Model accuracy in forecasting pathogen spread using climatic data

Annakate Schatz[1], Andrew Kramer[2]
[1]Mount Holyoke College, [2]University of Georgia, Odum School of Ecology

Introduction

When we discover a new emerging disease, one of the most urgent questions is where else it will spread. This project sought to answer that question with a model selection case study focused on time-sensitive predictive ability. Previous research conducted on model accuracy given data in time steps worked either with a single model fitting method or with a simulated disease (Vaclavik and Meentemeyer 2012, Patel unpublished). We extend such investigations by modeling a real disease, *Batrachochytrium dendrobatidis* (*Bd*), using multiple methods and subsets of the available occurrence data. In doing so, we gain a better sense of how quickly we can determine the potential extent of an emerging disease. If we can predict spread for *Bd*, then we can use similar modeling techniques to identify and monitor areas vulnerable to other diseases.

Objectives

1. Determine the amount of model training data needed to forecast *Bd* spread with acceptable accuracy.
2. Test multiple models to identify whether any method provides an acceptably accurate prediction of the observed spread of *Bd*.

Hypotheses

1. Prediction accuracy, as measured by several model evaluation statistics, will improve for all models with more training data.
2. Boosted regression trees (BRT) and Maximum Entropy (MaxEnt) will provide the most accurate predictions. This expectation is based on the methods' high levels of performance in Elith and Graham (2009) and Patel (unpublished) and MaxEnt's status as one of the most widely used species distribution models.

Methods

- data: Bd-Maps and WorldClim databases.
- models: BRT, MaxEnt, generalized linear model (GLM), k-Nearest Neighbor (k-NN), random forest (rF), Plug and Play Gaussian (PPG), and range bagging (RB).
- training and testing data: initial training set consisted of *Bd* data points from 1980-1995, after which data was added in four-year sets up to 2007; testing sets used data from all years following the end point of the training set.
- evaluations: area under the receiver operating characteristic curve (AUC), Cohen's kappa, True Skill Statistic (TSS), and false negative rate (FNR).

Results

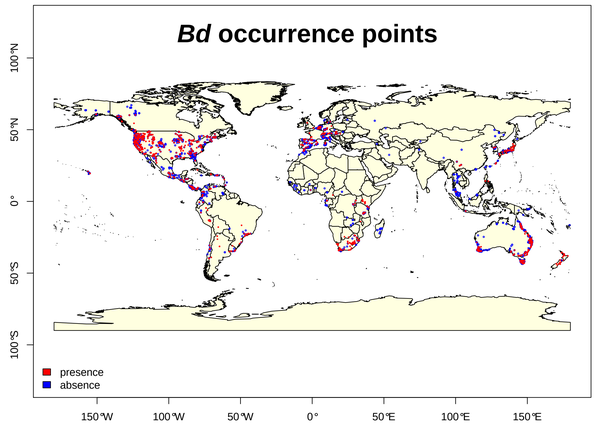

Figure 1. Bd presence and absence points. Areas of overlap might explain some error in predictions, specifically FNRs. Overlap can occur due to differently susceptible species in one location or multiple samples from a location over time.

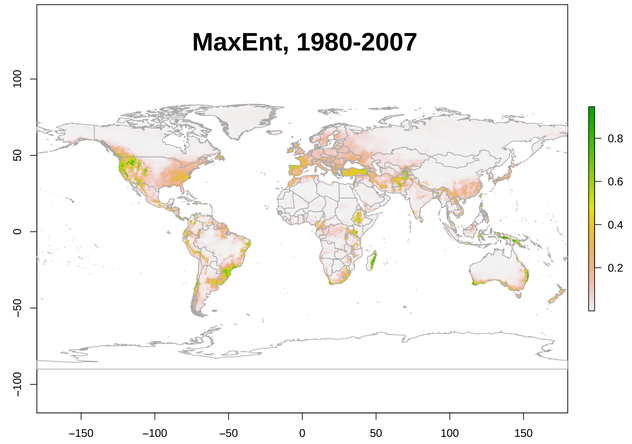

Figure 2. Relative probability of *Bd* occurrence from MaxEnt prediction, trained on 1980-2007 data and tested on 2007-2011 data. MaxEnt performed best overall on time-step predictions.

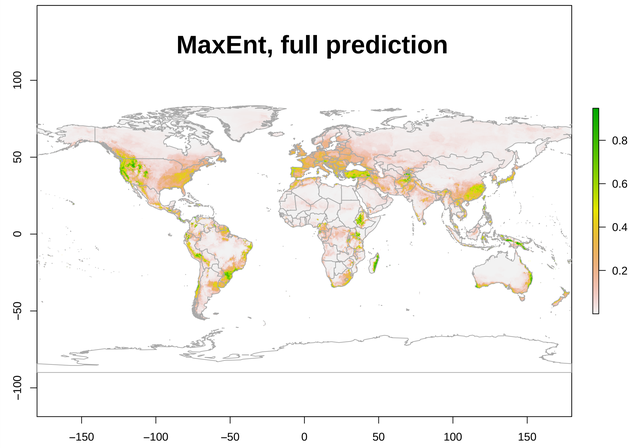

Figure 3. MaxEnt prediction, trained on 80% of all data and tested on the remaining 20%. This map's similarity to fig. 2 supports analysis of MaxEnt as the best overall modeling method.

Area under the receiving operator characteristic curve

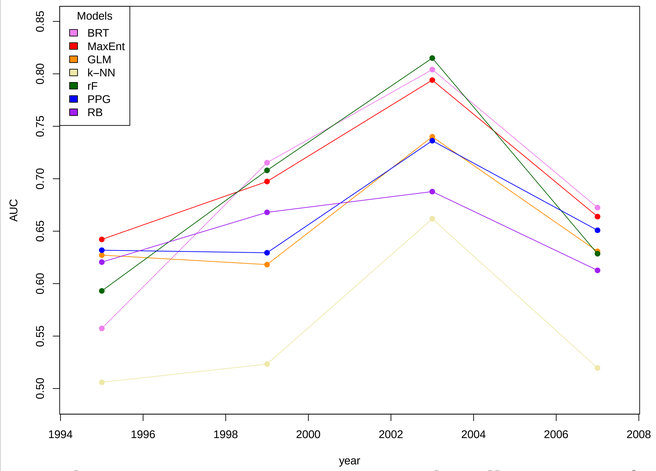

Figure 4. AUC scores for all seven model types across time-steps, where 0.5 indicates prediction no better than random and 1.0 indicates perfect prediction. We expected the initial upward trend in scores; more presence training data allows more refined predictions. The universal drop at the final time-step could be due to too few presence points in the final test set.

Conclusions

- *Hypothesis:* Prediction accuracy, as expressed by our evaluations, will improve for all models with more training data.
- *Finding:* This hypothesis was not supported by our results. All models received their highest AUC score when trained on data from 1980-2003 (the third of four time-steps), and all except PPG received their highest kappa and TSS there as well.

- *Hypothesis:* BRT and MaxEnt will provide the most accurate predictions of *Bd* spread.
- *Finding:* MaxEnt did in fact provide the most accurate predictions, followed by BRT and rF. All three performed well on AUC evaluation. MaxEnt, however, had a lower FNR than both BRT and rF 50% of the time, and a lower FNR than at least one 100% of the time. Thus we selected MaxEnt as our best method overall.

Acknowledgements

We thank Dr. John Drake of the Odum School of Ecology at UGA for letting us use his unpublished Plug and Play regularized Gaussian model. We also thank Deeran Patel of the Drake lab for sharing a data preparation function and extensive modeling code.
Funded by NSF award number 1156707, Population Biology of Infectious Diseases REU.

References

1. Bd-Maps. http://www.bd-maps.net/.
2. Elith J, Graham CH. 2009. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. Ecography 32: 66-77.
3. Patel D, Drake JM. unpublished. The Simulation of Infectious Disease Transmission within a Virtual Environment.
4. Vaclavik T, Meentemeyer RK. 2012. Equilibrium or not? Modeling potential distribution of invasive species in different stages of invasion. Diversity Distrib. 18: 73-83.
5. WorldClim: Global Climate Data. http://www.worldclim.org/.