 Introduction

This document includes explanations and descriptions of our model training algorithm, generation of the synthetic data, parameter tuning process, and setup of our real-data experiments. In particular, the avian point count datasets of 5 bird species, including the full descriptions of the site and survey features, are detailed in this document. Additional simulation results and real-data experiments on four more bird species can also be found in this document. Sec. 7 also presents discussions on the real-data experiments and insights revealed by the outputs of the algorithms from an ecological study viewpoint.

2 Subgradient Algorithm

Recall that the maximum likelihood estimation problem is as follows:

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

where κ_i is a constant defined as $\kappa_i = \mathbb{1}\left(\sum_{t=1}^{T_i} y_{it} = 0\right)$. The regularized version of our cost function is given by

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

where the ℓ_2/ℓ_1 mixed norm for $\mathbf{Z} \in \mathbb{R}^{m \times n}$ is expressed as follows:

$$\|\mathbf{Z}\|_{2,1} = \sum_{j=1}^{n} \|\mathbf{Z}(:, j)\|_2.$$

As we mentioned, the mixed norm is often used in the literature for feature selection. To put together, our optimization criteria can be summarized as

$$\min_{\boldsymbol{\theta}_G, \boldsymbol{\theta}_F} \quad -\frac{1}{M} \sum_{i=1}^{M} \tilde{\mathcal{L}}_i(\boldsymbol{\theta}_G, \boldsymbol{\theta}_F) + \lambda_F \phi(\boldsymbol{\theta}_F) + \lambda_G \phi(\boldsymbol{\theta}_G), \tag{3}$$

where

$$\tilde{\mathcal{L}}_i(\boldsymbol{\theta}_G, \boldsymbol{\theta}_F) = \log \mathcal{L}_i, \quad \phi(\boldsymbol{\theta}_F) = \|\mathbf{U}_1\|_{2,1}, \quad \phi(\boldsymbol{\theta}_G) = \|\mathbf{V}_1\|_{2,1}.$$

The maximum likelihood estimation problem is unconstrained, and thus a simple subgradient descent algorithm can be naturally employed. Since the three terms in (2) are all non-differentiable (as the neural networks in our construction use the rectified linear unit (ReLU) activation functions), subgradient should be used in optimization, instead of gradient.

In iteration k, the update rule is as follows:

$$\boldsymbol{\theta}^{(k+1)} \leftarrow \boldsymbol{\theta}^{(k)} - \alpha^{(k)} \left(-\partial \tilde{\mathcal{L}}(\boldsymbol{\theta}^{(k)}) + \partial \phi(\boldsymbol{\theta}^{(k)}) \right)$$

where $\boldsymbol{\theta} = [\boldsymbol{\theta}_G^\top, \boldsymbol{\theta}_F^\top]^\top$, $\phi(\boldsymbol{\theta}) = \lambda_F \phi(\boldsymbol{\theta}_F) + \lambda_G \phi(\boldsymbol{\theta}_G)$ and the subgradient $\partial \tilde{\mathcal{L}} = \sum_{i=1}^{M} \partial \tilde{\mathcal{L}}_i$ is computed via the chain rule and backpropagation.

To reduce complexity, $\partial \tilde{\mathcal{L}}$ can be approximated by sample averaging:

$$\partial \tilde{\mathcal{L}}(\boldsymbol{\theta}^{(k)}) \approx \frac{1}{|\mathcal{B}^{(k)}|} \sum_{i \in \mathcal{B}^{(k)}} \partial \tilde{\mathcal{L}}_i(\boldsymbol{\theta}^{(k)}),$$

where $\mathcal{B}^{(k)}$ is a randomly sampled batch of sites such that $\mathcal{B}^{(k)} \subseteq [M]$.

3 Data Simulation Details

We simulated data to evaluate the models' ability to predict probabilities and observations as well as discover important features. Our data generation formula is a mixture of linear and nonlinear components. The equations below show how we generate synthetic data for each site i and survey t. In this simulation setting, we define 10 features for both sites and surveys, and only the first five features are used to generate the responses. That is, there are five irrelevant features in each sub-model.

$$\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}), \tag{4a}$$

$$\mathbf{w}_{it} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}), \tag{4b}$$

$$[\boldsymbol{\alpha}]_k \sim \mathcal{U}(-1, 1) \text{ if } k = 1, \ldots, 5,\ [\boldsymbol{\alpha}]_k = 0,\ \forall k > 5, \tag{4c}$$

$$[\boldsymbol{\beta}]_j \sim \mathcal{U}(-1, 1) \text{ if } j = 1, \ldots, 5,\ [\boldsymbol{\beta}]_j = 0,\ \forall j > 5, \tag{4d}$$

$$o_i = \frac{\exp((1-\rho)\cdot\boldsymbol{\alpha}^T\mathbf{x}_i + \rho\cdot\mathbf{x}_i^T\mathbf{A}\mathbf{x}_i)}{1+\exp((1-\rho)\cdot\boldsymbol{\alpha}^T\mathbf{x}_i + \rho\cdot\mathbf{x}_i^T\mathbf{A}\mathbf{x}_i)}, \tag{4e}$$

$$d_{it} = \frac{\exp((1-\rho)\cdot\boldsymbol{\beta}^T\mathbf{w}_{it} + \rho\cdot\mathbf{w}_{it}^T\mathbf{B}\mathbf{w}_{it})}{1+\exp((1-\rho)\cdot\boldsymbol{\beta}^T\mathbf{w}_{it} + \rho\cdot\mathbf{w}_{it}^T\mathbf{B}\mathbf{w}_{it})}, \tag{4f}$$

Here, $\boldsymbol{\alpha}$ is a coefficient vector on site features ($\mathbf{x}_i$) and $\boldsymbol{\beta}$ is a coefficient vector on survey features ($\mathbf{w}_{it}$), $\mathbf{A}$ and $\mathbf{B}$ are diagonal matrices of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. The value of $\rho = [0, 1]$ indicates the contribution of linear and nonlinear terms in generating synthetic data. When $\rho = 0$, the latent generative models for o_i and d_{it} are linear models, while $\rho = 1$ corresponds to nonlinear models. We sample covariates from the normal distribution to ensure that we have well-balanced probabilities. We sampled the coefficients from the uniform distribution to avoid unbounded values.

We generated training, validation, and test sets from the same formula. We generate different types of datasets with the size of sites (M) and visits (T) and ρ value. We use $M \in \{100, 1000\}$ and $T \in \{3, 10\}$ for training and validation sets and fix the site size for test sets with $M = 1000$ and the corresponding value of T. We also generate datasets using $\rho \in \{0, 1\}$. In total, we have 8 different types of datasets as described in Table 1.

idx	nSites	nVisits	ρ
1	100	3	0
2	100	3	1
3	100	10	0
4	100	10	1
5	1000	3	0
6	1000	3	1
7	1000	10	0
8	1000	10	1

Table 1: Synthetic datasets

4 Avian Point Count Dataset Details

We also analyzed data on bird distributions to evaluate the proposed method on a real dataset. We used 10,845 5-minute point count bird surveys from the Oregon 2020 database (Robinson et al., 2020). Surveys were conducted during the bird breeding season (May 15-July 10) by trained field ornithologists from 2011 to 2019. The survey locations were selected according to a stratified random design to distribute observations across Oregon. Within this design, 3-8 surveys were clustered within one randomly selected 1-square-mile section of each of Oregon's 36-square-mile township. During each survey, all birds were counted and identified to species.

We selected five common Oregon species for this analysis. Common Yellowthroat (*Geothlypis trichas*), Eurasian Collared-Dove (*Streptopelia decaocto*), Song Sparrow (*Melospiza melodia*), Western Meadowlark (*Sturnella neglecta*), and Pacific Wren (*Troglodytes pacificus*), vocalize frequently during the breeding season and have conspicuous, easily identifiable vocalizations. These species have very different habitat preferences. Common Yellowthroat is found in extremely wet vegetation with little canopy cover. Eurasian Collared-Dove is found in human-dominated habitats. Song Sparrow is more of a generalist, and is found in most habitats with rich ground-level vegetation. Western Meadowlark is found in grasslands. Pacific Wren is found in wet forests with rich undergrowth on the forest floor.

4.1 Environmental Features

We compiled features for the models representing both the surrounding environment and the observation conditions. We constructed environmental features from a time series of radiometrically consistent, gap-free Landsat satellite image composites. We aggregated all summertime (Julian days 183 - 243) Landsat Collection 1 Tier 1 surface reflectance images with less than 85% cloud cover and which intersected our study area for processing. We harmonized the Landsat Operational Land Imager data with the Landsat Thematic Mapper and the Landsat Enhanced Thematic Mapper plus data using the reduced major axis regression coefficients from Roy et al. 2016. We removed clouds and cloud shadows from the imagery using the quality assessment band produced by the FMask algorithm (Zhu and Woodcock, 2012; Zhu, Wang, and Woodcock, 2015). We composited each year's worth of satellite imagery into a single image using the medoid method (Flood, 2013). We computed a time series of normalized burn ratio (NBR) images from the annual composites (Key and Benson, 1999). The LandTrendr algorithm, with the NBR time series as input, derived a time series of gap-free, fitted imagery (see Kennedy et al. 2015 for details). We used Google Earth Engine (Gorelick et al., 2017) for all image processing. From the time-series of fitted images 34 spectral indices were computed. Specifically, we used three components (brightness, greenness, wetness) of Tasseled Cap - TCB, TCG, TCW - and Tasseled Cap Angle (TCA) which captures the angle between the TCG and TCB values.

ID	Environmental Features	ID	Environmental Features
1	aspect mean 75	15	aspect stdDev 300
2	aspect stdDev 75	16	TCA stdDev 300
3	elevation mean 75	17	TCB stdDev 300
4	elevation stdDev 75	18	TCW stdDev 300
5	slope stdDev 75	19	aspect mean 600
6	TCA mean 75	20	aspect stdDev 600
7	TCA stdDev 75	21	TCB stdDev 600
8	TCB mean 75	22	TCW stdDev 600
9	TCB stdDev 75	23	aspect mean 1200
10	TCG stdDev 75	24	aspect stdDev 1200
11	TCW stdDev 75	25	TCB stdDev 1200
12	aspect stdDev 150	26	TCW stdDev 1200
13	TCB stdDev 150	27	aspect mean 2400
14	TCW stdDev 150	28	aspect stdDev 2400

Table 2: 28 environmental features used in this paper's experiments. The feature name indicates the land cover index, statistics (mean/stdDev), and radius scale.

4.2 Observation Features

The observation-related features were year, day, and time of observation, to capture time-varying detectability. In the real data experiments, the detection model had both the observation-related features and the environmental features as inputs. Even though the environmental features did not vary across surveys, they could affect detectability (e.g., vegetation affects how the sound of bird calls carries through forest). The feature selection layer of the neural networks provided a mechanism for choosing a sparser set of features.

5 Parameter Tuning Details

The hyper-parameters used for each model and the number and range of values tried per hyper-parameter are described in Tab. 3. The optimal values are selected based on AUPRC performance on the validation set. In this work, we assumed that the regularization weights λ_F (for occupancy component) and λ_G (for detection component) in `StatEcoNet` share the same value ($\lambda_F = \lambda_G = \lambda$).

Tuning Parameter	`OD-LR`	`OD-1NN`	`StatEcoNet`
learningRate	$\{0.0001, 0.001, 0.01\}$		
nEpoch	$[1 - 2000]$ for synthetic datasets, $[1 - 1000]$ for bird datasets		
batchSize		$\{32, all\}$	
nNeurons		$\{8, 16, 32\}$ for synthetic datasets, $\{16, 32, 64\}$ for bird datasets	
nLayers			$\{1, 3\}$
$\ell_{2,1}$-norm weight (λ)			$\{0, 0.001, 0.01\}$

Tuning Parameter	`OD-BRT`
shrinkage	$[0.1 - 1]$
bagFraction	$[0.1 - 1]$
nTrees	$[1 - 1000]$
treeDepth	$[2 - 10]$

Table 3: Tuning parameter values. For the first five rows, we explored combinations of these discrete values in a grid search. For the `OD-BRT` parameters in the bottom three rows, we explored these ranges with Bayesian optimization.

We found that tuning the `OD-BRT` parameters was computationally intensive, so we selected parameters via Bayesian optimization (Snoek, Larochelle, and Adams, 2012), as implemented in the R package `rBayesianOptimization` (Yan, 2016). Since grid search evaluates every combination of the set of tuning parameters, it surely finds the best combination of those values; however, it can be inefficient to evaluate all possible combinations. In contrast, Bayesian optimization searches for parameter values in a range, potentially evaluating parameter values beyond the fixed values used in grid search. This allows for the possibility of finding better combinations of parameter values than those specified by grid search, though it may not always find the optimal values among all possibilities. We found that the Bayesian optimization method found tuning parameter values with higher AUPRC than grid search in less time.

6 Simulation Results

6.1 Linear Latent Model ($\rho = 0$)

6.1.1 Optimal parameters

Model	Hyper-parameter	Optimal Values			
		100x3	100x10	1000x3	1000x10
`OD-LR`	*learningRate*	0.01	0.01	0.01	0.01
`OD-1NN`	*learningRate*	0.001	0.001	0.001	0.001
	batchSize	32	32	32	32
	nNeurons	16	16	16	32
`StatEcoNet`	*learningRate*	0.001	0.001	0.001	0.001
	batchSize	*all*	32	32	32
	nNeurons	8	8	8	8
	nLayers	1	3	1	3
	λ	0	0.01	0.01	0.01
`OD-BRT`	*shrinkage*	0.2399	0.3629	0.1407	0.123
	bagFraction	0.6279	0.5107	0.8759	0.4817
	treeDepth	2	2	4	6

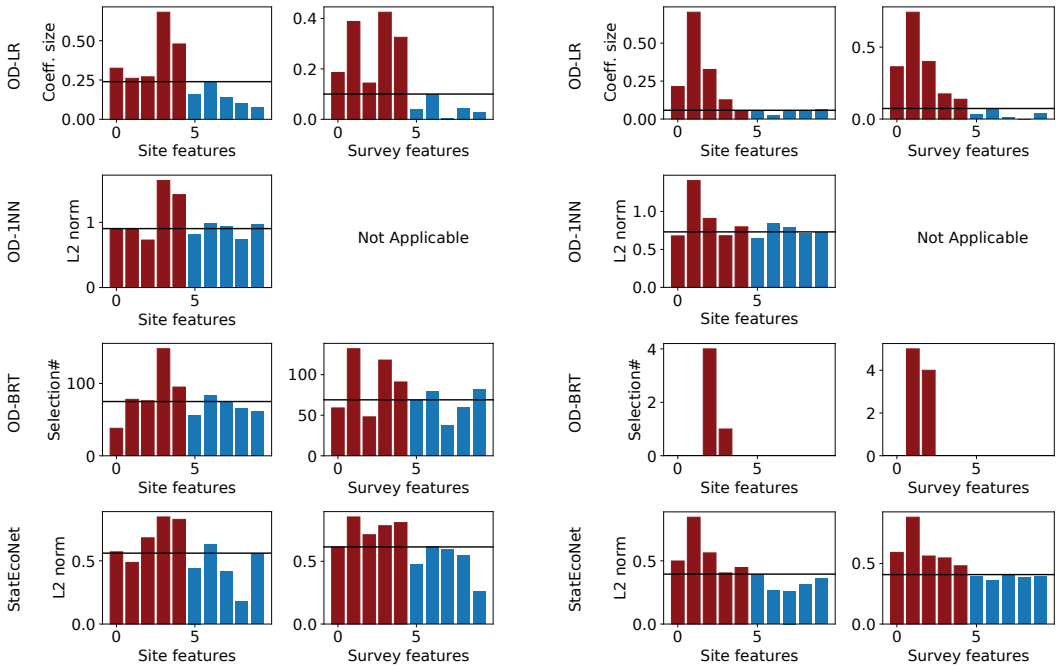
Table 4: Optimal parameters for linear latent models.

6.1.2 Predictive performance

Our model comparisons on simulated data with linear feature combinations indicates that the linear model, `OD-LR`, performs best on linear data, as expected. However, it is rare that all feature relationships would be linear *and* that the modeler would know this in advance. Considering the more general case with unknown feature relationships, the results show that `StatEcoNet` performs similarly to `OD-LR` for recovering the true model probabilities (Tab. 5 correlation columns), predicting new data (Tab. 5 AUPRC and AUROC columns), and selecting the correct features (Fig. 1c). The `OD-1NN` and `OD-BRT` models exhibit problems on some datasets, notably with detection probability correlations (Tab. 5) and occupancy feature selection (Fig. 1b).

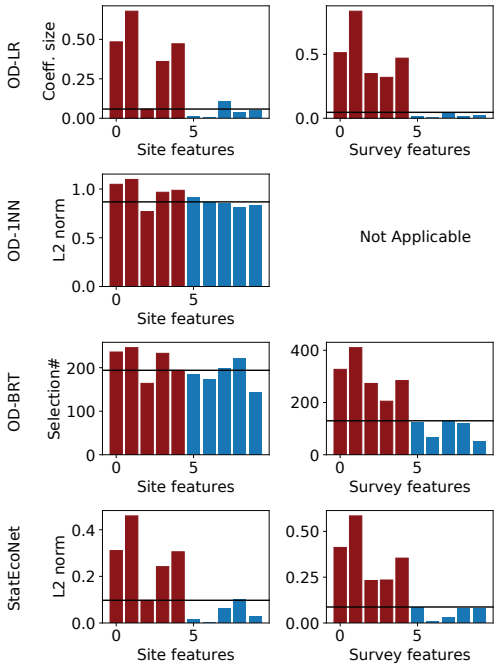
Data size	Method	Training Time	Occ.Prob.Corr.	Det.Prob.Corr.	AUPRC	AUROC
$M = 100$ $T = 3$	`OD-LR`	4.58 ± 3.62	**0.91** ± 0.03	**0.96** ± 0.02	**0.63** ± 0.004	**0.84** ± 0.002
	`OD-1NN`	9.87 ± 2.54	0.86 ± 0.03	0.82 ± 0.02	0.59 ± 0.01	0.81 ± 0.004
	`OD-BRT`	**3.44** ± 2.47	0.78 ± 0.02	0.84 ± 0.02	0.57 ± 0.01	0.80 ± 0.01
	`StatEcoNet`	11.29 ± 3.62	0.87 ± 0.02	0.95 ± 0.01	0.62 ± 0.01	0.83 ± 0.01
$M = 100$ $T = 10$	`OD-LR`	6.16 ± 3.53	**0.93** ± 0.01	**0.98** ± 0.01	**0.71** ± 0.003	**0.87** ± 0.002
	`OD-1NN`	10.10 ± 5.78	0.92 ± 0.03	0.93 ± 0.01	0.67 ± 0.01	0.85 ± 0.01
	`OD-BRT`	**2.66** ± 4.07	0.81 ± 0.03	0.88 ± 0.05	0.62 ± 0.03	0.81 ± 0.02
	`StatEcoNet`	3.51 ± 0.85	**0.93** ± 0.02	0.97 ± 0.01	0.70 ± 0.01	**0.87** ± 0.004
$M = 1000$ $T = 3$	`OD-LR`	20.65 ± 7.60	**0.99** ± 0.0001	**1.00** ± 0.0002	**0.68** ± 0.0003	**0.86** ± 0.0001
	`OD-1NN`	6.75 ± 0.67	0.98 ± 0.002	0.98 ± 0.004	0.66 ± 0.004	**0.86** ± 0.001
	`OD-BRT`	**1.19** ± 0.85	0.77 ± 0.05	0.75 ± 0.02	0.53 ± 0.03	0.76 ± 0.02
	`StatEcoNet`	9.79 ± 5.57	0.98 ± 0.002	**1.00** ± 0.001	**0.68** ± 0.001	**0.86** ± 0.001
$M = 1000$ $T = 10$	`OD-LR`	12.82 ± 8.33	**0.99** ± 0.003	**1.00** ± 0.0004	**0.68** ± 0.001	**0.87** ± 0.001
	`OD-1NN`	6.06 ± 1.21	0.99 ± 0.002	0.99 ± 0.001	0.67 ± 0.001	0.86 ± 0.001
	`OD-BRT`	5.88 ± 1.23	0.86 ± 0.02	0.79 ± 0.01	0.53 ± 0.01	0.78 ± 0.01
	`StatEcoNet`	**5.04** ± 2.34	**0.99** ± 0.003	0.99 ± 0.001	**0.68** ± 0.002	0.86 ± 0.001

Table 5: Performance metrics (mean ± st. dev.) on simulated data with linear relationships.

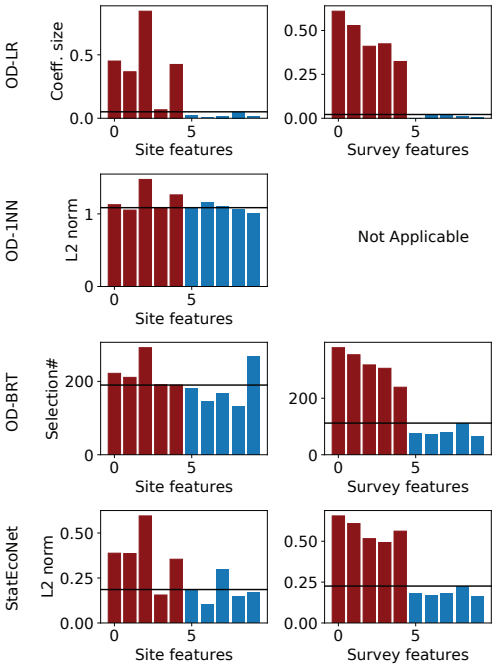


(a) M=100, T=3 (λ=0)

(b) M=100, T=10 (λ=0.01)



(c) M=1000, T=3 (λ=0.01)



(d) M=1000, T=10 (λ=0.01)

Figure 1: Selected features by each method for the synthetic dataset with linear relationships. The red bars correspond to relevant features, and the blue bars irrelevant features. M is the number of training sites and T is the number of visits per site. λ is the optimal regularization weights for λ_F and λ_G. The second plot of `OD-1NN` is not available here because survey features are combined with outputs of a hidden layer from that method. The horizontal black line indicates the top 5 features according to the importance scores (y-axis).

6.2 Nonlinear Latent Model ($\rho = 1$)

6.2.1 Optimal parameters

Model	Hyper-parameter	Optimal Values			
		100x3	100x10	1000x3	1000x10
OD-LR	*learningRate*	0.01	0.01	0.01	0.01
OD-1NN	*learningRate*	0.001	0.001	0.001	0.001
	batchSize	*all*	*all*	32	32
	nNeurons	16	16	32	8
StatEcoNet	*learningRate*	0.001	0.001	0.001	0.001
	batchSize	32	32	*all*	*all*
	nNeurons	32	8	16	16
	nLayers	3	1	3	3
	λ	0.001	0.001	0.01	0.01
OD-BRT	*shrinkage*	0.9440	0.3320	0.5149	0.4040
	bagFraction	0.1435	0.6444	0.7826	0.7499
	treeDepth	5	9	2	3

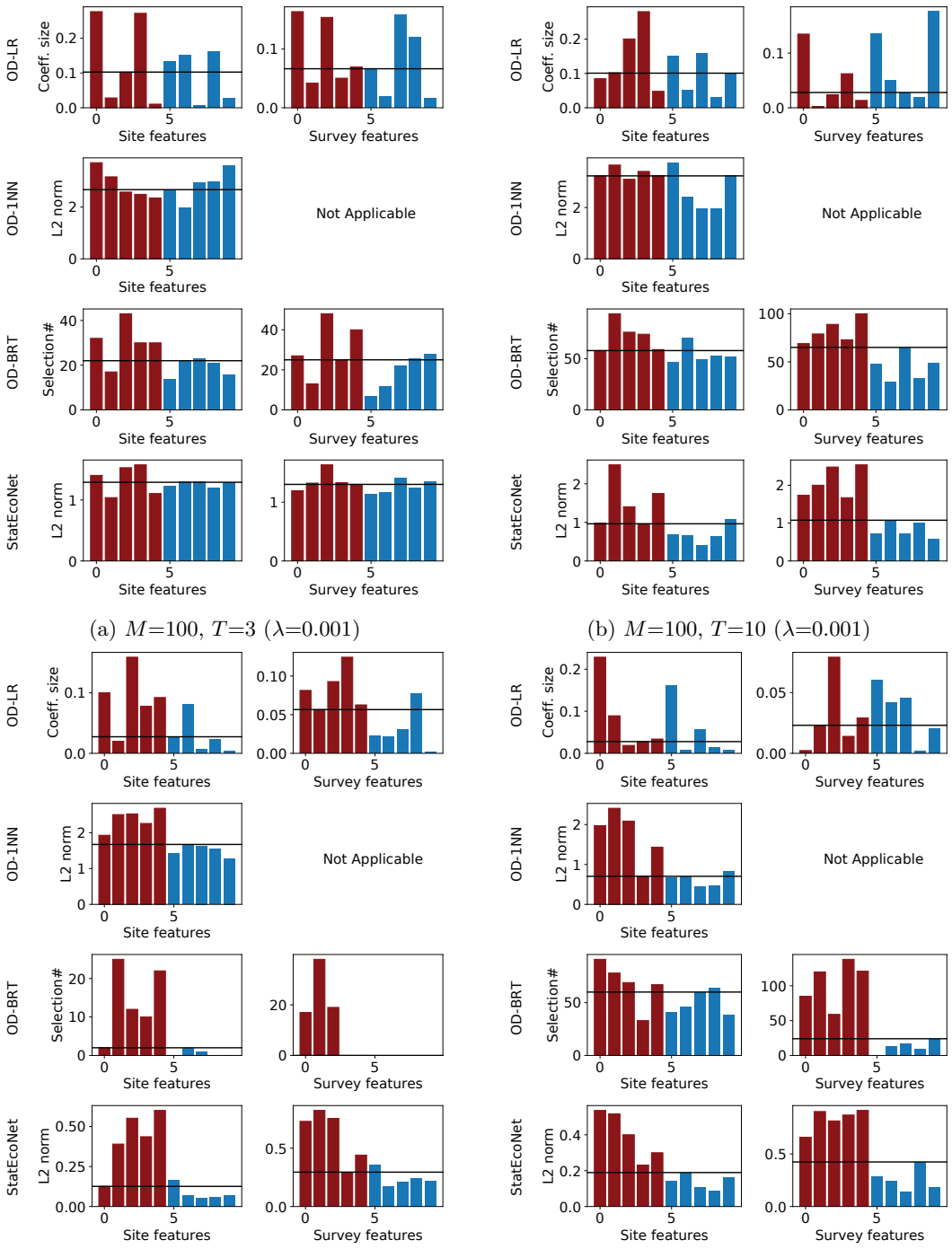
Table 6: Optimal parameters for nonlinear latent models.

6.2.2 Predictive performance

On the simulation experiments where the data generation uses nonlinear feature combinations, StatEcoNet performs well. On only the smallest datasets ($M = 100$), it is outperformed by OD-BRT in terms of recovering the occupancy and detection probabilities as well as predicting new data (Tab. 7). On the larger datasets ($M = 1000$), StatEcoNet performs as well or better than OD-BRT, and the training time starts to favor StatEcoNet heavily as dataset sizes increase. On these larger datasets, StatEcoNet also has an advantage for feature selection.

Data size	Method	Training Time	Occ.Prob.Corr.	Det.Prob.Corr.	AUPRC	AUROC
$M = 100$ $T = 3$	OD-LR	**0.40** ± 0.50	-0.002 ± 0.02	0.003 ± 0.01	0.29 ± 0.004	0.50 ± 0.004
	OD-1NN	5.77 ± 12.33	0.05 ± 0.07	-0.01 ± 0.02	0.29 ± 0.01	0.50 ± 0.02
	OD-BRT	1.48 ± 1.07	**0.38** ±0.11	**0.55** ±0.15	**0.37** ± 0.03	**0.60** ± 0.02
	StatEcoNet	4.91 ± 6.85	0.1 ± 0.12	0.16 ± 0.21	0.31 ± 0.03	0.53 ± 0.04
$M = 100$ $T = 10$	OD-LR	**0.90** ± 0.52	-0.003 ± 0.02	0.01 ± 0.004	0.39 ± 0.01	0.51 ± 0.01
	OD-1NN	16.13 ± 16.08	0.11 ± 0.15	0. ± 0.01	0.39 ± 0.01	0.52 ± 0.01
	OD-BRT	8.07 ± 0.88	**0.59** ± 0.01	**0.80** ± 0.01	**0.53** ± 0.002	**0.66** ± 0.01
	StatEcoNet	39.57 ± 50.71	0.03 ± 0.08	0.31 ± 0.38	0.42 ± 0.06	0.55 ± 0.07
$M = 1000$ $T = 3$	OD-LR	**1.21** ± 1.21	-0.02 ± 0.04	-0.01 ± 0.03	0.35 ± 0.01	0.50 ± 0.01
	OD-1NN	18.96 ± 2.47	0.73 ± 0.02	-0.02 ± 0.01	0.42 ± 0.01	0.59 ± 0.01
	OD-BRT	28.77 ± 22.9	**0.79** ± 0.03	0.88 ± 0.04	**0.55** ± 0.01	**0.70** ± 0.01
	StatEcoNet	25.85 ± 14.16	0.54 ± 0.04	**0.90** ± 0.03	0.53 ± 0.02	**0.70** ± 0.01
$M = 1000$ $T = 10$	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 7: Performance metrics (mean ± st. dev.) on simulated data with nonlinear relationships.



(a) M=100, T=3 (λ=0.001)

(b) M=100, T=10 (λ=0.001)

(c) M=1000, T=3 (λ=0.01)

(d) M=1000, T=10 (λ=0.01)

Figure 2: Selected features by each method for the synthetic dataset with nonlinear relationships. The red bars correspond to relevant features, and the blue bars irrelevant features. M is the number of training sites and T is the number of visits per site. λ is the optimal regularization weights for λ_F and λ_G. The second plot of `OD-1NN` is not available here because survey features are combined with outputs of a hidden layer from that method. The horizontal black line indicates the top 5 features according to the importance scores.

7 Avian Point Count Results

Each subsection below reports more detailed results for each of the five species. Each section gives a histogram of the learned occupancy and detection probabilities, feature importances for the occupancy and detection models across methods, and the optimal hyperparameters that resulted from the tuning process.

There are a few overall trends in the avian point count study results to point out. First, the differences between methods in terms of AUPRC and AUROC are minor (Table 8), even though the interpretations of the learned models vary substantially (visualizations in species-specific subsections below). (The results of `OD-1NN` and `StatEcoNet` have been updated from the main paper with some small changes to the parameter tuning; the main trends are unchanged.) Second, `OD-BRT` sometimes produces probability histograms that are concentrated around 0.5. These seem unrealistic and appear underfit, despite careful parameter tuning. Third, note that the detection feature importance plots are missing for `OD-1NN` for all species because this inference is not available from that method due to the architecture of the neural network.

There are also a few things to consider when viewing the probability histograms below. First, the upper left corners should be interpreted loosely. As the occupancy probability for a given point approaches zero, the contribution of the detection model for that point gets less influence in the likelihood function. Second, variation in detection probability is often biologically plausible (with some exceptions). Finally, for specialists, a low or bimodal distribution of occupancy probabilities for the non-detections makes sense, since sites will be obviously suitable or unsuitable, and some suitable sites may have non-detections.

Species	Metric	`OD-LR`	`OD-1NN`	`OD-BRT`	`StatEcoNet`
COYE	AUPRC	0.375 ± 0.0614	0.376 ± 0.0495	0.369 ± 0.0458	**0.383** ± 0.0519
EUCD	AUPRC	0.208 ± 0.0462	0.272 ± 0.0462	0.183 ± 0.0453	**0.283** ± 0.0610
SOSP	AUPRC	0.563 ± 0.0230	0.567 ± 0.0311	0.558 ± 0.0322	**0.571** ± 0.021
WEME	AUPRC	0.559 ± 0.132	0.545 ± 0.1269	**0.634** ± 0.0665	0.593±0.1049
PAWR	AUPRC	0.474 ± 0.0382	0.461 ± 0.0311	0.473 ± 0.0348	**0.496** ± 0.0314
COYE	AUROC	0.834 ± 0.0355	**0.836** ± 0.0229	0.834 ± 0.0404	0.828 ± 0.0375
EUCD	AUROC	0.756 ± 0.0325	**0.809** ± 0.03	0.72 ± 0.0709	**0.809** ± 0.021
SOSP	AUROC	0.797 ± 0.0175	0.801 ± 0.0192	0.802 ± 0.0185	**0.803** ± 0.0152
WEME	AUROC	0.881 ± 0.0516	0.891 ± 0.0416	**0.912** ± 0.0283	0.910 ± 0.0292
PAWR	AUROC	0.858 ± 0.0178	0.865 ± 0.0218	0.868 ± 0.0309	**0.875** ± 0.026

Table 8: Predictive performance of methods for five bird species.

7.1 Common Yellowthroat (COYE)

Common Yellowthroat (COYE) is found in extremely wet vegetation with little canopy cover. Like all songbirds, it sings more in the early morning than later in the day, so it is more frequently detected on early surveys.

Figure 3 shows two-dimensional histograms of the occupancy and detection probabilities for all positive species reports (detections, $y = 1$) in the top row, and all negative species reports (non-detections, $y = 0$) in the bottom row. The concentration of negatives in the lower left corner of the histograms of `StatEcoNet` and `OD-1NN` may reflect the fact that much of the surveyed points are not suitable habitat for this species, so many occupancy probabilities should be low. In contrast, `OD-LR` is less believable, with many negatives having high occupancy probability, implying that the species was missed more frequently than is realistic.

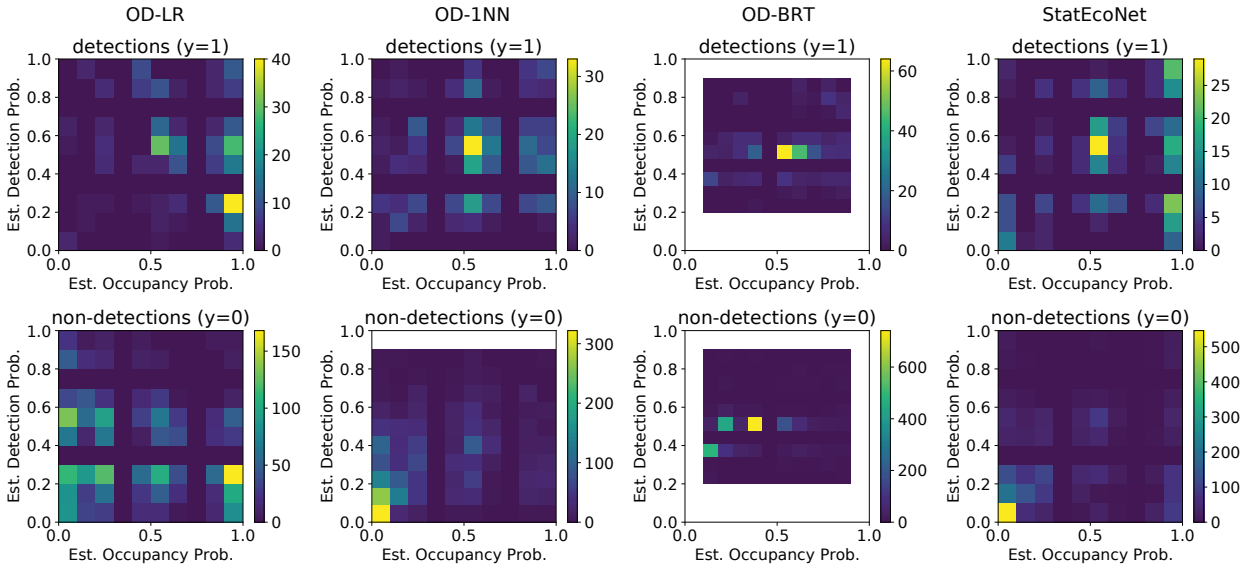


Figure 3: Histograms for Common Yellowthroat. The top row shows the occupancy and detection probability histograms for positives (detections, $y = 1$), and the bottom row shows the same for negatives (non-detections, $y = 0$).

Figures 4 and 5 show the top five most important variables learned by each method for COYE. Mean elevation was consistently among the top site features, which fits with field observations that this species utilizes wetland and riparian habitats. Such habitats of sufficient size for this species are often found at lower elevations. Inclusion of standard deviations of TCA and TCW probably relate to the contrast between reflectance of water versus adjacent wetland habitats.

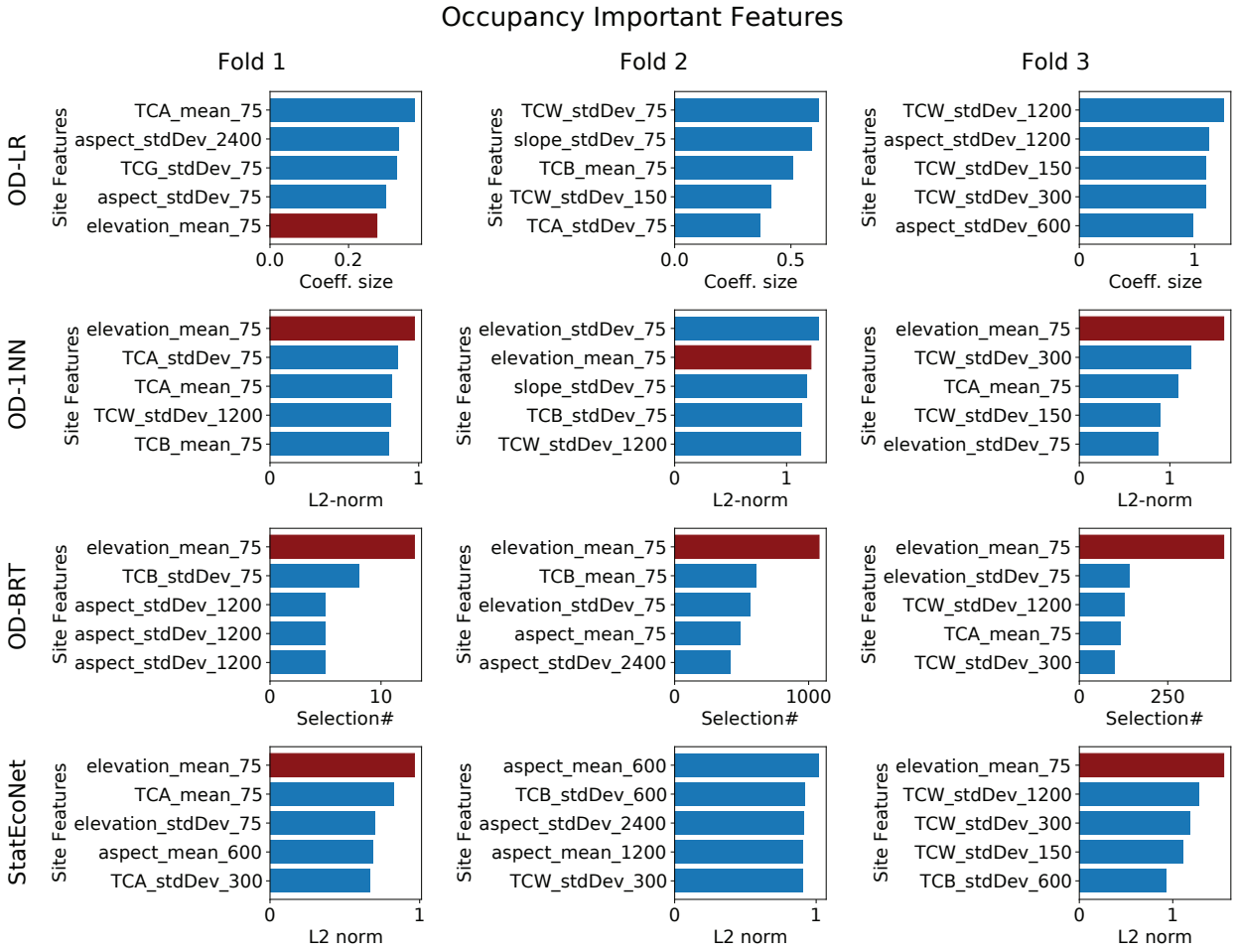


Figure 4: Occupancy feature importances for Common Yellowthroat. The top five features per method per fold are plotted. Note that the x-axes differ across methods. The feature corresponding to the mean elevation at the 75 m scale (chosen as an example feature that is important for `StatEcoNet`) is shaded red to highlight differences across methods.

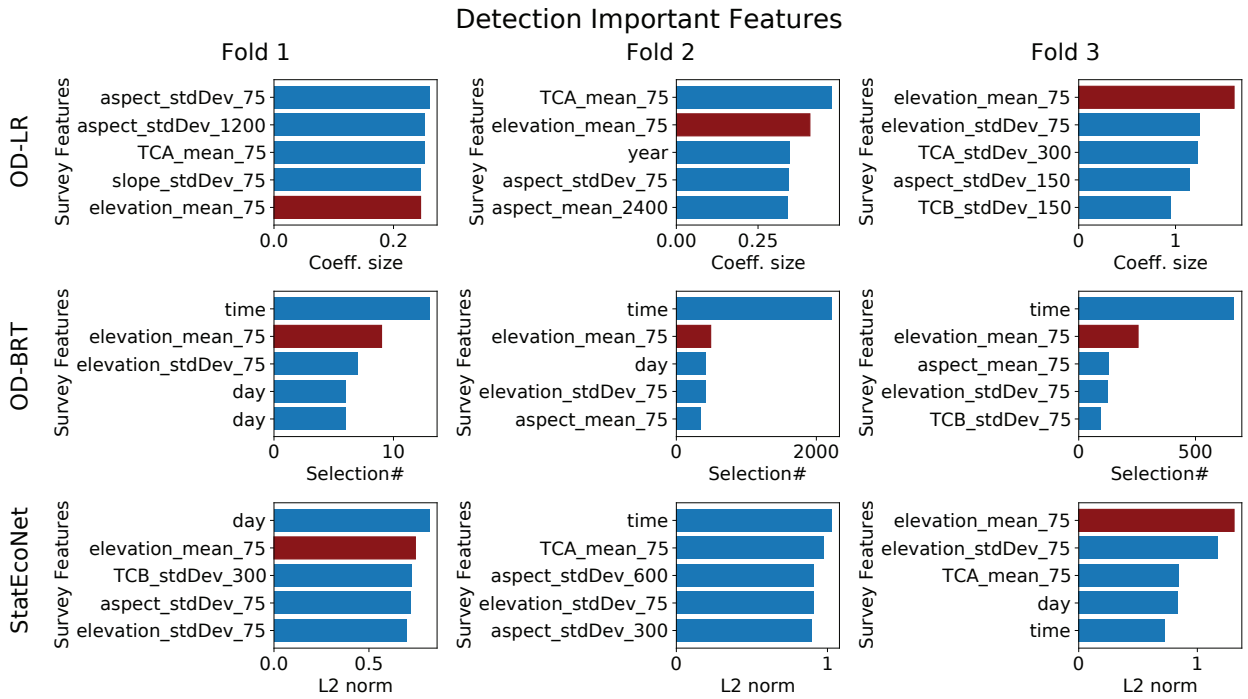


Figure 5: Detection feature importances for Common Yellowthroat. The top five features per method per fold are plotted. Note that the x-axes differ across methods. The feature corresponding to the mean elevation at the 75 m scale (chosen as an example feature that is important for `StatEcoNet`) is shaded red to highlight differences across methods. `OD-1NN` is not included here because the importance of environmental features to the detection model is not available from that method.

Model	Hyper-parameter	Optimal Values		
		Fold 1	Fold 2	Fold 3
`OD-LR`	$learningRate$	0.01	0.01	0.01
`OD-1NN`	$learningRate$	0.001	0.001	0.001
	$batchSize$	32	32	32
	$nNeurons$	32	64	16
`StatEcoNet`	$learningRate$	0.001	0.001	0.001
	$batchSize$	32	32	32
	$nNeurons$	32	64	32
	$nLayers$	1	3	1
	λ	0	0.001	0.001
`OD-BRT`	$shrinkage$	0.7274	0.4756	0.2038
	$bagFraction$	0.4633	0.9526	0.9429
	$treeDepth$	3	10	10

Table 9: Optimal parameters per fold for Common Yellowthroat

7.2 Eurasian Collared-Dove (EUCD)

Eurasian Collared-Dove (EUCD) is found in human-dominated habitats. When present, it is usually easy to identify both visually and aurally. However, in noisy urban areas, its calls may be drowned out by other sounds.

Figure 6 shows two-dimensional histograms of the occupancy and detection probabilities for all positive species reports (detections, $y = 1$) in the top row, and all negative species reports (non-detections, $y = 0$) in the bottom row. Here, the bimodality of the occupancy probabilities (`OD-LR`, `OD-1NN`, `StatEcoNet`) makes sense, as human-dominated habitats are relatively easy to distinguish. `StatEcoNet` shows most detections as having high occupancy and detection probabilities, most non-detections with low occupancy and detection probabilities; this makes sense for a highly-detectable bird with an easily distinguishable habitat. The secondary concentration of sites that are highly likely to be occupied but with very low detection probabilities could be sites where noise pollution impedes detection.

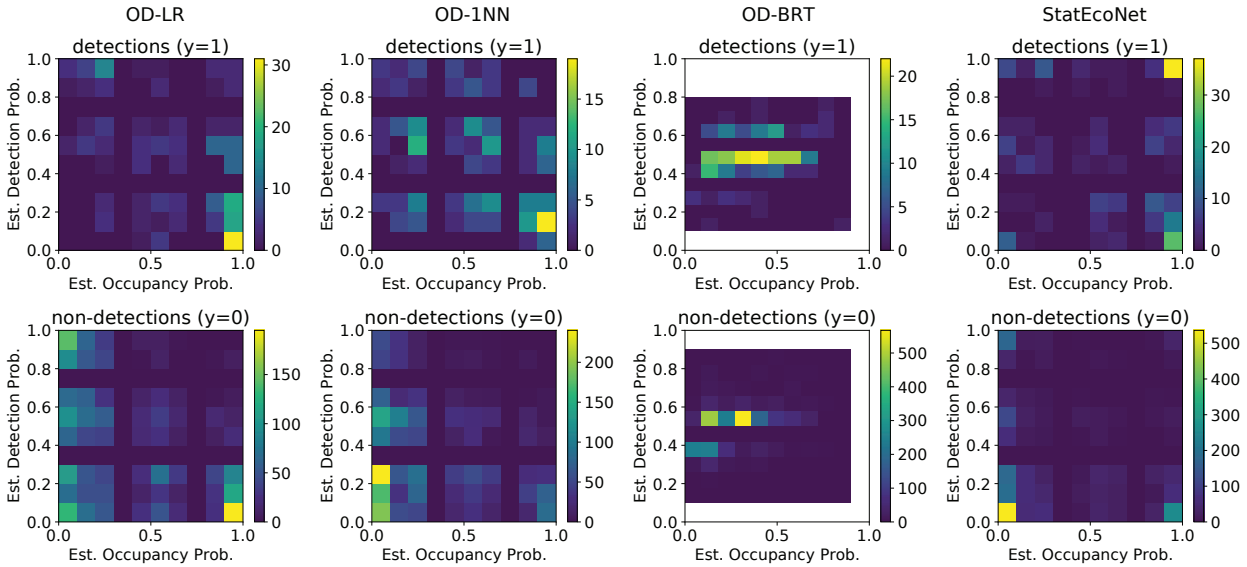


Figure 6: Histograms for Eurasian Collared-Dove.

Figures 7 and 8 show the top five most important variables learned by each method for EUCD. Eurasian Collared-Doves tend to be most numerous around small homesteads (barns, homes) surrounded by agricultural habitats, which is reflected in the identification of TCW standard deviations as important site features. They also are numerous in suburbanized settings, which are captured well by TCA and TCB.

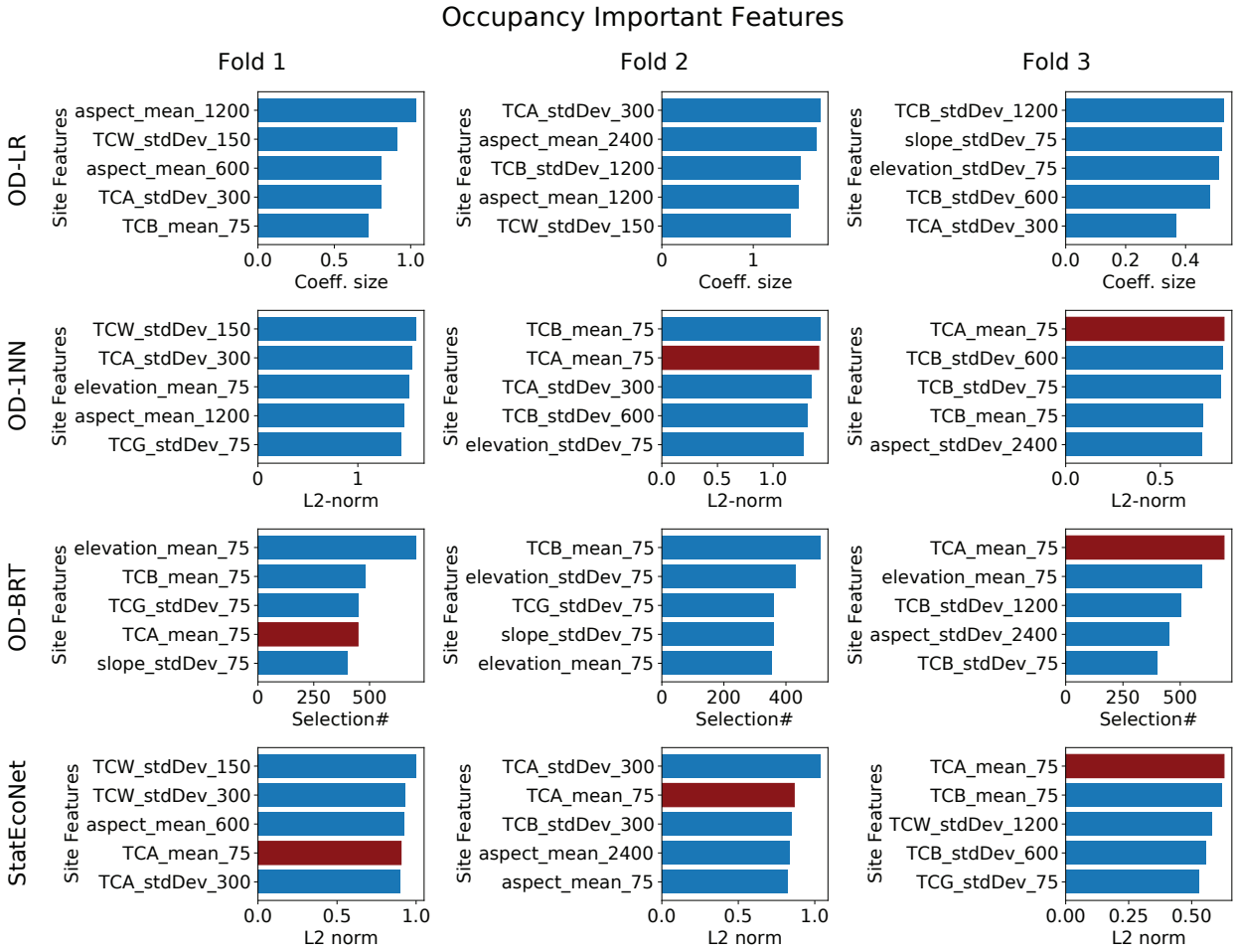


Figure 7: Occupancy feature importances for Eurasian Collared-Dove. The top five features per method per fold are plotted. Note that the x-axes differ across methods. The feature corresponding to the mean TCA at the 75 m scale (chosen as an example feature that is important for `StatEcoNet`) is shaded red to highlight differences across methods.

Figure 8: Detection feature importances for Eurasian Collared-Dove. The top five features per method per fold are plotted. Note that the x-axes differ across methods. The feature corresponding to the year (chosen as an example feature that is important for `StatEcoNet`) is shaded red to highlight differences across methods. `OD-1NN` is not included here because the importance of environmental features to the detection model is not available from that method.

Model	Hyper-parameter	Optimal Values		
		Fold 1	Fold 2	Fold 3
`OD-LR`	*learningRate*	0.01	0.01	0.01
`OD-1NN`	*learningRate*	0.001	0.001	0.001
	batchSize	*all*	32	32
	nNeurons	64	64	16
`StatEcoNet`	*learningRate*	0.001	0.001	0.001
	batchSize	32	32	32
	nNeurons	64	32	16
	nLayers	3	3	3
	λ	0.001	0	0.001
`OD-BRT`	*shrinkage*	0.8073	0.4250	0.8064
	bagFraction	0.7758	0.7055	0.9508
	treeDepth	10	7	8

Table 10: Optimal parameters per fold for Eurasian Collared-Dove

7.3 Song Sparrow (SOSP)

Song Sparrow (SOSP) is found in most habitats with rich ground-level vegetation. It is usually in wet areas, occasionally restricted to riparian zones, but also found in residential areas with lush vegetation. It can be quite prevalent in some habitats.

Figure 9 shows two-dimensional histograms of the occupancy and detection probabilities for all positive species reports (detections, $y = 1$) in the top row, and all negative species reports (non-detections, $y = 0$) in the bottom row. Here, the `OD-BRT` histogram of non-detections concentrating on very low occupancy probabilities seems to imply that almost all of the occupied sites had detections; this is improbable. For the other models, the bimodality of the non-detection occupancy probabilities indicates that the models are finding good separation between habitat and non-habitat and explaining non-detections in good habitat with low detection probabilities.

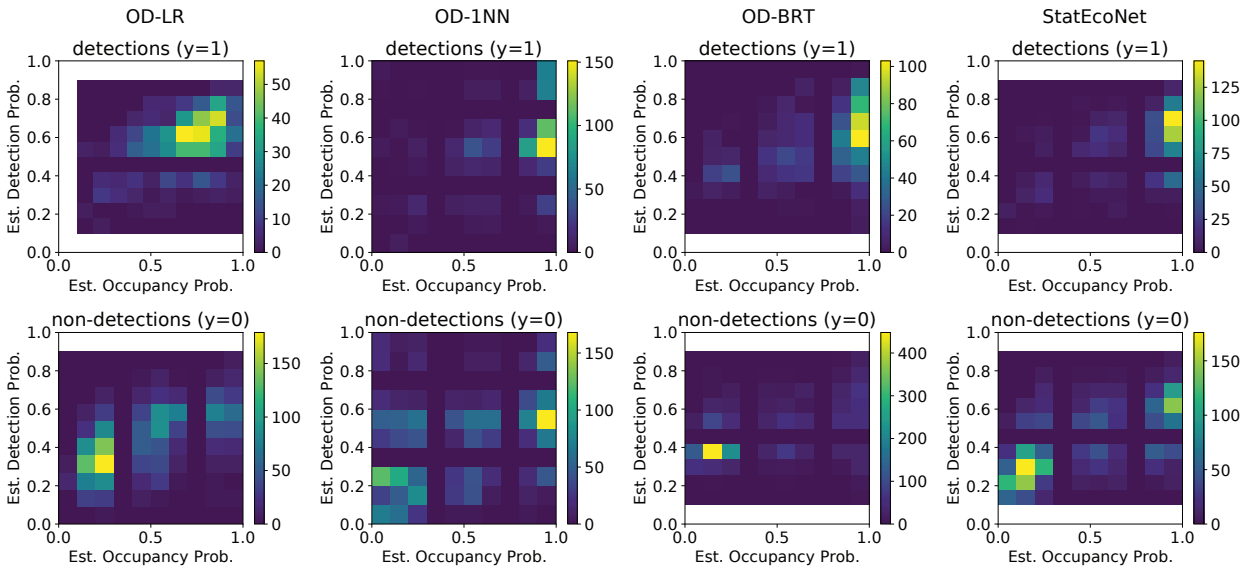


Figure 9: Histograms for Song Sparrow.

Figures 10 and 11 show the top five most important variables learned by each method for SOSP. Song Sparrows are widely distributed common species associated with riparian habitats, suburban habitats and early successional habitats. Most approaches accurately detected that Song Sparrows most often occur at lower elevations. Because they occupy a wide variety of habitats, specific habitat reflectance features did not consistently emerge across the four analytical approaches, although consistency across the 3 folds was better for `StatEcoNet`.

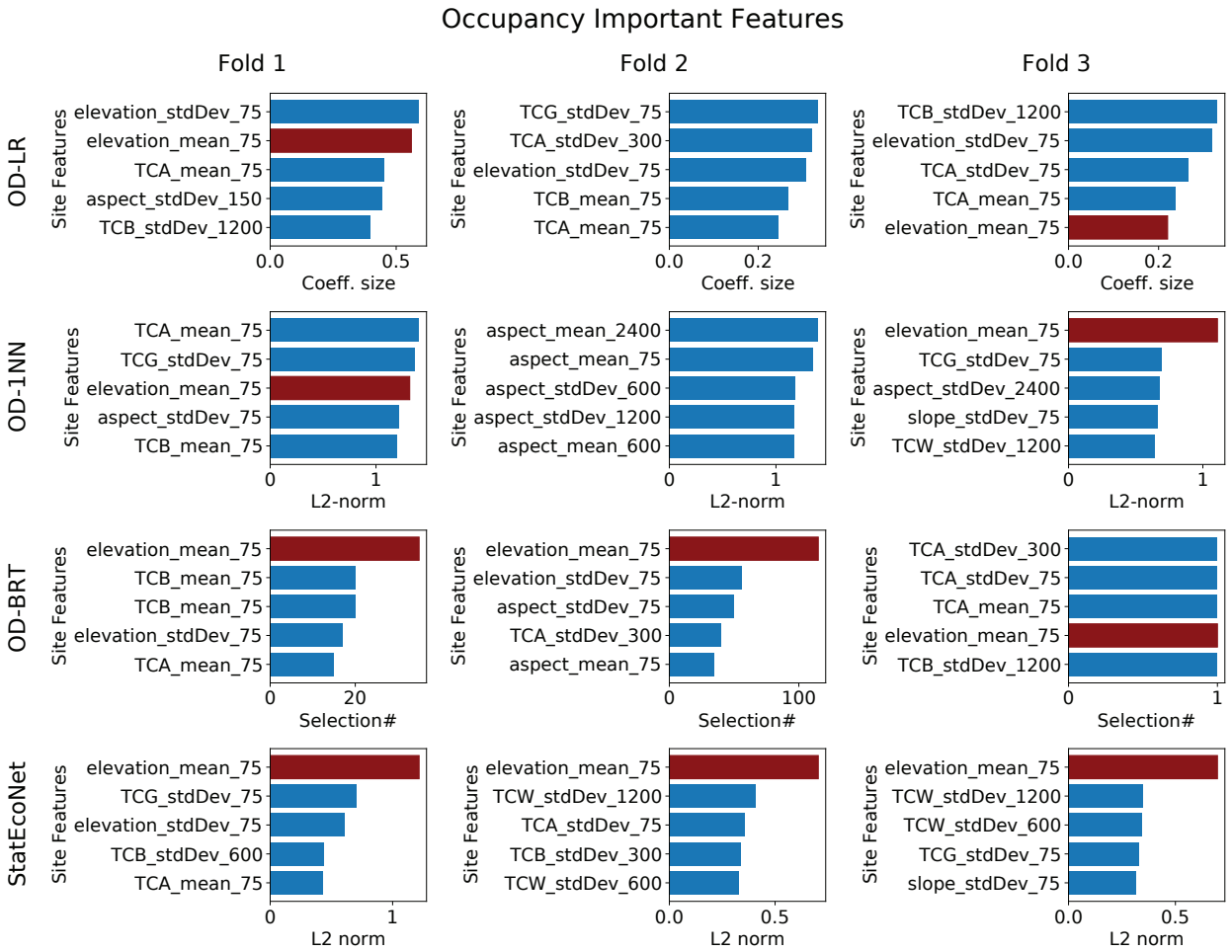


Figure 10: Occupancy feature importances for Song Sparrow. The top five features per method per fold are plotted. Note that the x-axes differ across methods. The feature corresponding to the mean elevation at the 75 m scale (chosen as an example feature that is important for `StatEcoNet`) is shaded red to highlight differences across methods.

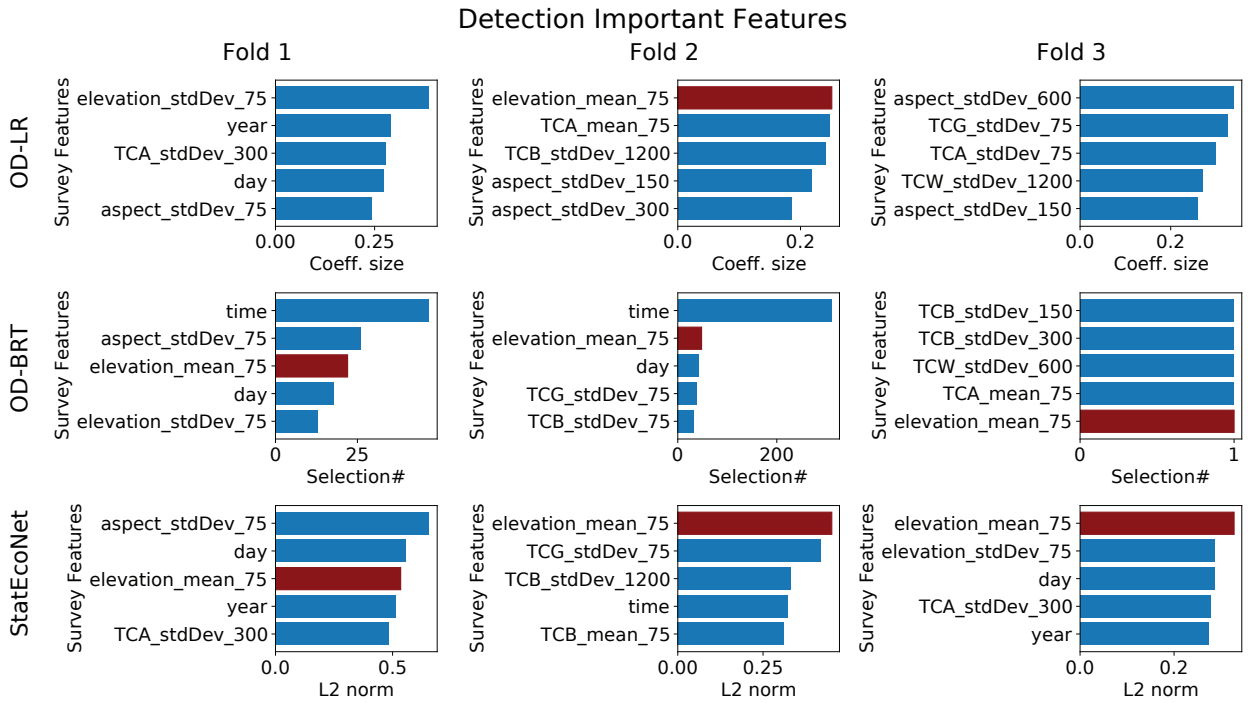


Figure 11: Detection feature importances for Song Sparrow. The top five features per method per fold are plotted. Note that the x-axes differ across methods. The feature corresponding to the mean elevation (chosen as an example feature that is important for `StatEcoNet`) is shaded red to highlight differences across methods. `OD-1NN` is not included here because the importance of environmental features to the detection model is not available from that method.

Model	Hyper-parameter	Optimal Values		
		Fold 1	Fold 2	Fold 3
`OD-LR`	*learningRate*	0.01	0.01	0.01
`OD-1NN`	*learningRate*	0.001	0.001	0.001
	batchSize	32	*all*	32
	nNeurons	16	32	16
`StatEcoNet`	*learningRate*	0.001	0.001	0.001
	batchSize	32	32	32
	nNeurons	64	16	16
	nLayers	1	3	3
	λ	0.01	0.01	0.01
`OD-BRT`	*shrinkage*	0.6779	0.6801	0.8664
	bagFraction	0.9158	0.2259	0.7236
	treeDepth	3	4	5

Table 11: Optimal parameters per fold for Song Sparrow

7.4 Western Meadowlark (WEME)

Western Meadowlark (WEME) strongly specializes on grasslands. Grassland habitat should be more easily distinguishable from our remotely sensed features than some other habitat types (e.g., different types of forest). WEME is one of the most available species for detection in the early morning and can be heard from 1 km away. Since all counts in this dataset were conducted in the morning, high detection probabilities for positive observations make sense.

Figure 12 shows two-dimensional histograms of the occupancy and detection probabilities for all positive species reports (detections, $y = 1$) in the top row, and all negative species reports (non-detections, $y = 0$) in the bottom row. The `StatEcoNet` histograms here are quite concentrated, but this may reflect the high detectability of this species and the ease with which its habitat is distinguished by the remote sensing features. The non-detections with high occupancy probability and low detection probability (lower right corner) may be areas of the Willamette valley that do have grassland habitat, but that do not host WEME because they are only small patches of grassland; these would appear to the model as highly likely to be occupied but with low detection probability.

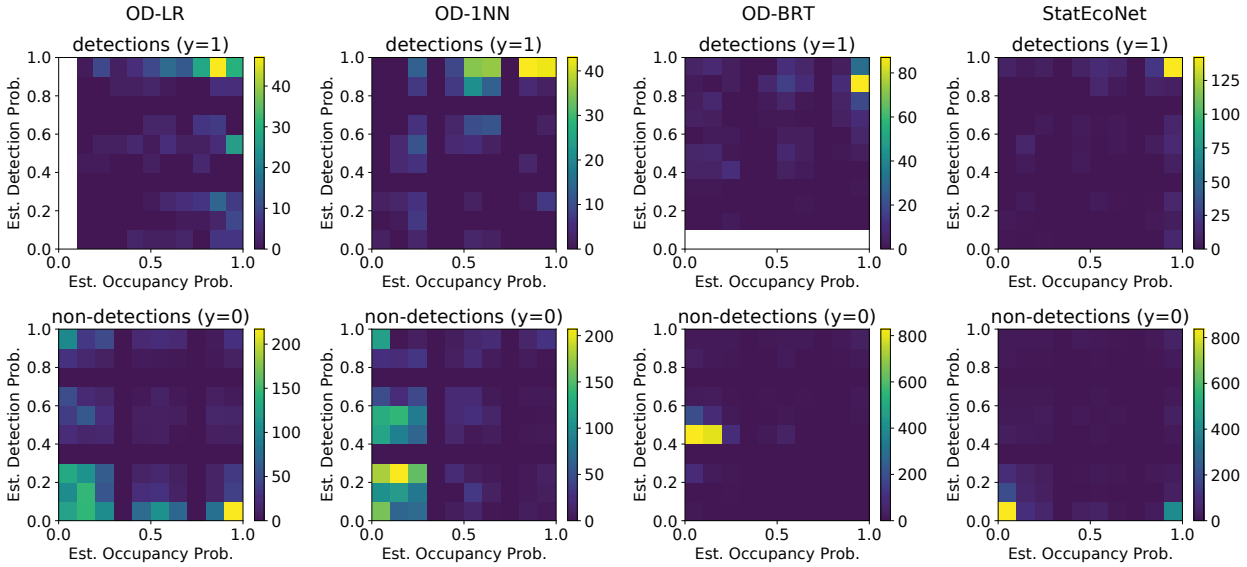


Figure 12: Histograms for Western Meadowlark.

Figures 13 and 14 show the top five most important variables learned by each method for WEME. Western Meadowlarks inhabit grasslands and sagebrush of large extent, avoiding smaller patches or tracts composed largely of agricultural grasslands. Emergence of mean TCA at small buffers (75 m) possibly is related to habitat quality as greenness (moist, productive grasslands) is an important contributor to TCA.

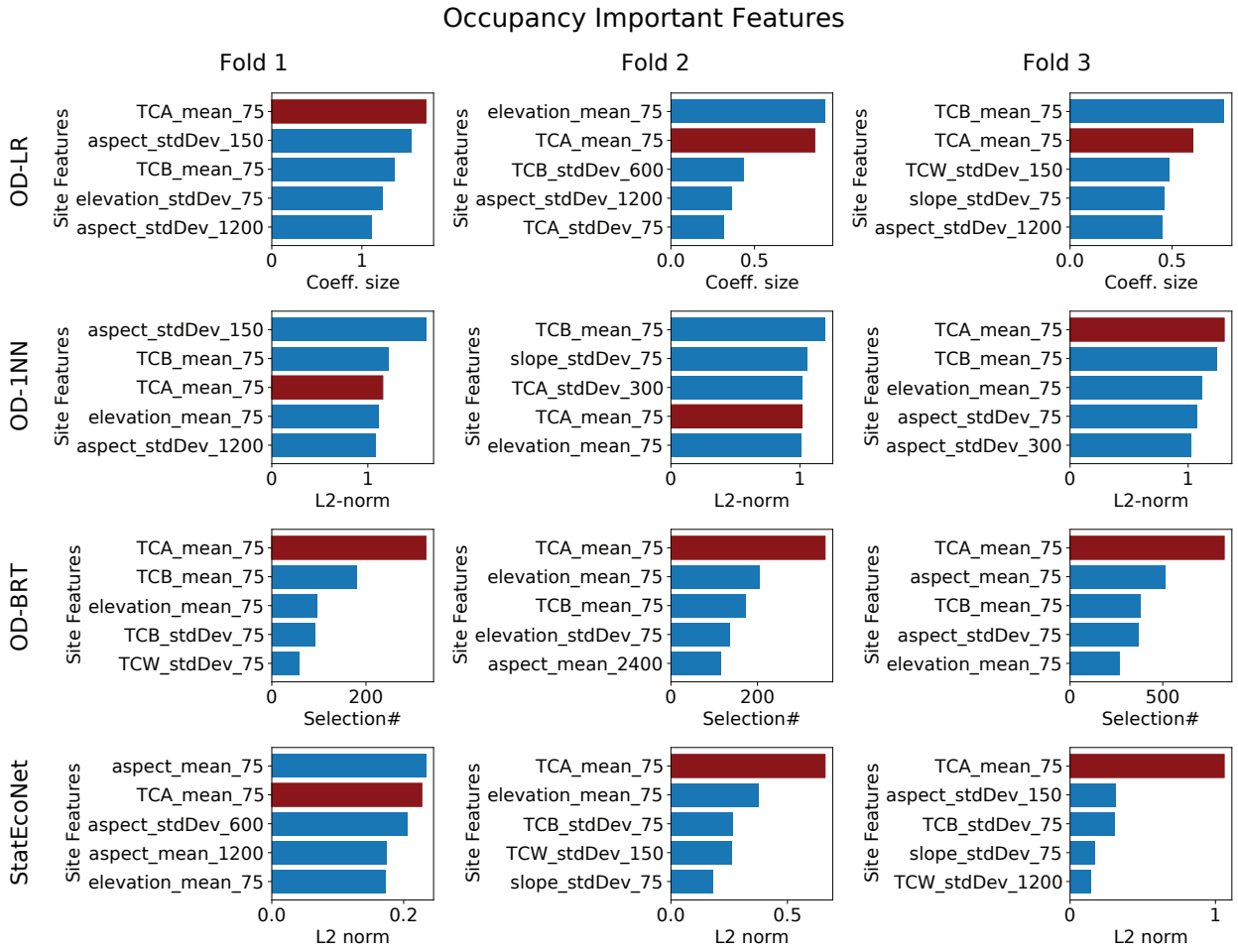


Figure 13: Occupancy feature importances for Western Meadowlark. The top five features per method per fold are plotted. Note that the x-axes differ across methods. The feature corresponding to the mean TCA at the 75 m scale (chosen as an example feature that is important for `StatEcoNet`) is shaded red to highlight differences across methods.

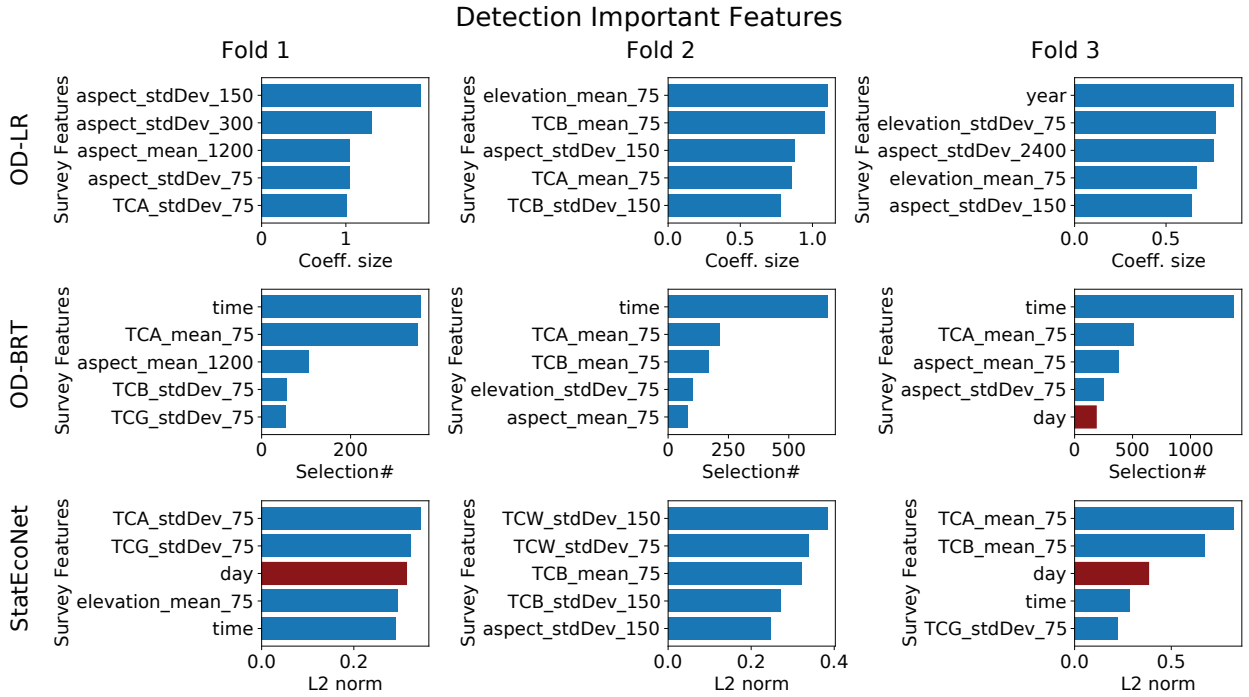


Figure 14: Detection feature importances for Western Meadowlark. The top five features per method per fold are plotted. Note that the x-axes differ across methods. The feature corresponding to the day (chosen as an example feature that is important for `StatEcoNet`) is shaded red to highlight differences across methods. `OD-1NN` is not included here because the importance of environmental features to the detection model is not available from that method.

Model	Hyper-parameter	Optimal Values		
		Fold 1	Fold 2	Fold 3
`OD-LR`	*learningRate*	0.01	0.01	0.01
`OD-1NN`	*learningRate*	0.001	0.001	0.001
	batchSize	*all*	32	32
	nNeurons	16	64	64
`StatEcoNet`	*learningRate*	0.001	0.001	0.001
	batchSize	32	32	32
	nNeurons	32	64	32
	nLayers	3	1	1
	λ	0.01	0.01	0.01
`OD-BRT`	*shrinkage*	0.2121	0.7199	0.4600
	bagFraction	0.8853	0.7763	0.2401
	treeDepth	2	6	10

Table 12: Optimal parameters per fold for Western Meadowlark

7.5 Pacific Wren (PAWR)

This is the example from the main text, repeated here for completeness. In Fig. 15, the `OD-BRT` plots show that many of the model probabilities 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 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 ties the detected/undetected events 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. This is more likely to be the case in practice.

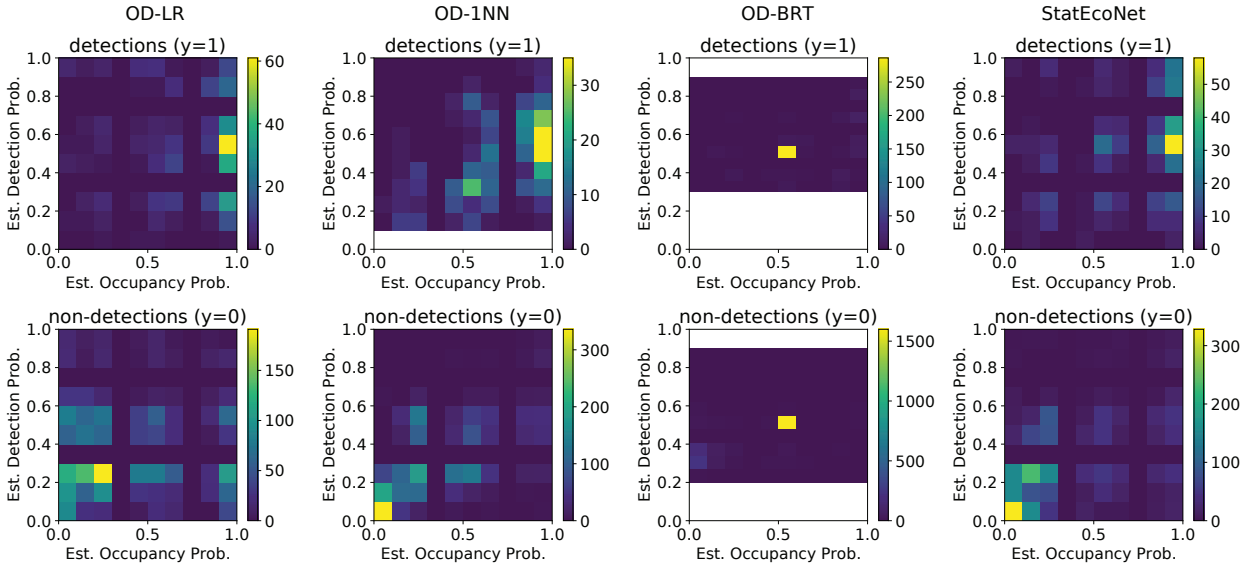


Figure 15: Histograms for Pacific Wren.

Figures 16 and 17 show the top five most important variables learned by each method for PAWR. Inhabiting moist forests, often near riparian zones, Pacific Wrens occupy north-facing slopes that retain moisture later into the dry Pacific Northwest summers. The inclusion of TCA, which captures greenness and brightness, and TCW, capturing correlates of moisture, fits well. The occurrence of aspect also suggests non-random selection of locations in mountainous landscapes by Pacific Wrens.

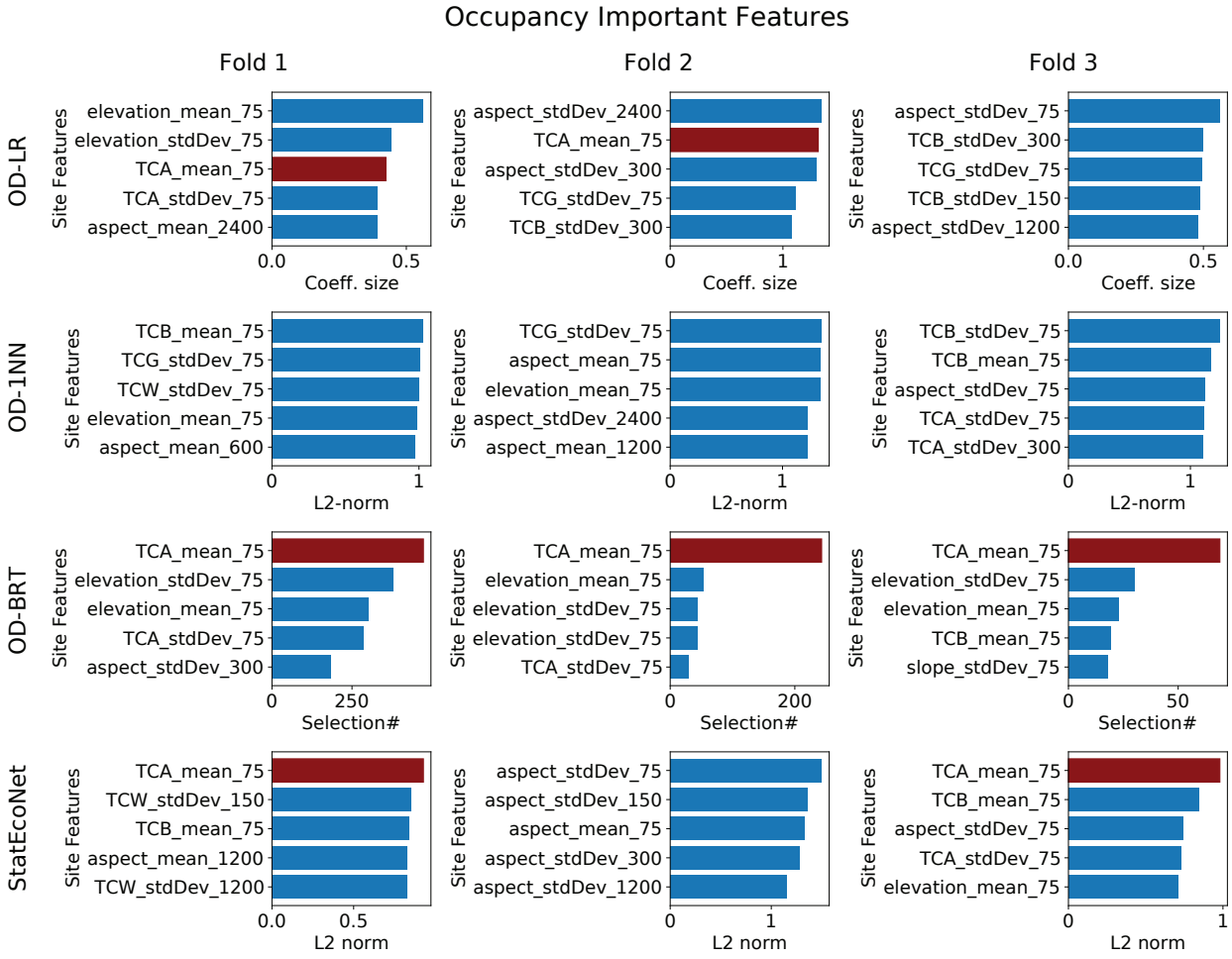


Figure 16: Occupancy feature importances for Pacific Wren. The top five features per method per fold are plotted. Note that the x-axes differ across methods. The feature corresponding to the mean TCA at the 75 m scale (chosen as an example feature that is important for `StatEcoNet`) is shaded red to highlight differences across methods.

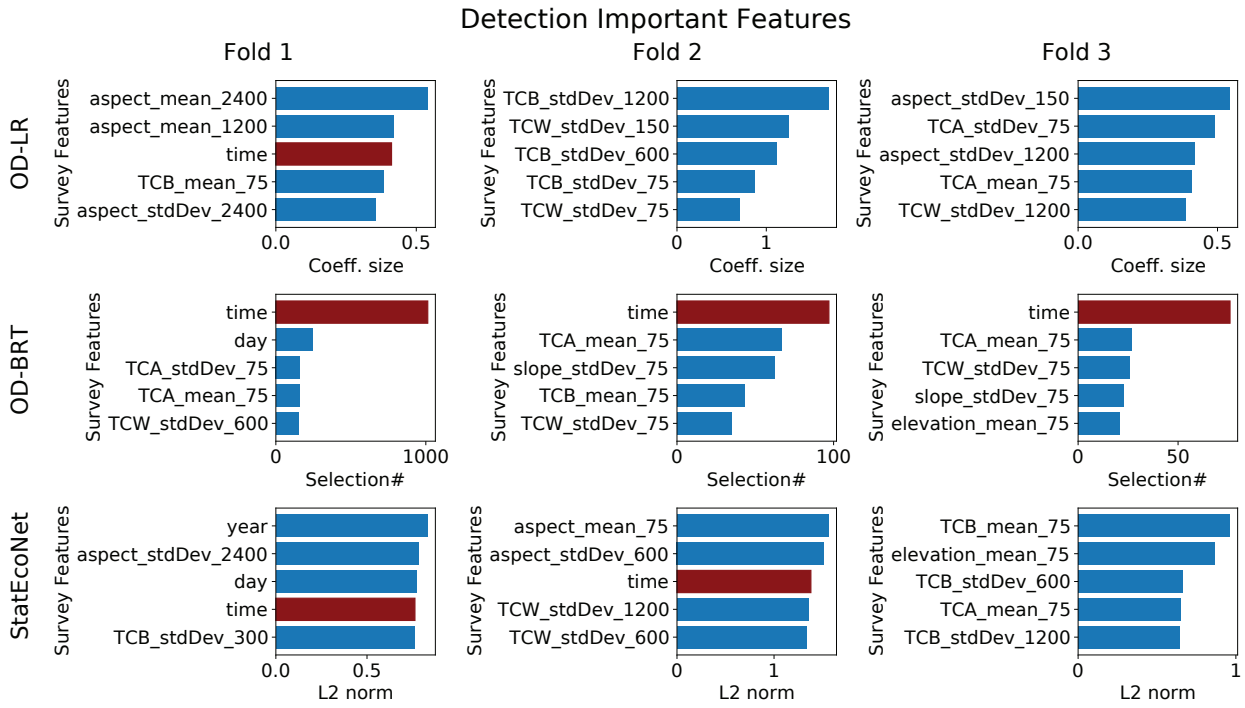


Figure 17: Detection feature importances for Pacific Wren. The top five features per method per fold are plotted. Note that the x-axes differ across methods. The feature corresponding to the time (chosen as an example feature that is important for `StatEcoNet`) is shaded red to highlight differences across methods. `OD-1NN` is not included here because the importance of environmental features to the detection model is not available from that method.

Model	Hyper-parameter	Optimal Values		
		Fold 1	Fold 2	Fold 3
`OD-LR`	*learningRate*	0.01	0.01	0.01
`OD-1NN`	*learningRate*	0.001	0.001	0.001
	batchSize	*all*	*all*	*all*
	nNeurons	64	64	32
`StatEcoNet`	*learningRate*	0.001	0.001	0.001
	batchSize	32	*all*	32
	nNeurons	64	16	16
	nLayers	3	1	1
	λ	0.01	0	0.001
`OD-BRT`	*shrinkage*	0.100	0.100	0.4628
	bagFraction	1.0000	0.4268	0.3946
	treeDepth	4	2	3

Table 13: Optimal parameters per fold for Pacific Wren

8 Computing Infrastructure

Hardware	CPU	# of Cores	4
		# of Threads	8
		Model	Intel(R) Xeon(R) CPU E3-1230 v5 @ 3.40GHz
	Memory	16 GB	
	Operating System	CentOS Linux 7	
Software	Python		3.8.3
	Python libraries	torch	1.5.1
		numpy	1.19.1
		pandas	1.0.5
		matplotlib	3.3.0
		tqdm	4.48.0
		scikit-learn	0.23.1
		scipy	1.5.2
		jupyterlab	2.2.1
		import-ipynb	0.1.3
	R		4.0.2
	R libraries	grt	0.2.1
		reshape2	1.4.4
		PRROC	1.3.1
		Metrics	0.1.4
		paramtest	0.1.0
		Rcpp	1.0.5
		scales	1.1.1
		dplyr	1.0.1
		ggplot2	3.3.2
		patchwork	1.0.1
		blockCV	2.1.1
		raster	3.3-13
		sf	0.9-5

Table 14: Computing infrastructure specification.

References

Flood, N. 2013. Seasonal composite Landsat TM/ETM+ images using the medoid (a multi-dimensional median). *Remote Sensing* 5(12): 6481–6500.

Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; and Moore, R. 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment* 202: 18–27.

Kennedy, R. E.; Yang, Z.; Braaten, J.; Copass, C.; Antonova, N.; Jordan, C.; and Nelson, P. 2015. Attribution of disturbance change agent from Landsat time-series in support of habitat monitoring in the Puget Sound region, USA. *Remote Sensing of Environment* 166: 271–285.

Key, C. H.; and Benson, N. C. 1999. The Normalized Burn Ratio (NBR): A Landsat TM radiometric measure of burn severity. *United States Geological Survey, Northern Rocky Mountain Science Center.(Bozeman, MT)* .

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

Roy, D. P.; Kovalskyy, V.; Zhang, H.; Vermote, E. F.; Yan, L.; Kumar, S.; and Egorov, A. 2016. Characterization of Landsat-7 to Landsat-8 reflective wavelength and normalized difference vegetation index continuity. *Remote sensing of Environment* 185: 57–70.

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.

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

Zhu, Z.; Wang, S.; and Woodcock, C. E. 2015. Improvement and expansion of the Fmask algorithm: Cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and Sentinel 2 images. *Remote Sensing of Environment* 159: 269–277.

Zhu, Z.; and Woodcock, C. E. 2012. Object-based cloud and cloud shadow detection in Landsat imagery. *Remote sensing of environment* 118: 83–94.