Incorporating Boosted Regression Trees into Ecological Latent Variable Models

Rebecca A. Hutchinson, Li-Ping Liu, Thomas G. Dietterich

School of EECS, Oregon State University

Motivation

- Species Distribution Modeling (SDM)
  - SDMs characterize the geographic distribution of a species in terms of a set of environmental variables.
  - Supervised classification problem from features $\mathbf{X}$ to species observations $\mathbf{y}$.

$$\{(\boldsymbol{X}_i, y_i)\}_{i=\{1,\ldots,N\}} \rightarrow y = f(\boldsymbol{X})$$

- Goals
  - Mapping current distribution
  - Understanding habitat requirements
  - Predicting distribution
- SDMs can be used as input to reserve design algorithms

Outline

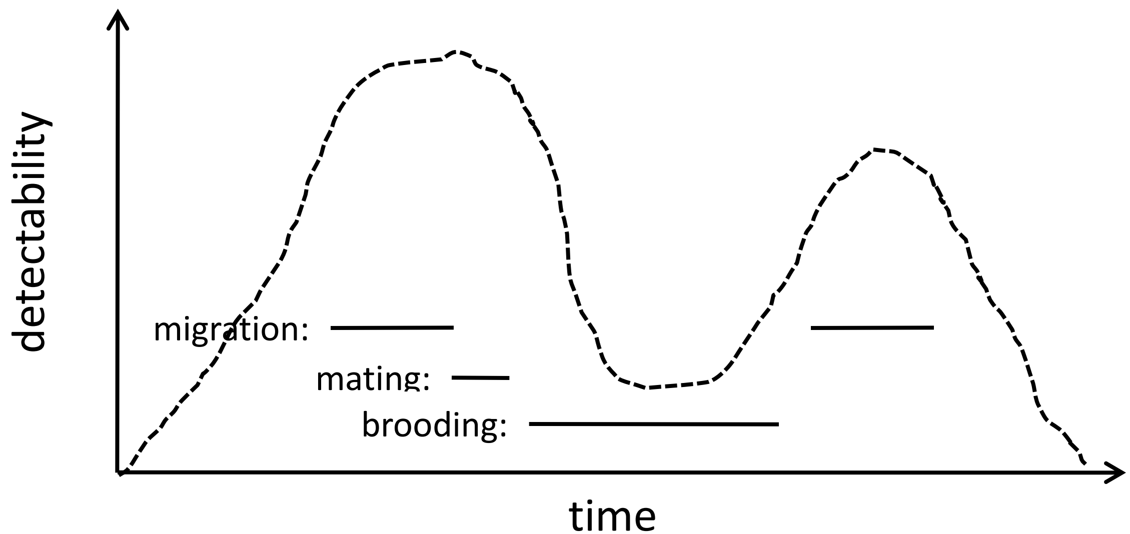
- Background on 2 challenges for SDM:
  - Imperfect detection
    - Existing solution: hierarchical probabilistic approach called site-occupancy or occupancy-detection models [MacKenzie et al 2006]
  - Complexity of ecological systems
    - Existing solution: boosted regression trees have been successful in SDM [Elith et al 2006]

- Our work: A hybrid approach
  - Site-occupancy models fit with an ensemble of regression trees via functional gradient descent [Friedman 2001]

- Experimental results on eBird data

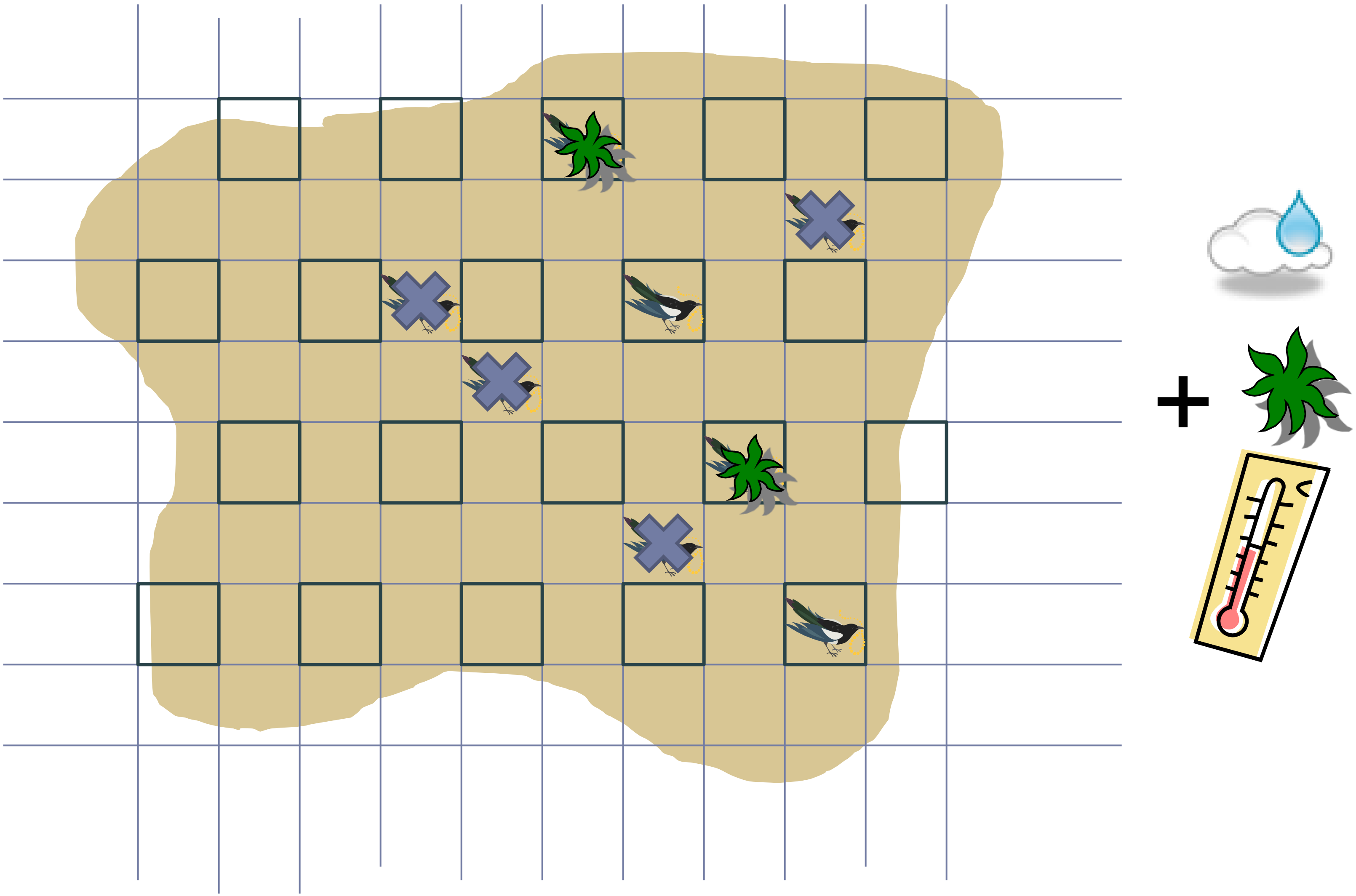
Challenge #1: Imperfect Detection

- Problem: many species are hard to detect even when present, so their data contain false negatives

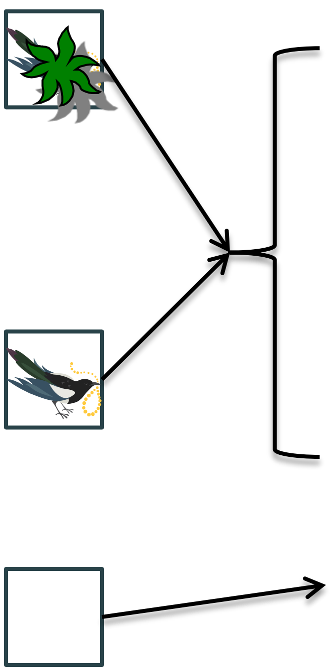


- Solution:
  - Survey sites multiple times
  - Use a hierarchical model to describe the data collection process explicitly and correct for false zeros

Data Collection: Toy Example



Data: Detection Histories



		Detection History		
Site	*True occupancy (latent)*	Visit 1 (rainy day, 12pm)	Visit 2 (clear day, 6am)	Visit 3 (clear day, 9am)
A (forest, elev=400m)	*1*	0	1	1
B (forest, elev=300m)	*1*	0	0	0
C (grassland, elev=200m)	*0*	0	0	0

* Assumption: experts never report false positives.

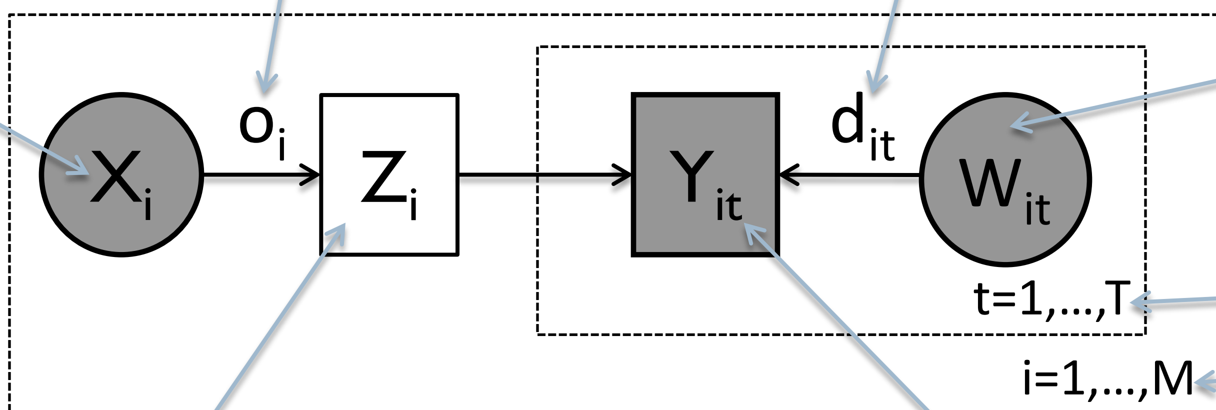
Occupancy-Detection Models

Covariates of occupancy (e.g. elevation, vegetation)

Probability of occupancy (function of X_i, α)

Probability of detection (function of W_{it}, β)

Covariates of detection (e.g. time of day, effort)



Visits

Sites

True (latent) presence/absence
$Z_i \sim \text{Bern}(o_i)$

Observed presence/absence
$Y_{it} \mid Z_i \sim \text{Bern}(Z_i d_{it})$

MacKenzie, et al, 2006

Typical Usage

$$\text{logit}(o_i) = F(X_i) = \alpha \cdot X_i$$

$$\text{logit}(d_{it}) = G(W_{it}) = \beta \cdot W_{it}$$

- Gradient search methods can be applied to find the maximum likelihood values of α and β
- Model selection:
  - construct models including different sets of occupancy and detection covariates
  - evaluate fit with AIC
  - hypothesis tests/confidence intervals

Challenge #2: Complexity!

- We may lack prior knowledge of the complex relationships underpinning ecological systems

- Constructing an occupancy-detection model may entail:
  - Dealing with missing inputs
  - Rescaling/centering inputs
  - Linearizing suspected nonlinear relationships (e.g., via log or sqrt transforms)
  - Transforming ordinal variables and nominal variables
  - Selecting interaction terms to include in the model

- Solution: Boosted regression trees have been successful in SDM [Elith et al 2006]
  - But they don't account for imperfect detection

Incorporating Regression Trees into Occupancy-Detection Models

$$\text{logit}(o_i) = F(X_i) = \sum_{j=1}^{J} \rho_j^{(o)} tree_j^{(o)}(X_j)$$

$$\text{logit}(d_{it}) = G(W_{it}) = \sum_{j=1}^{J} \rho_j^{(d)} tree_j^{(d)}(W_{it})$$

- How to fit this version of occupancy models?

Previous Work

- Friedman (2001): L2-Tree-Boost
  - Fit a logistic regression as a weighted sum of regression trees:
  - $\log \frac{P(y=1|x)}{P(y=0|x)} = \rho_0 + \rho_1 \mathrm{tree}_1(x) + \cdots + \rho_L \mathrm{tree}_L(x)$
  - Fit via functional gradient descent (a form of boosting)
- Dietterich et al. (2004): TreeCRF
  - Fit a Conditional Random Field model using weighted sum of regression trees
- Both cases assume fully-observed outputs (although input features may be missing)

- Can we extend tree boosting to latent variable models?

Fitting Boosted Regression Trees in Occupancy-Detection Models

- $F^{(0)} = G^{(0)} = 0$
- For $j = 1, \ldots, J$
  - For each site i, compute
    $$\tilde{z}_i = \partial \ell_i / \partial F \big|_{F=F^{(j-1)}(x_i)}$$
  - Fit regression tree f_j to $\{\langle x_i, \tilde{z}_i \rangle\}_{i=1}^{M}$
  - Let $F^{(j)} = F^{(j-1)} + \rho_j f_j$

  Occupancy Sub-Model

  - For each visit t to site i, compute
    $$\tilde{y}_{it} = \partial \ell_i / \partial G \big|_{G=G^{(j-1)}(w_{it})}$$
  - Fit regression tree g_j to $\{\langle w_{it}, \tilde{y}_{it} \rangle\}_{i=1,t=1}^{M,T_i}$
  - Let $G^{(j)} = G^{(j-1)} + \nu_j g_j$

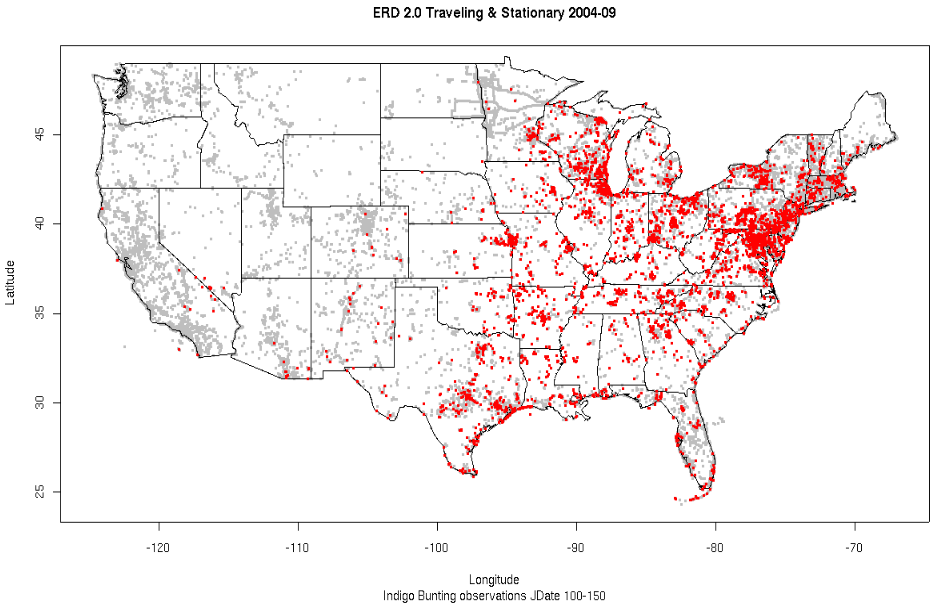
  Detection Sub-Model

Experiment: 4 models

	Supervised (S) $(x_i, w_{it}) \rightarrow y_{it}$	**Occupancy-Detection (OD)** (latent variable models)
Linear (LR)	S-LR Logistic regression	OD-LR F and G as logistic regressions
Tree-based (BRT)	S-BRT Boosted Regression Trees	OD-BRT F and G as regression tree ensembles

eBird Data

- "Citizen Science" Data:
  - 12 bird species
  - 3 synthetic species
  - 3124 observations from New York State, May-July 2006-2008
  - Pre-processing for occupancy models to group records into sites
  - 19 occupancy features, 4 detection features



Synthetic Species

- Synthetic Species 1: F and G linear

$$\text{logit}(o_i) = -2x_i^{(4)} + 2x_i^{(13)}$$
$$\text{logit}(d_{it}) = w_{it}^{(2)} + w_{it}^{(3)} - 1$$

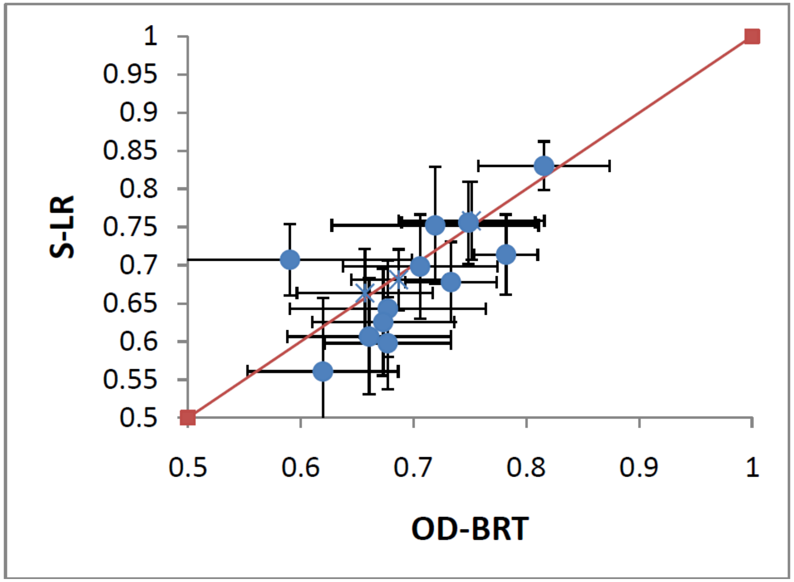
- Synthetic Species 2: F and G nonlinear

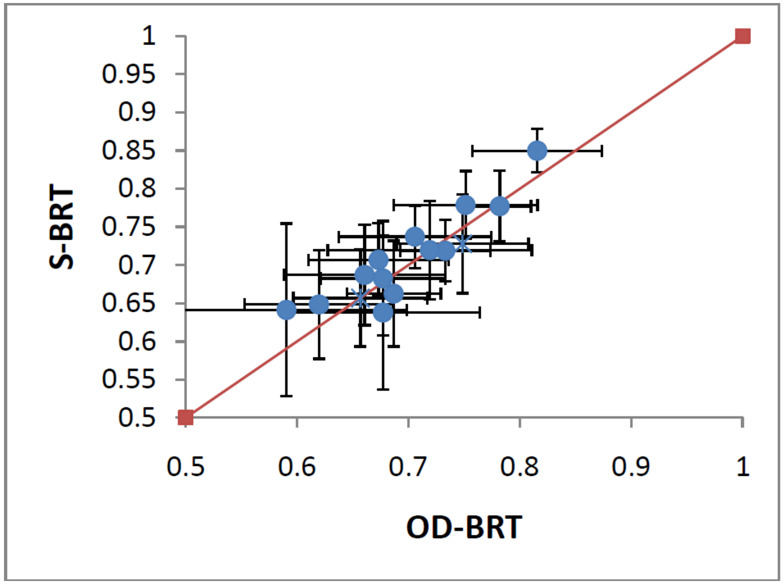
$$\text{logit}(o_i) = -2\left[x_i^{(1)}\right]^2 + 3\left[x_i^{(2)}\right]^2 - 2x_i^{(3)}$$
$$\text{logit}(d_{it}) = \exp(-0.5w_{it}^{(4)}) + \sin(1.25w_{it}^{(1)} + 5)$$

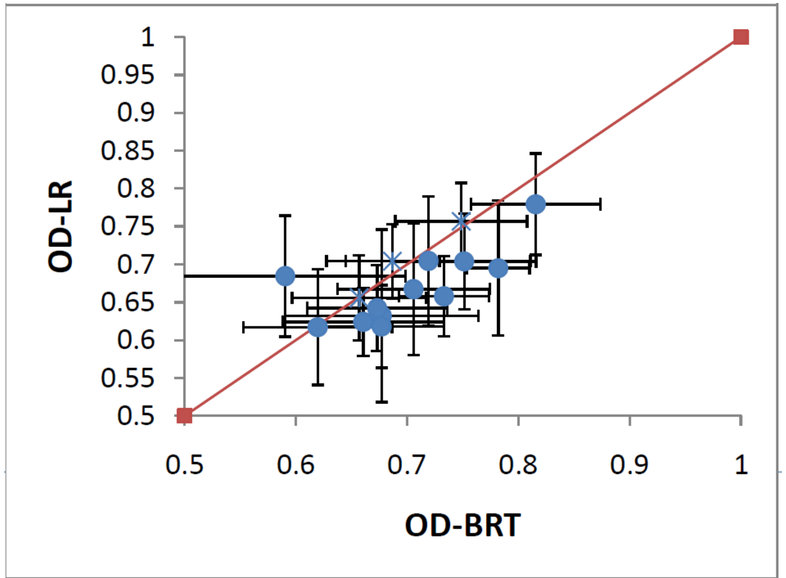
- Synthetic Species 3: F and G nonlinear with interactions

$$\text{logit}(o_i) = -\exp\left(-x_i^{(4)}x_i^{(12)}\right) - 2x_i^{(1)} - 0.5$$
$$\begin{aligned}\text{logit}(d_{it})\\ &= \exp(-0.5w_{it}^{(4)}) \cdot \sin\left(1.25w_{it}^{(1)} + 5\right) + \exp\left(-0.5w_{it}^{(4)}\right)\\ &+ \sin\left(1.25w_{it}^{(1)} + 5\right)\end{aligned}$$

Results for AUC of y_{it}: No Significant Differences

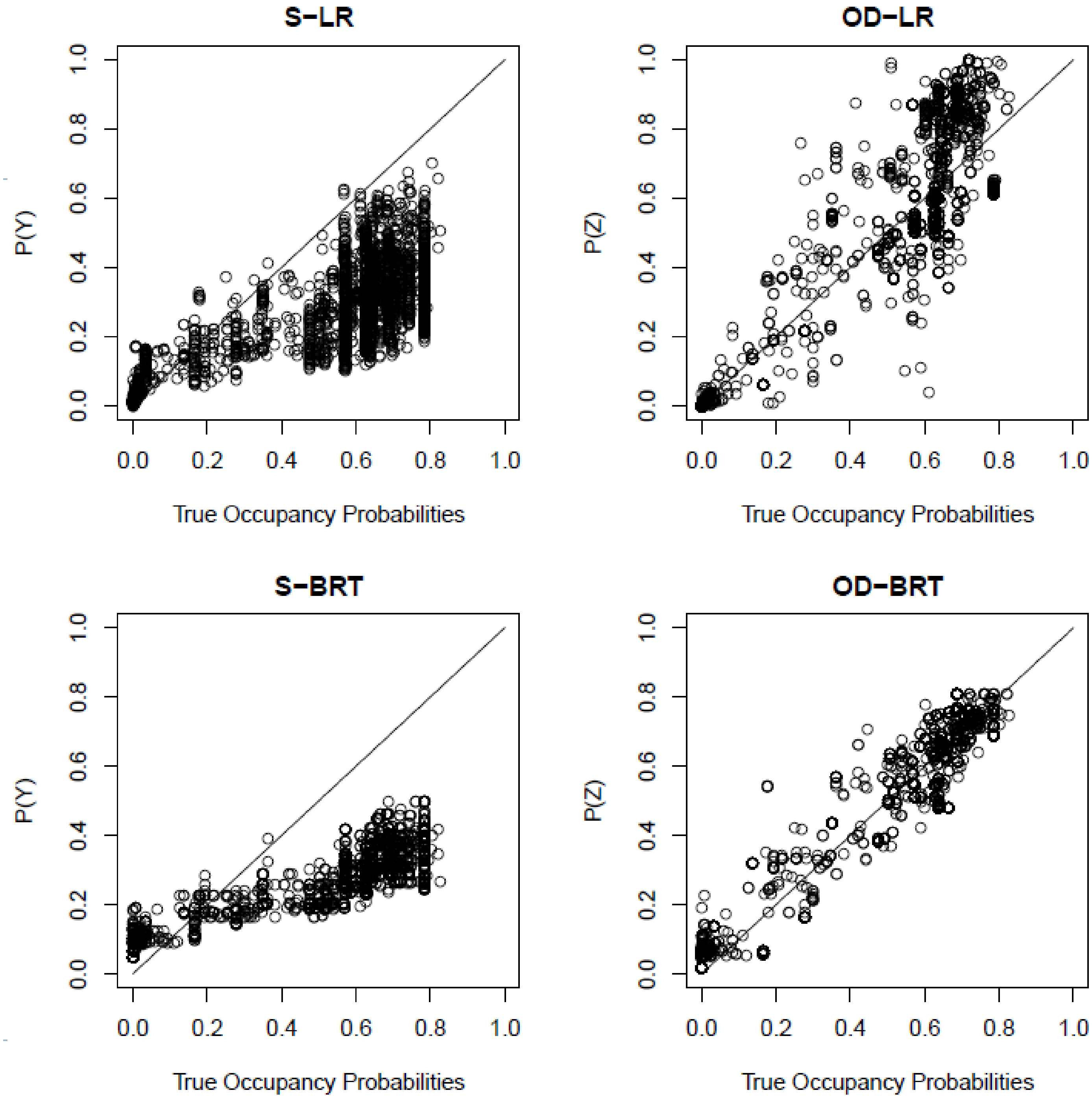






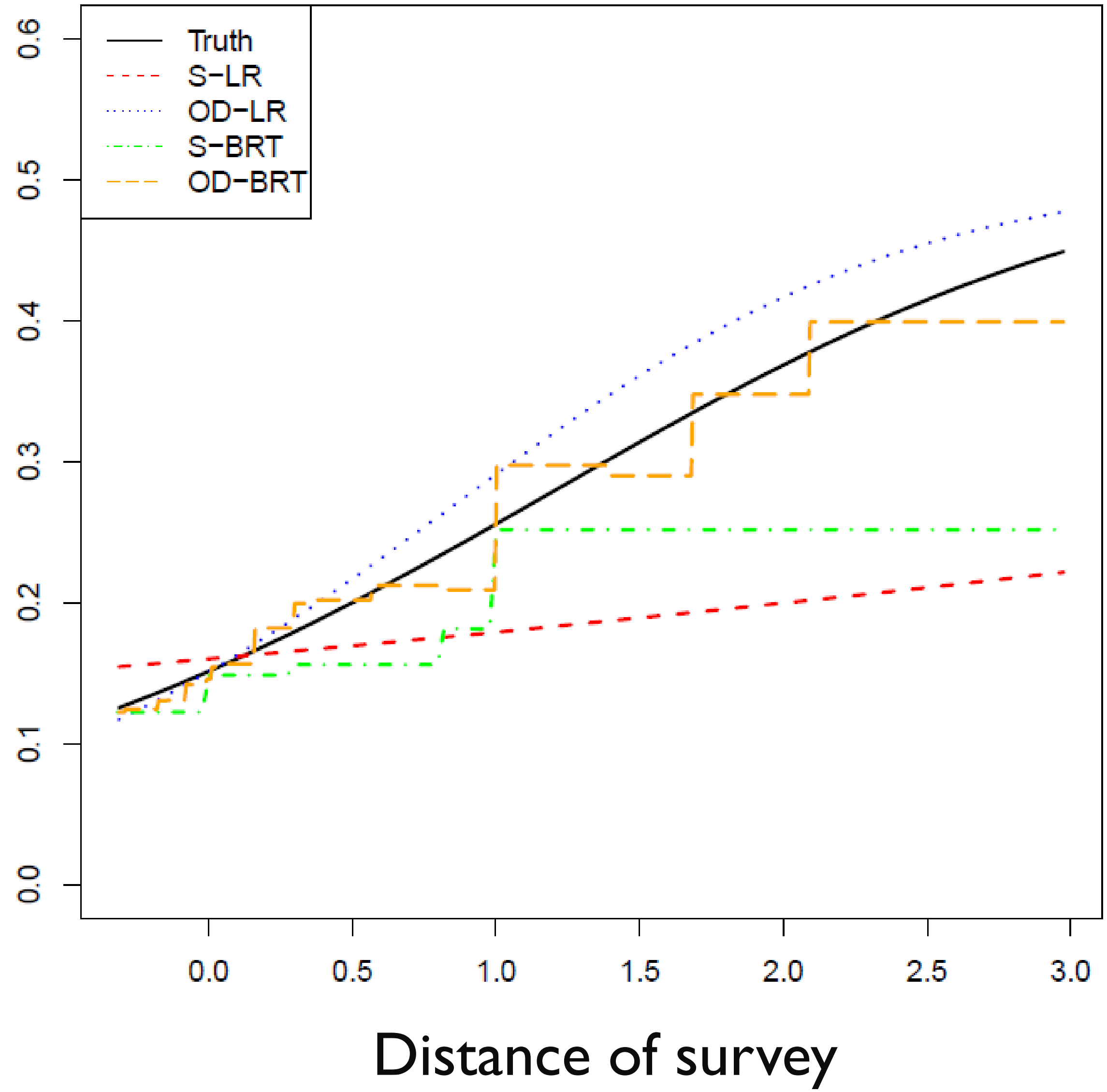
Predicting Occupancy

Synthetic Species 2



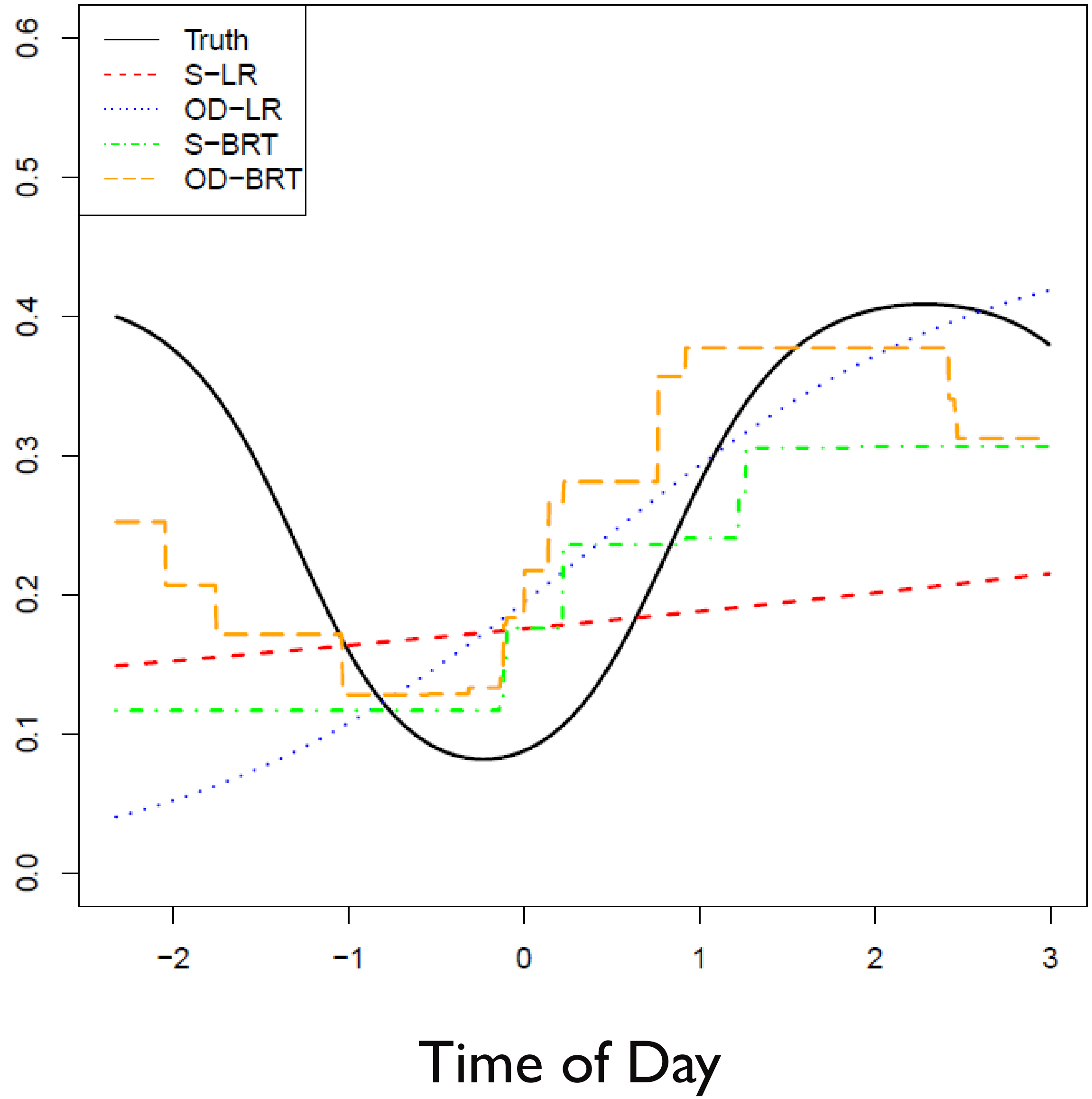
Partial Dependence Plot
Synthetic Species 1

- OD-BRT has the least bias

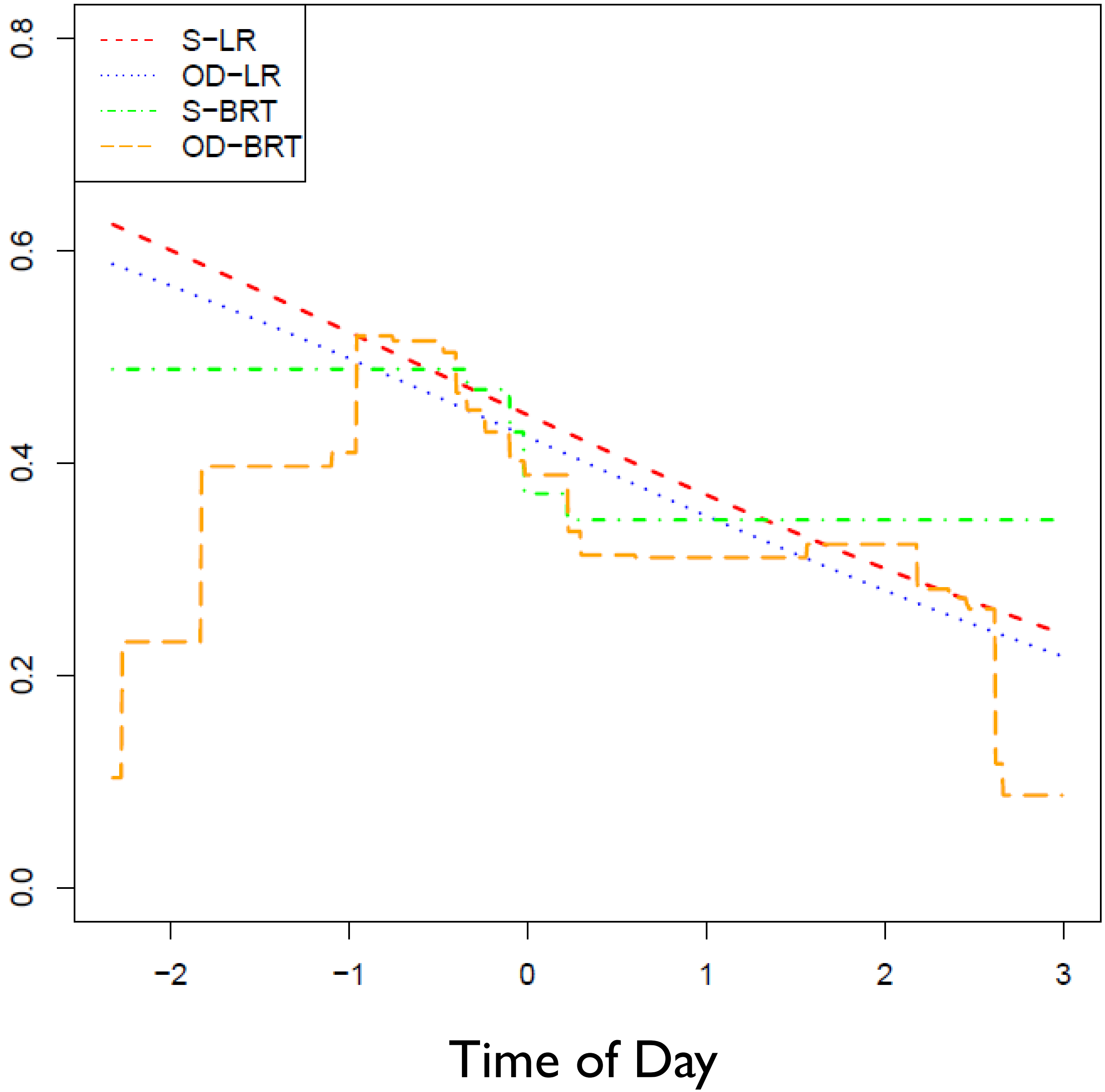


Partial Dependence Plot
Synthetic Species 3

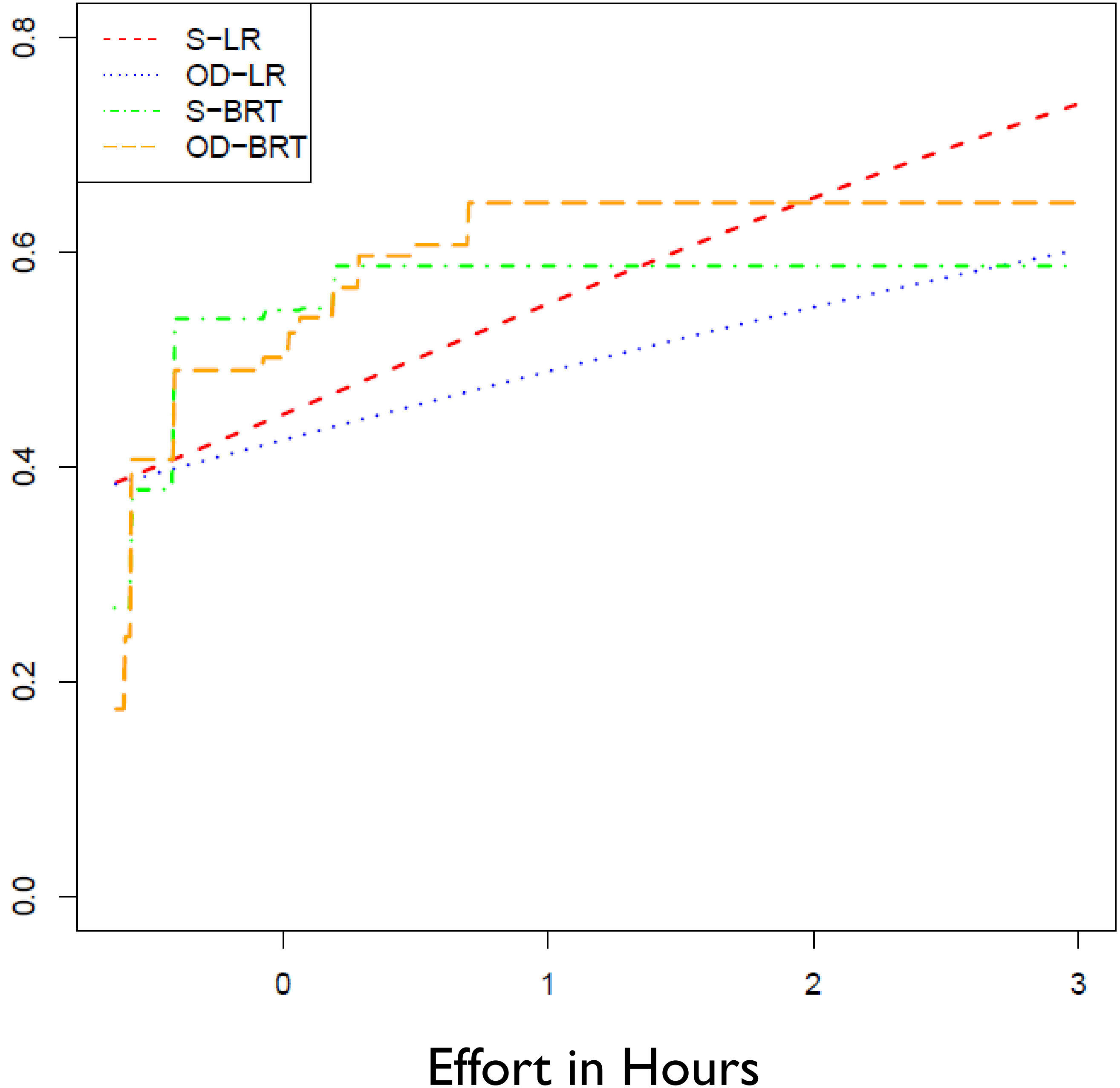
- OD-BRT has the least bias and correctly captures the bimodality



Partial Dependence Plot Blue Jay vs. Time of Day



Partial Dependence Plot Blue Jay vs. Duration of Observation



Conclusions and Contributions

- We have succeeded in incorporating BRTs into the occupancy-detection models
  - Accurate predictions of both the observations and the latent variables
  - The fitted trees are correctly capturing nonlinearities

- Machine learning: case study for doing functional gradient descent in latent variable models
- Ecology: allows two major modeling challenges to be addressed simultaneously

Next Steps

- Preparing an R package
- Collaborating with ecologists to apply OD-BRT to more datasets
- Experiments to validate interaction discovery
- Extending the method to models with more complex latent structure

References and Acknowledgements

Elith J, Graham CH, Anderson RP, et al. **Novel methods improve prediction of species' distributions from occurrence data.** *Ecography*. 2006;29(2):129-151.

Dietterich, TG, Ashenfelter, A, and Bulatov, Y. **Training Conditional Random Fields via Gradient Tree Boosting**. *International Conference on Machine Learning*, 2004.

Friedman J. **Greedy function approximation: a gradient boosting machine.** *Annals of Statistics*. 2001;29(5):1189-1232.

MacKenzie DI, Nichols JD, Royle JA, et al. ***Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence.*** Elsevier, San Diego, USA; 2006.

This research was funded by the National Science Foundation under grant number NSF-0832804.

Thanks!

- Questions?