Creating detailed endangered species distribution maps using inductive modeling methods

Emilie Henderson[1], Timothy Howard[2]

[1]Oregon State University, Institute for Natural Resources, [2]New York Natural Heritage Program – The Nature Conservancy
Contact: Emilie.Henderson@oregonstate.edu

Summary

We studied how species distribution models might be created for all federally listed species in the United States and how they might be most useful. Illustrated here are modeled suitable habitats within New York for two widely distributed animals: the Karner blue butterfly (*Lycaeides melissa samuelis*), and and the bog turtle (*Clemmys muhlenbergii*), and fine-scale patterns for two narrowly distributed plant species in Oregon: Bradshaw's desertparsley (*Lomatium bradshawii*), and Willamette daisy (*Erigeron decumbens*). Unified maps and rigorous model validation are important components for providing effective tools to end users.

Background

Transportation projects are often hindered by insufficient information on potential locations of federally-listed threatened and endangered species. The current approach of using known observations of endangered species for project review often falls short. Known observations cannot highlight likely locations that have not yet been surveyed. Both transportation planners and regulators seek information on the potential or likely locations of at-risk species, rather than just known locations.

Methods

Data

This method requires two types of input data: points of species presence and absence, and rasters describing environmental factors that constrain the species' distribution.

Natural Heritage Program databases provide species presence locations and help experts interpret expected absence points. Rasters were all resampled to 10m resolution (30m in NY). They included spectral reflectance and spatial pattern descriptors from airphotos (National Air Photo Inventory Program) and LANDSAT images, elevation and topography, and climate descriptors (PRISM).

Modeling

For each species, we modeled the relationship between its presence/absence points (y) and the environmental rasters (x). We used Random Forest [1,3], a machine learning technique that extends classification trees by leveraging the predictive power of multiple trees. The technique yields a prediction that is analogous (but not identical) to a probability of habitat suitability. We built a prediction for each pixel in our map, and simplified the raster surface into categories relating to model certainty using the precision-recall F-measure [5].

Map Integration

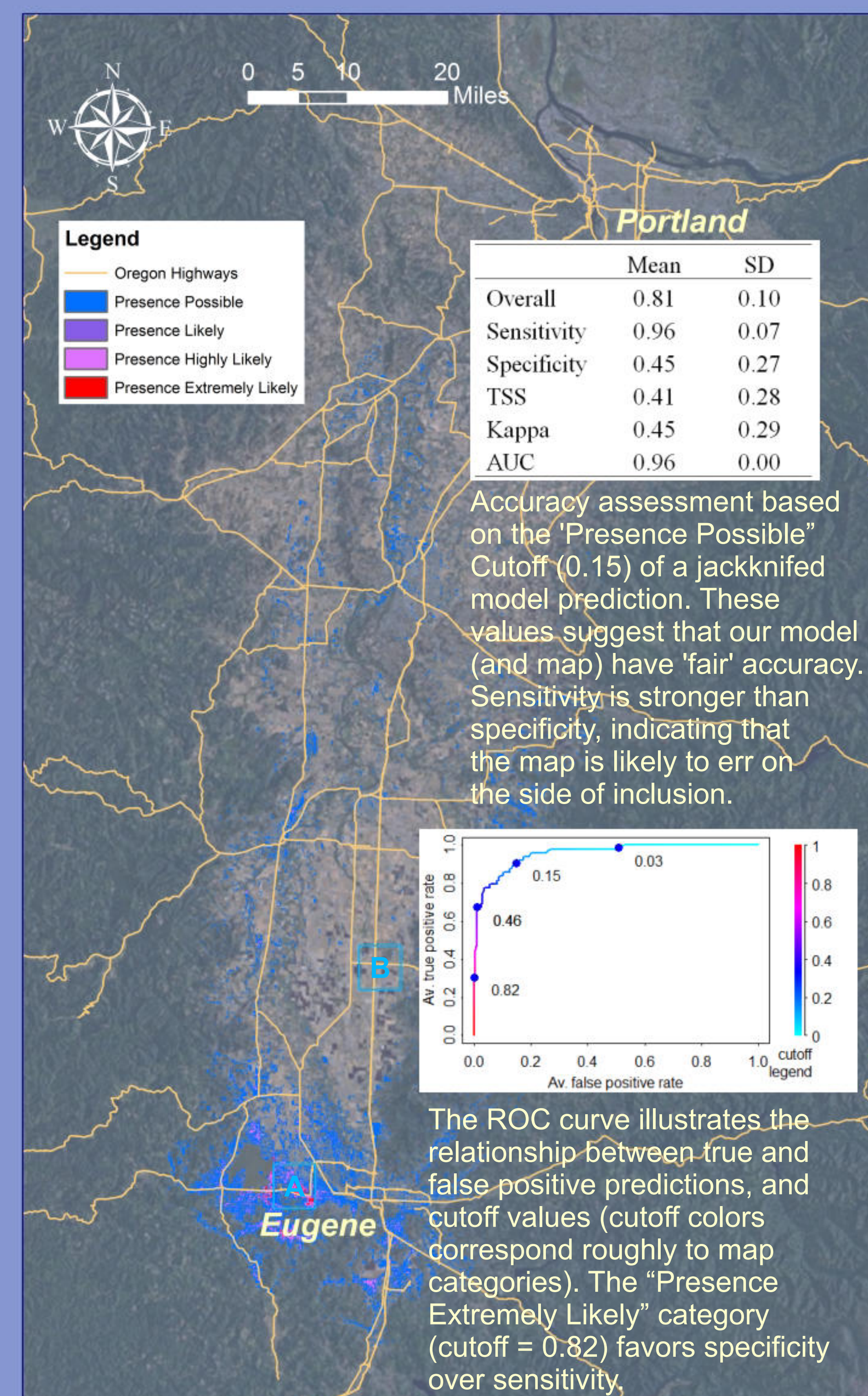
To integrate maps, we simplified the single-species maps to presence-absence (0 or 1), and then summed them, highlighting areas likely to contain both endangered species.

Accuracy Assessment

We assessed map accuracy by evaluating the ability to predict locations left out of a subsample model and then jackknifing all locations[2]. Standard accuracy assessments [2,5] were calculated on jackknifed predictions.

Results

Bradshaw's Desertparsley

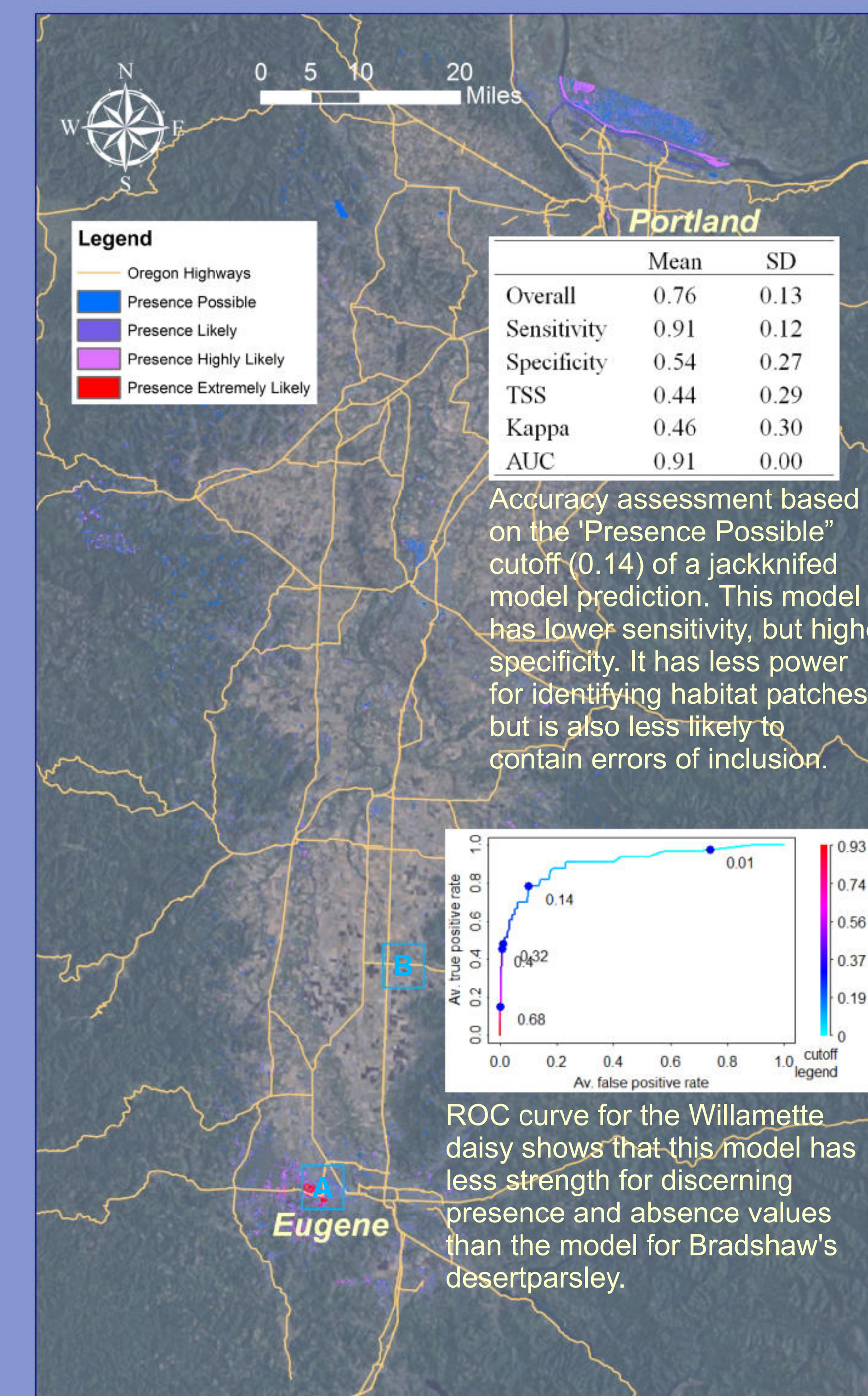


	Mean	SD
Overall	0.81	0.10
Sensitivity	0.96	0.07
Specificity	0.45	0.27
TSS	0.41	0.28
Kappa	0.45	0.29
AUC	0.96	0.00

Accuracy assessment based on the 'Presence Possible" Cutoff (0.15) of a jackknifed model prediction. These values suggest that our model (and map) have 'fair' accuracy. Sensitivity is stronger than specificity, indicating that the map is likely to err on the side of inclusion.

The ROC curve illustrates the relationship between true and false positive predictions, and cutoff values (cutoff colors correspond roughly to map categories). The "Presence Extremely Likely" category (cutoff = 0.82) favors specificity over sensitivity.

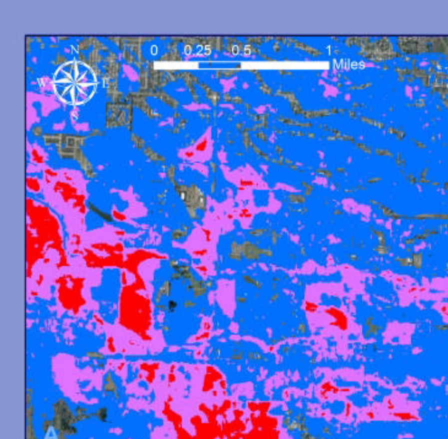
Willamette Daisy



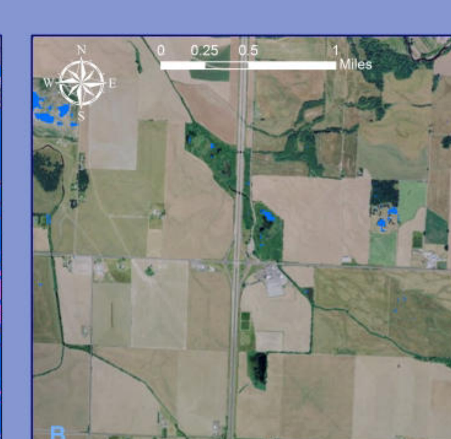
	Mean	SD
Overall	0.76	0.13
Sensitivity	0.91	0.12
Specificity	0.54	0.27
TSS	0.44	0.29
Kappa	0.46	0.30
AUC	0.91	0.00

Accuracy assessment based on the 'Presence Possible" cutoff (0.14) of a jackknifed model prediction. This model has lower sensitivity, but higher specificity. It has less power for identifying habitat patches, but is also less likely to contain errors of inclusion.

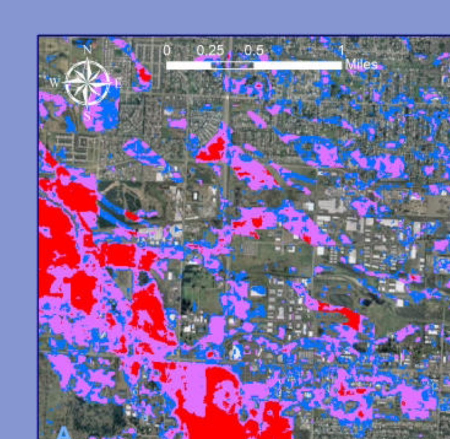
ROC curve for the Willamette daisy shows that this model has less strength for discerning presence and absence values than the model for Bradshaw's desertparsley.



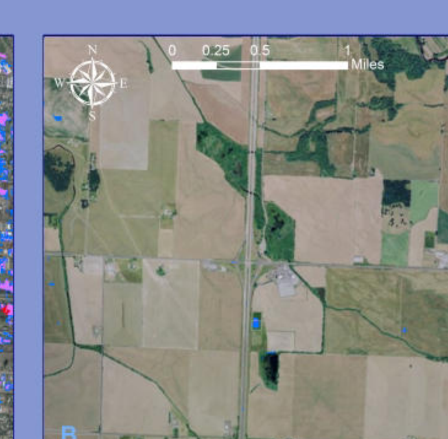
Bradshaw's desertparsley is abundant in and around Eugene's wetlands. In this closeup, we can see the model's strengths and weaknesses. The large area shown in blue may reflect the low specificity of the underlying model, which has a tendency to overmap likely habitat. (same legend as above)



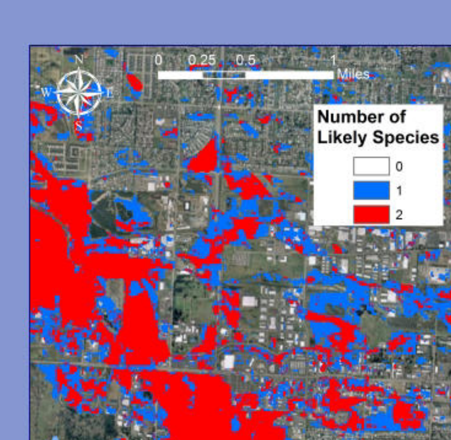
Further North, potential desertparsley habitat is uncommon. The map highlights areas to potentially survey on the ground for this species. Given the low specificity of the model, it is likely that some of the patches here contain no desertparsley.



Willamette daisy is common in Eugene's wetlands. Its distribution is similar to that of the desertparsley, although this habitat map shows less area in the "Presece Possible" category (blue) due to the higher specificity of the model. Because of the lower model sensitivity, a few habitat patches may be missed.



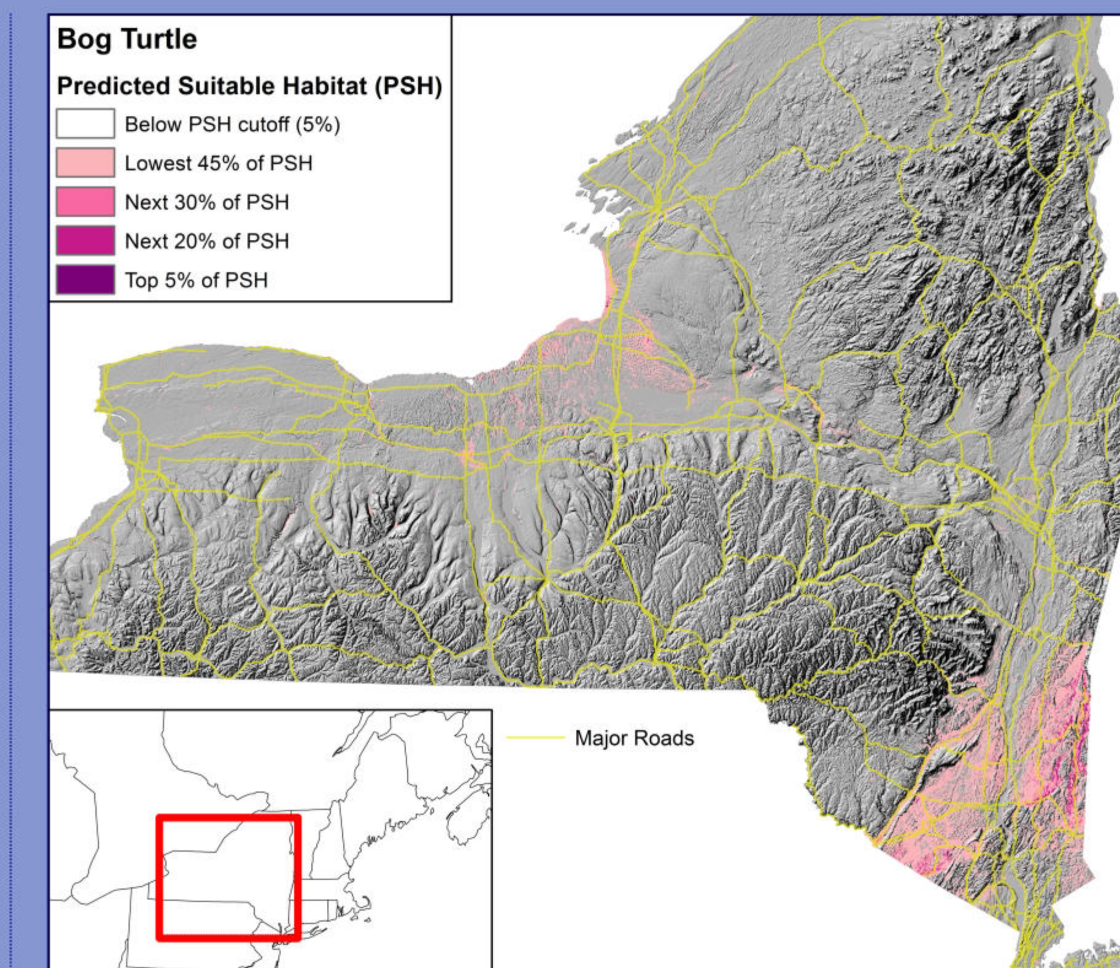
In this location, only a few small Patches of the Willamette daisy are likely. Because this model has higher specificity, it is more likely that highlighted habitat is accurately mapped than it was for the desertparsley.



By reclassifying each model to show a single value (1) for 'presence', we can sum layers for a map of diversity hotspots. Here, we an see how much the desertparsley and the daisy overlap in Eugene. A value of one here means that one of the above plants was ranked as 'Likely Present' (purple and red) by the maps above. A value of two indicates that both plants were ranked as such.

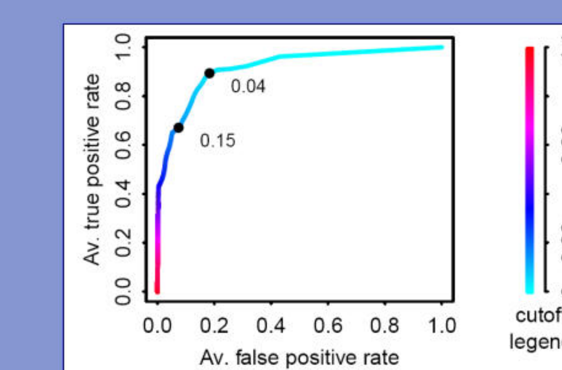
Careful attention to the model and map accuracy is vital for useful map integration. A species diversity map built from a set of unsatisfactory models will aggregate errors from each map.

Bog Turtle



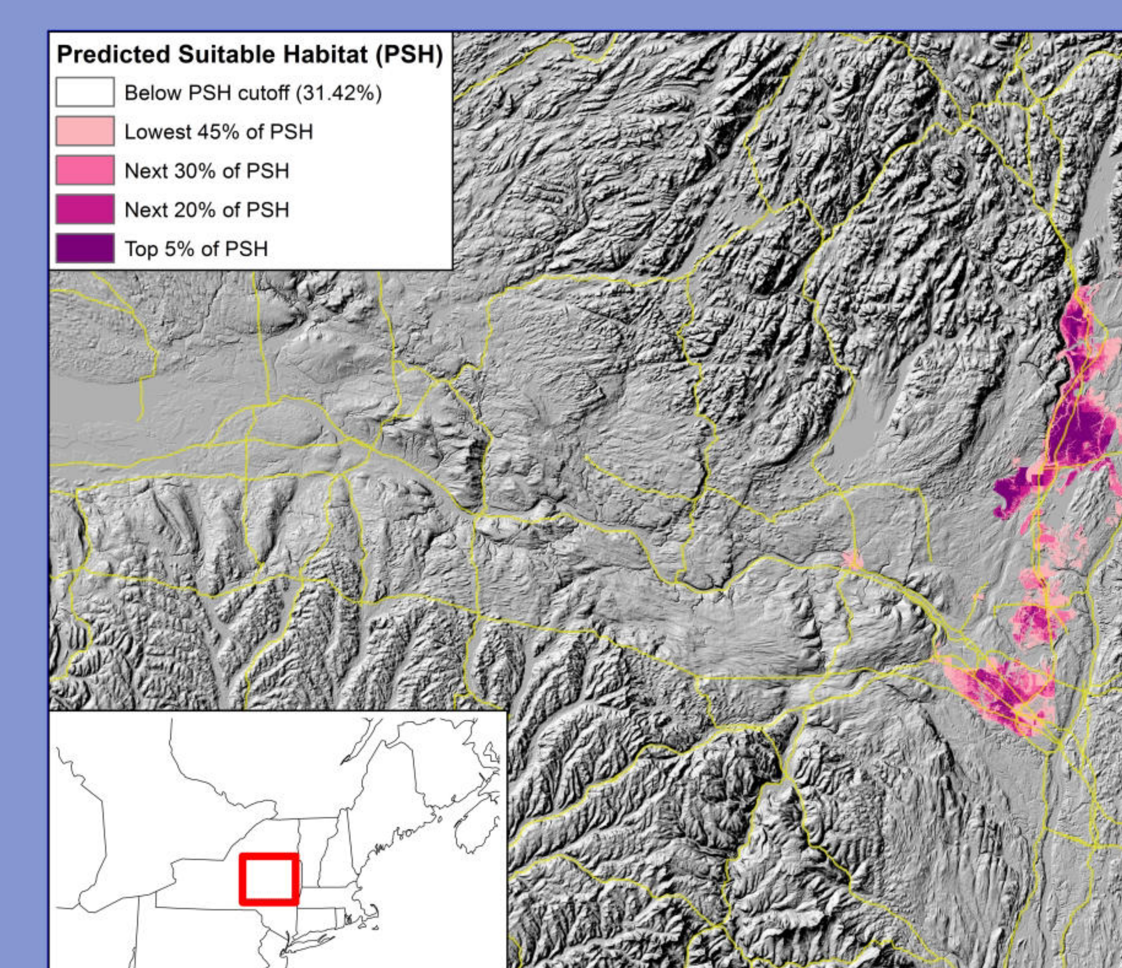
Name	Mean	SD	SEM
Overall Accuracy	0.91	0.16	0.03
Specificity	0.94	0.06	0.01
Sensitivity	0.87	0.32	0.06
TSS	0.81	0.32	0.06
Kappa	0.81	0.32	0.06
AUC	0.97	0.07	0.01

Validation statistics for 31 jackknife trials. Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve. This model is quite strong, with both high specificity and sensitivity. Correspondingly, the measures that integrate these two factors (TSS, Kappa and AUC) have high values.



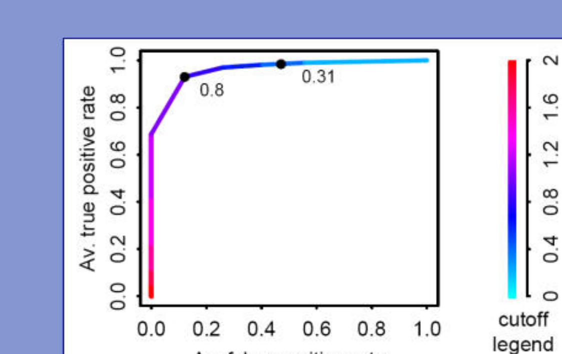
ROC curve, for the bog turtle model, averaged over 31 validation runs

Karner Blue Butterfly

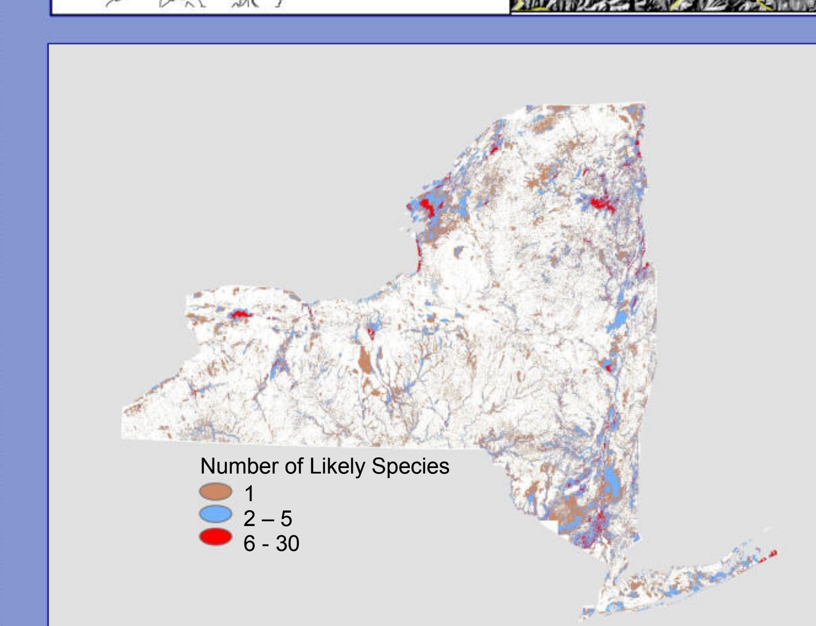


Name	Mean	SD	SEM
Overall Accuracy	0.96	0.14	0.01
Specificity	0.99	0.05	0.00
Sensitivity	0.92	0.27	0.03
TSS	0.91	0.27	0.03
Kappa	0.91	0.27	0.03
AUC	1.00	0.00	0.00

Validation statistics for 100 jackknife trials. Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve. This model is even stronger than the one behind the bog turtle map above.



ROC curve for the Karner Blue model, averaged over 31 validation runs.



An example of applying the same methodology described to the left for depicting suitable habitat for each species and then stacking them to create diversity hotspots.

Here we show the results of combining models for 317 rare species (not all Federally listed!) throughout New York. Some locations show overlaps of many species, yet most of New York contains no predicted suitable habitat.

Discussion

In order for these data to be truly transformative for transportation planning, the U.S. Fish and Wildlife Service has to assist in identifying reasonable probability cutoffs to flag sites to be surveyed for the species in a Section 7 review. Because of this need, the work also explored ways to create partnerships between the U.S. Fish and Wildlife Service and NatureServe to complete these maps, and the Federal Highways Administration (FHWA) to assist in funding their completion and making them accessible to transportation planners.

Conclusions

Maps built from species distribution models can provide information on endangered species previously not available to planners. Used during planning, these maps have potential to save time and expense during the implementation and planning phase of a transportation project.

These maps depict modeled suitable habitat and may over-predict some locations and under-predict others. They do not replace field surveys, but rather represent a new tool that can help planners avoid endangered species impacts. The maps serve as an information source during early transportation planning phases, and as a means for directing and prioritizing field visits.

Used properly, they can improve conservation outcomes while increasing the effectiveness of environmentally conscious planning, all in a more cost effective manner.

Acknowledgements

In Oregon, the Transportation Research Board & FHWA funded the plant models and New York's Department of Environmental Conservation and DEC's Hudson River Estuary Program funded the model development in New York. Jimmy Kagan facilitated the Oregon work and fostered the collaboration between New York & Oregon, and Eric Nielsen created the Oregon base input data.

References

1. Breiman, L. 2001. Random Forests. Machine Learning 45:5-32.
2. Fielding, A. H., and J. F. Bell, 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24:38-49.
3. Liaw, A. and M. Wiener, 2002. Classification and Regression by randomForest. R News 2:18-22.
4. R Development Core Team 2011. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
5. Sing, T., O. Sander, N. Beerenwinkel, T. Lengauer, 2005. ROCR: visualizing classifier performance in R. Bioinformatics 21:3940-3941.