Incorporating Boosted Regression Trees into Ecological Latent Variable Models

Rebecca A. Hutchinson and **Li-Ping Liu** and **Thomas G. Dietterich**
School of Electrical Engineering and Computer Science
Oregon State University
Corvallis, OR 97331-5501
{rah,liuli,tgd}@eecs.oregonstate.edu

Abstract

Important ecological phenomena are often observed indirectly. Consequently, probabilistic latent variable models provide an important tool, because they can include explicit models of the ecological phenomenon of interest and the process by which it is observed. However, existing latent variable methods rely on hand-formulated parametric models, which are expensive to design and require extensive preprocessing of the data. Nonparametric methods (such as regression trees) automate these decisions and produce highly accurate models. However, existing tree methods learn direct mappings from inputs to outputs—they cannot be applied to latent variable models.

This paper describes a methodology for integrating nonparametric tree methods into probabilistic latent variable models by extending functional gradient boosting. The approach is presented in the context of occupancy-detection (OD) modeling, where the goal is to model the distribution of a species from imperfect detections. Experiments on 12 real and 3 synthetic bird species compare standard and tree-boosted OD models (latent variable models) with standard and tree-boosted logistic regression models (without latent structure). All methods perform similarly when predicting the observed variables, but the OD models learn better representations of the latent process. Most importantly, tree-boosted OD models learn the best latent representations when nonlinearities and interactions are present.

For many problems in ecology and ecosystem management, the phenomena of interest are not directly observed. Examples range from the basic spatial distribution of species to more complex phenomena such as dispersal, migration, and species interactions (mating, predation, etc.). Instead of direct observations, we often can obtain only indirect information such as animal sightings, abandoned nests, animal droppings, and so on. A fundamental challenge for data-driven modeling in ecology is to construct models of the phenemona of interest from such indirect information.

In this paper, we consider a particular instance of this problem: modeling the habitat requirements of a species. A habitat model is a function $f : X \mapsto Y$, where $x \in X$ describes the habitat at a site and $y \in \{0, 1\}$ indicates whether the habitat is suitable or unsuitable for the species. To construct such a model using machine learning methods, one would like to visit a variety of sites and measure whether the species is present or absent at those sites. The resulting data could be applied to train a habitat model. Unfortunately, many species are difficult to detect (e.g., because they actively hide from people, they are camouflaged, or they roam over large ranges). So an observer may report that the species is absent at a site when in fact it is present. Supervised learning applied to such observations will try to fit these "false zeroes" and seriously underestimate the distribution of the species (MacKenzie et al. 2002), which can lead to errors in scientific understanding and in the design of conservation strategies.

A general approach to solving such problems is to formulate a probabilistic latent variable model in which the true presence or absence of the species is represented by a latent variable z, and the observations y are produced by a stochastic *observation process* in which, when the species is present ($z = 1$), $y = 1$ according to a *detection probability* and $y = 0$ otherwise. This model is known in ecology as the "occupancy model", but we will refer to it by the more accurate name of "occupancy-detection model" (abbreviated OD).

Recently, the OD model has begun to be applied in ecology and wildlife studies (MacKenzie et al. 2002; 2006). However, the OD model exhibits several drawbacks common to all parametric probabilistic models. The modeler must carefully design the model so that it includes relevant environmental features. If there are interactions or nonlinearities, then terms must be included in the model to capture these. If some features are missing, then their values must be imputed or the relevant records must be ignored. Finally, the data must be transformed and standardized to match the model assumptions (e.g., linearity, gaussianity, etc.). In cases where the system is already well-understood and the goal is hypothesis testing, this is acceptable. But these drawbacks make parametric probabilistic models unsuitable for exploratory and predictive modeling, where the goal is to *discover* a good model and apply it to make accurate predictions.

One of the most important contributions of machine learning to statistical modeling has been the development of robust, easy-to-use, nonparametric modeling methods such as

boosted trees and support vector machines. Classification and regression trees (Breiman et al. 1984), in particular, can be applied to data without preprocessing, because they are invariant to rescaling and other monotonic data transformations. They can handle missing values in the input features, and they automatically capture nonlinearities and feature interactions. Within ecology, boosted regression trees (e.g., as implemented by the R package `gbm` (Ridgeway 2007)) have been shown to produce extremely accurate species distribution models (Elith et al. 2006) in the fully-observed case.

An exciting direction for machine learning is to find ways to integrate nonparametric methods into probabilistic graphical models, and especially into latent variable models. A first step in this direction was achieved by Friedman (2001), who showed how to incorporate boosted regression trees into generalized linear models such as logistic regression and Poisson regression. A second step was the work of Dietterich, et al. (2008), who showed how to integrate boosted regression trees into structured output models such as conditional random fields.

In this paper, we show how to integrate boosted regression trees into the OD model. The method is general and can be applied to any probabilistic graphical model for which the necessary functional gradients can be computed. It is most appropriate in cases where it is desired to condition parts of the model on potentially large sets of input features. Combining nonparametric methods with graphical models allows us to obtain the best aspects of both. The modeler can specify the qualitative structure of the graphical model (e.g., to introduce appropriate latent variables) and then the conditional probability distributions in the model can be fit using flexible boosted regression trees.

The OD Model

Figure 1 shows a plate diagram of the OD model. The outer plate represents M sites (indexed by $i = 1, \ldots, M$). The variable x_i denotes a vector of occupancy features, and $z_i \in \{0, 1\}$ denotes the true occupancy of site i. Site i is visited T_i times (indexed by $t = 1, \ldots, T_i$). The variable w_{it} is a vector of detection features, and $y_{it} \in \{0, 1\}$ indicates whether the species was detected ($y_{it} = 1$) on visit t.

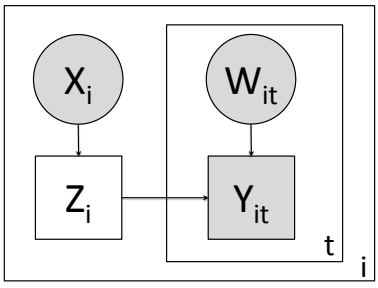

Figure 1: Plate diagram of the Occupancy-Detection model.

In the standard parametric form, the OD model has the following structure. To generate the observations for site i, a logistic regression formula $\text{logit}(o_i) = F(x_i; \alpha)$ (the occupancy or habitat model) is first evaluated to compute the probability o_i that site i is occupied by the species. Then the true occupancy z_i is generated by drawing from a Bernoulli random variable with parameter o_i. Next, the site is visited T_i times. At each visit t, a second logistic regression formula $\text{logit}(d_{it}) = G(w_{it}; \beta)$ is evaluated to compute the detection probability d_{it}. Finally, the observation y_{it} is generated by drawing a Bernoulli random variable with parameter $z_i d_{it}$. The effect is that if $z_i = 0$, then $y_{it} = 0$, but if $z_i = 1$, then $y_{it} = 1$ with probability d_{it} and 0 otherwise. In this formulation, F and G are linear models with coefficient vectors α and β, respectively.

From this description, we can see that the OD model makes two assumptions:

- Survey sites are visited multiple times over a period of *population closure* during which the occupancy status (z_i) of a site does not change.
- Observers are trained sufficiently well that it is reasonable to assume there are no false positives in the data. There may be false negatives, but the observers never mistake one species for another or otherwise falsely report the presence of the species.

The log likelihood function for the OD model can be written as follows ($\mathbb{1}(\cdot)$ is 1 if its argument is true and 0 otherwise, and $\theta = \{\alpha, \beta\}$):

$$\ell(\mathbf{Y}|\mathbf{x}, \mathbf{w}; \theta) = \sum_{i=1}^{M} \log \ell_i(y_{i\cdot}|x_i, \mathbf{w}_{i\cdot}; \theta) = \sum_{i=1}^{M} \log \ell_i, \quad (1)$$

$$\begin{aligned} \ell_i &= \sum_{z \in \{0,1\}} P(z_i = z) P(y_{i\cdot}|z_i = z) \\ &= \sum_{z \in \{0,1\}} o_i^z (1 - o_i)^{1-z} \prod_{t=1}^{T_i} (z d_{it})^{y_{it}} (1 - z d_{it})^{1-y_{it}} \\ &= o_i \prod_{t=1}^{T_i} \left[d_{it}^{y_{it}} (1 - d_{it})^{1-y_{it}} \right] + \\ &\quad (1 - o_i) \mathbb{1} \left(\sum_{t=1}^{T} y_{it} = 0 \right). \end{aligned} \quad (2)$$

We will refer to this standard approach as OD-LR, because it employs log-linear models for F and G. To fit OD-LR to data, the parameters of the model, α and β, are adjusted via gradient ascent to maximize Eq. 1.

Integrating Boosted Regression Trees into the OD Model

To integrate boosted regression trees into the OD model, we can replace the functions F and G that determine the occupancy and detection probabilities with weighted sums of regression trees trained via coordinate functional gradient descent. We will call this the OD-BRT model. The functions F and G are initialized to be $F^{(0)} = G^{(0)} = 0$. Then each gradient descent step $j = 1, \ldots, J$ consists of the following:

1. For each site i, compute the partial derivative of the log likelihood wrt F evaluated at its current value $F^{(j-1)}(x_i)$. The computed derivative, $\tilde{z}_i$, is called the *pseudo-target* for F at site i.
2. Fit a regression tree f_j to the training examples $\{\langle x_i, \tilde{z}_i \rangle\}$. Choose a step size ρ_j. (We employed constant step sizes, chosen via a holdout set.) Let $F^{(j)} = F^{(j-1)} + \rho_j f_j$.
3. For each visit t to each site i, compute the partial derivative of the log likelihood wrt G evaluated at its current value $G^{(j-1)}(w_{it})$. The computed derivative, $\tilde{y}_{it}$ is called the *pseudo-target* for G at (i, t).
4. Fit a regression tree g_j to the training examples $\{\langle w_{it}, \tilde{y}_{it} \rangle\}$. Choose a step size ν_j. Let $G^{(j)} = G^{(j-1)} + \nu_j g_j$.

While we have written this algorithm in terms of a specific ecological example, we note that the functional gradient descent algorithm could be applied to a variety of other problems in which prior knowledge indicates a particular probabilistic structure but the model probabilities are unknown functions of a potentially large set of input variables. In this case, we used a standard application of functional gradient ascent, since the latent variables can be marginalized out easily. For models with more complex latent structure, it may be necessary to merge this algorithm with Expectation-Maximization (Dempster, Laird, and Rubin 1977).

Experimental Design and Data

To test the effectiveness of the OD-LR and OD-BRT models, we designed an experiment to compare them against the standard methods of logistic regression (which we will refer to as S-LR) and boosted regression trees (which we will denote by S-BRT). The S in S-LR and S-BRT stands for "supervised", because these algorithms formulate the task as a standard supervised learning problem. This four-way comparison allows us to measure the effect of introducing boosted trees to both the supervised and OD models.

We applied these four models to a portion of the eBird Reference Data Set (Munson et al. 2010). eBird is a citizen science program (`www.eBird.org`) run by the Cornell Lab of Ornithology and the National Audobon Society in which bird-watchers report their observations to a central database. Our data consists of 3124 detection/non-detection observations of 12 species from the breeding seasons (May-July) of 2006-2008 in the state of New York. Table 1 lists the features accompanying these observations, which include variables describing both the surrounding habitat and the observation conditions for each report. In order to apply S-LR and OD-LR, all of the features were standardized to $\mathcal{N}(0, 1)$; this is unnecessary for the tree methods.

A drawback of using real data is that we do not have ground truth for the latent occupancy variables z_i nor do we know the true structure of the occupancy and detection models. To address this, we designed three synthetic species (Species 13-15) and generated latent variables and observations for each of them (using the same features as for the real species). The functions were chosen arbitrarily to showcase a variety of relationships. For Species 13, the occupancy and detection functions are linear, and hence match the assumptions of OD-LR:

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

Species 14 introduces nonlinear occupancy and detection components, which should favor OD-BRT:

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

Species 15 introduces interactions between nonlinear components, which should further favor OD-BRT:

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

$X^{(1)}$	Human population per sq. mile
$X^{(2)}$	Number of housing units per sq. mile
$X^{(3)}$	Percentage of housing units vacant
$X^{(4)}$	Elevation
$X^{(5)} \dots X^{(19)}$	Percent of surrounding 22,500 hectares in each of 15 habitat classes from the National Land Cover Dataset (NLCD)[1]
$W^{(1)}$	Time of day
$W^{(2)}$	Observation duration
$W^{(3)}$	Distance traveled during observation
$W^{(4)}$	Day of year

Table 1: Input variables for the species distribution models.

The S-LR and S-BRT algorithms treat each observation as iid, so they can be applied directly to the eBird observations. The OD-LR and OD-BRT methods take into account the non-iid nature of the data—namely that multiple visits to the same site i are influenced by a single value of z_i. To reveal the site structure of the data, we aggregated observations within 0.16 kilometers of each other into sites, condensing 3124 observations to 314 sites. The number of visits per site ranged from 1 to 81. Observations from the same location in different years were treated as different "sites" so that the population closure assumption only applied within a year and not across years. While OD models are more commonly applied to data collected specifically to match their assumptions (repeated visits, population closure, etc.), they have also been applied to citizen science data using similar preprocessing (Kéry, Gardner, and Monnerat 2010).

Each of the four models has tuning parameters to set. The S-LR parameters were regularized with an L2 penalty with

[1]NLCD classes are: open water, developed open space, developed low intensity, developed medium intensity, developed high intensity, barren land, deciduous forest, evergreen forest, mixed forest, shrub/scrub, grassland/herbaceous, pasture/hay, cultivated crops, woody wetlands, emergent herbaceous wetlands

the regularization parameter $\lambda \in \{0, 0.001, 0.01, 0.1, 1\}$. The S-BRT models were fit using the `gbm` package in R (Ridgeway 2007). The tuning parameters for S-BRT were the number of trees ($nTrees \in \{100, 200, 400, 800, 1600\}$), the step size ($shrinkage \in \{0.001, 0.01, 0.1\}$), and the depth of the trees ($treeDepth \in \{1, 2, 5, 10\}$). In addition to tuning these parameters, we set other `gbm` parameters to match the OD-BRT implementation. The log-linear components of OD-LR were also L2-regularized, but the occupancy and detection components had separate penalties with independent regularization parameters: $\lambda_O \in \{0, 0.001, 0.01, 0.1\}$ and $\lambda_D \in \{0, 0.001, 0.01, 0.1\}$. The OD-BRT models were tuned with the same parameters and values as `gbm`, but due to the already large number of tuning parameter combinations, we used the same settings for both the occupancy and detection components. For each combination of model, species, and training set, the tuning parameters were set to the values that produced the best AUC in predicting the observations on an independent validation set.

To divide the eBird data into training and test sets, we placed a 9×16 checkerboard pattern over the state of New York (each grid cell was roughly 50km $\times$ 33km). We then performed two-fold cross-validation with the white and black squares defining the folds. Within the training set, we further subdivided each square into a 2×2 grid and used one diagonal pair as the training set and the other as a validation set for choosing tuning parameters. We repeated this 2-fold cross-validation for ten random placements of the bottom left corner of the checkerboard, for a total of 20 splits of the data. This is the same validation scheme that was used by Yu et al. (2010). For visualization purposes, we also trained one instance of each model using the entire training set (with the tuning parameters set to the mode of the tuning parameters from the cross-validation runs).

Results

When methods without latent occupancy structure (like S-LR and S-BRT) are applied in species distribution modeling, they are typically evaluated based on their ability to predict held-out observations. Figure 2 presents the means and standard deviations (across the 20 train/test splits) of the AUCs for all four methods on all 15 species for predicting the observations Y. These results indicate that while the tree-based methods tend to slightly outperform the linear methods, there is no significant and consistent difference between these four methods for the task of predicting Y. However, the predictions of Y do not answer the ecological question of interest for species distribution models in the presence of imperfect detection. Instead, they only predict what will be observed at a new location rather than whether the location is truly inhabited by the species of interest.

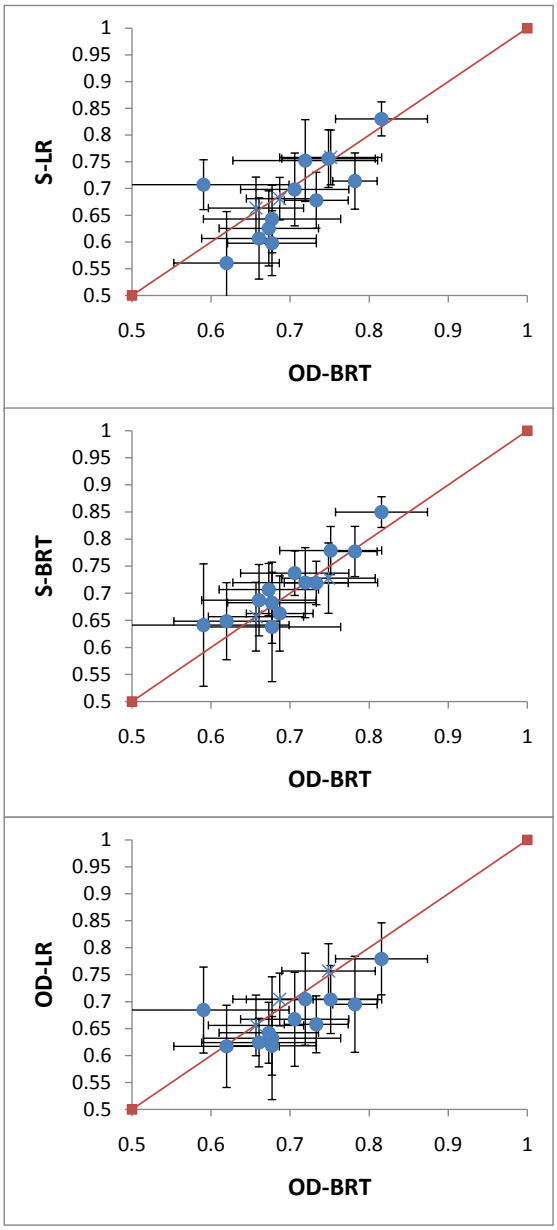

Figure 2: Mean AUCs (error bars are $\pm$ one standard deviation) for predicting held-out observations Y. The mean and standard deviation are taken across 20 train/test splits. Circles represent real species and crosses represent synthetic species. While the tree-based methods tend to slightly outperform the linear methods, there is no significant and consistent difference between these four methods for the task of predicting Y. However, the predictions of Y do not answer the ecological question of interest for species distribution models in the presence of imperfect detection.

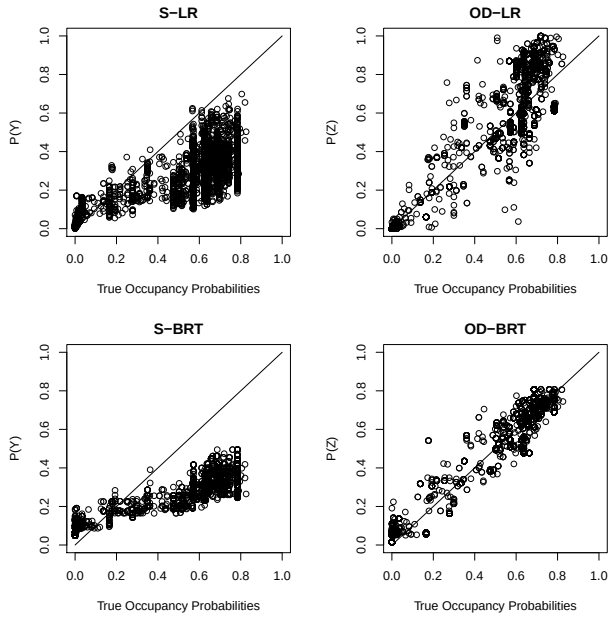

Figure 3: True occupancy probabilities for Species 14 versus model predictions.

Since predictive performance on the observations is not our primary goal, we would like to evaluate the accuracy of the model in predicting the true occurrence pattern of the species. Unfortunately, we almost never have ground truth for the latent variables in occupancy-detection models, but we can address this question with the synthetic species. Figure 3 shows scatter plots of probabilities from the four models against the true occupancy probabilities used to generate the data for Species 14. The OD-LR and OD-BRT plots show the estimated probabilities of occupancy, $P(Z)$, on the y-axis. The S-LR and S-BRT model plots show the estimated probabilities of observation, $P(Y)$ on the y-axis. As expected, the S-LR and S-BRT predictions are biased low, since these models must interpret all zeroes in the data as true absences, whereas the occupancy-detection models can interpret some zeroes as false absences. While the S-LR and S-BRT plots may not seem like a fair comparison, we emphasize that these models cannot make a prediction about $P(Z)$. Nonetheless their predicted probabilities of observations are frequently interpreted as predictions of occupancy when the detection issue is ignored.

We also note that the nonlinear OD-BRT model produces better estimates of $P(Z)$ than the OD-LR model, due to its ability to model the nonlinear relationships underlying the data for Species 14. In this case, the OD-LR model could be tailored to represent these relationships by adding the appropriate quadratic terms to the occupancy and detection functions, if the model designer suspected these relationships in advance. On the other hand, the OD-BRT model discovers and represents the nonlinearities automatically, which is useful for exploratory modeling in which the relevant variables and relationships are unknown.

Partial dependence plots (Friedman 2001) provide a visualization of the effect of a variable on a predicted response after averaging over all of the other variables. Figure 4 shows the effects of $w^{(1)}$ on the functions estimated by the four methods, along with the true, bimodal relationship used in generating the data for Species 15. This bimodality is an example of a nonlinearity that cannot easily be represented by the OD-LR model even if it is known to the model designer, whereas the OD-BRT model again discovers and represents the relationship automatically. Figure 5 shows a similar plot for $w^{(3)}$ in Species 13, a relationship that is truly linear. This plot shows that while the (correct) OD-LR model estimates the linear effect slightly better, the OD-BRT model also does a reasonable job of capturing the effect.

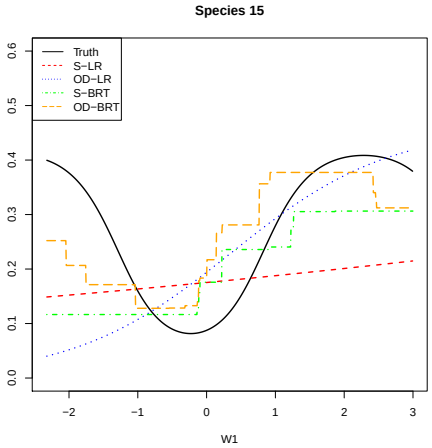

Figure 4: Partial dependence of time of day for Species 15.

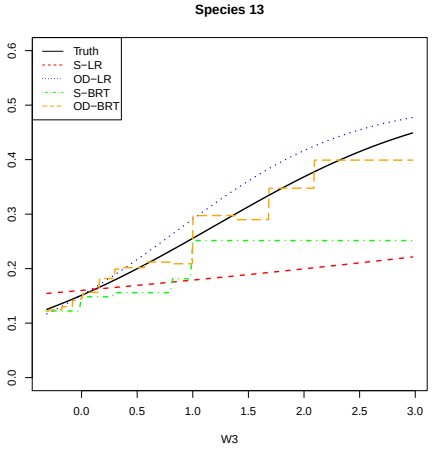

Figure 5: Partial dependence of distance of survey for Species 13.

Note that these partial dependence plots display the relationship between an input variable and the function estimating $P(Y)$, since all four methods can compute this quantity. OD-LR and OD-BRT can also produce partial dependence plots for the occupancy and detection functions.

While we cannot validate partial dependence plots for the

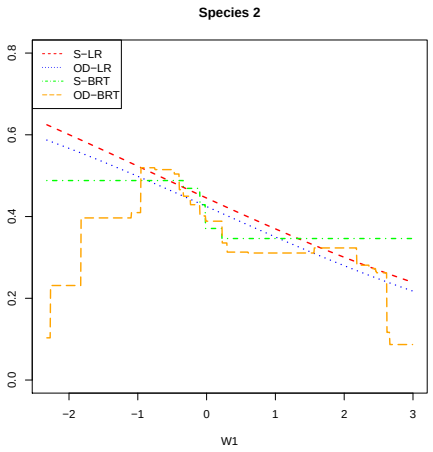

Figure 6: Partial dependence of time of day for Species 2.

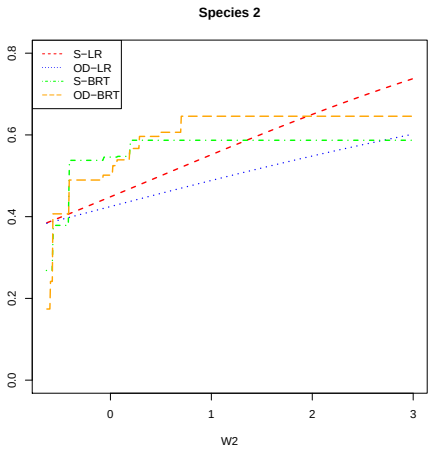

Figure 7: Partial dependence of survey duration for Species 2.

real species against ground truth, they are nonetheless a good tool for visualizing trained models. For example, the partial dependence plots for the OD-BRT model on Species 2 suggest plausible nonlinear effects of the detection conditions. The effect of time of day in Figure 6 has a peak early in the day followed by a plateau before the effect drops off rapidly at night. The effect of the duration of the survey shown in Figure 7 shows a rapid increase as duration starts to increase, but indicates diminishing returns after a period of time.

Conclusions

This paper has presented a method for combining boosted regression trees and probabilistic occupancy-detection models from the ecology literature. Our results indicate that this model can produce accurate estimates of the true occupancy probabilities for synthetic species without sacrificing the ability to predict observations, and that it performs favorably in comparison with logistic regression, boosted regression trees without latent structure, and occupancy-detection models without trees. We have also shown that tree-boosted occupancy-detection models can automaticaly discover and represent the relationships with the input variables that generated the synthetic species data, and we have given examples applying these models to the eBird data.

This work makes important contributions in both machine learning and ecology. Our incorporation of regression trees into occupancy-detection models represents the first application of functional gradient descent to probabilistic models with latent variables in machine learning. In ecology, merging boosted regression trees and occupancy models resolves an existing false dichotomy in species distribution modeling: that we can account for imperfect detection or build flexible models, but not both.

In future work, we will extend our work on functional gradient descent with latent variables to more complex models in which the latent variables cannot be marginalized away. We expect to achieve this by combining our current algorithm with an Expectation-Maximization approach and by developing appropriate initialization and regularization strategies to guide the optimization. We also plan to apply our methods to additional datasets, and we are developing an R package to make our methods conveniently available.

References

Breiman, L.; Friedman, J. H.; Olshen, R. A.; and Stone, C. J. 1984. *Classification and Regression Trees*. Wadsworth.

Dempster, A. P.; Laird, N. M.; and Rubin, D. B. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *JRSS Ser. B* 39(1):1–38.

Dietterich, T. G.; Hao, G.; and Ashenfelter, A. 2008. Gradient Tree Boosting for Training Conditional Random Fields. *JMLR* 9:2113–2139.

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

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

Kéry, M.; Gardner, B.; and Monnerat, C. 2010. Predicting species distributions from checklist data using site-occupancy models. *J. Biogeo.* 37:1851–1862.

MacKenzie, D. I.; Nichols, J. D.; Lachman, G. B.; Droege, S.; Royle, J. A.; 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. 2006. *Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence*. Elsevier.

Munson, M. A.; Webb, K.; Sheldon, D.; et al. 2010. The eBird Reference Dataset, Version 1.0.

Ridgeway, G. 2007. Generalized Boosted Models: A guide to the gbm package.

Yu, J.; Wong, W.-K.; and Hutchinson, R. A. 2010. Modeling Experts and Novices in Citizen Science data for Species Distribution Modeling. In *ICDM 2010*.