Maxent Modeling Ideas for BIEN3

Cory Merow

April 29, 2014

1 Possible Changes

1.1 Easy

1. Reduce run time
2. Do we have enough background points?
3. Increase regularization penalty?
4. Remove threshold features?
5. Do we need all 40 (bioclim + spatial) predictors?

2 Results for easy changes

2.1 Number of Background Points

Summary:

1. 10k background isn't enough
2. 50k background points might be an ok default, but it still probably doesn't lead to convergence, is WAY slower (around 30 min processor time), and requires lots more RAM (>2gb in many cases)
3. More presence samples suggests the need for more background points
4. it's not clear how much the background really affects the maps (AUC is already really high with 10k background in many cases), but it could be dangerous to take shortcuts if we're not actually monitoring each model

The following is based on a random sample of roughly 200 north american trees for which we have the Little range maps. I use the value of the gain function, the thing that maxent optimizes during model fitting as a measure of convergence. exp[gain] is interpreted as the ratio of (relative) probability of presence at an average presence location to (relative) probability of presence at a background location. The value of the gain is roughly between 2-7, depending on the species.

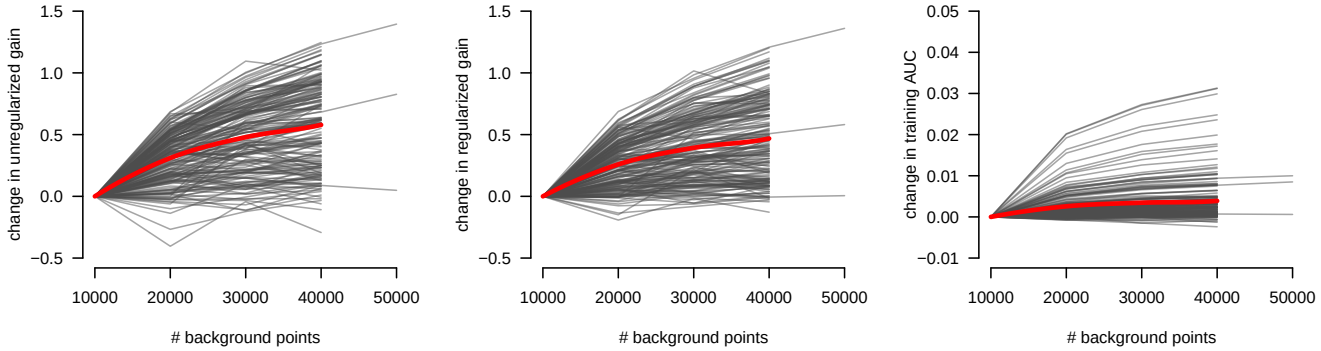

Figure A1: Regularized gain is optimized by maxent to fit the model. 10k background isn't enough for convergence. Red line is a loess fit. These plots show how much the gain is *improved* by increasing the number of background points (that is, because each species has a different gain for different qualities of models, i made all species comparable by looking only at the improvement in gain relative to its value for that species with 10k background points.

See [Fithian and Hastie, 2013] for issues with convergence of gain for maxent-like models

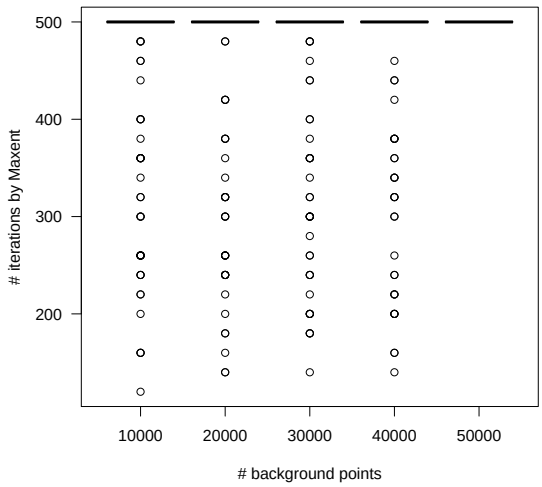

Figure A2: Yes, this is a box plot. Looks like Maxent runs up against its max iteration threshold (while optimizing gain), so we probably need to increase that to get convergence. I'd like to check whether increasing the iteration threshold for fewer background points would allow us to get better convergence of the gain without having to use so many background points (which is slow).

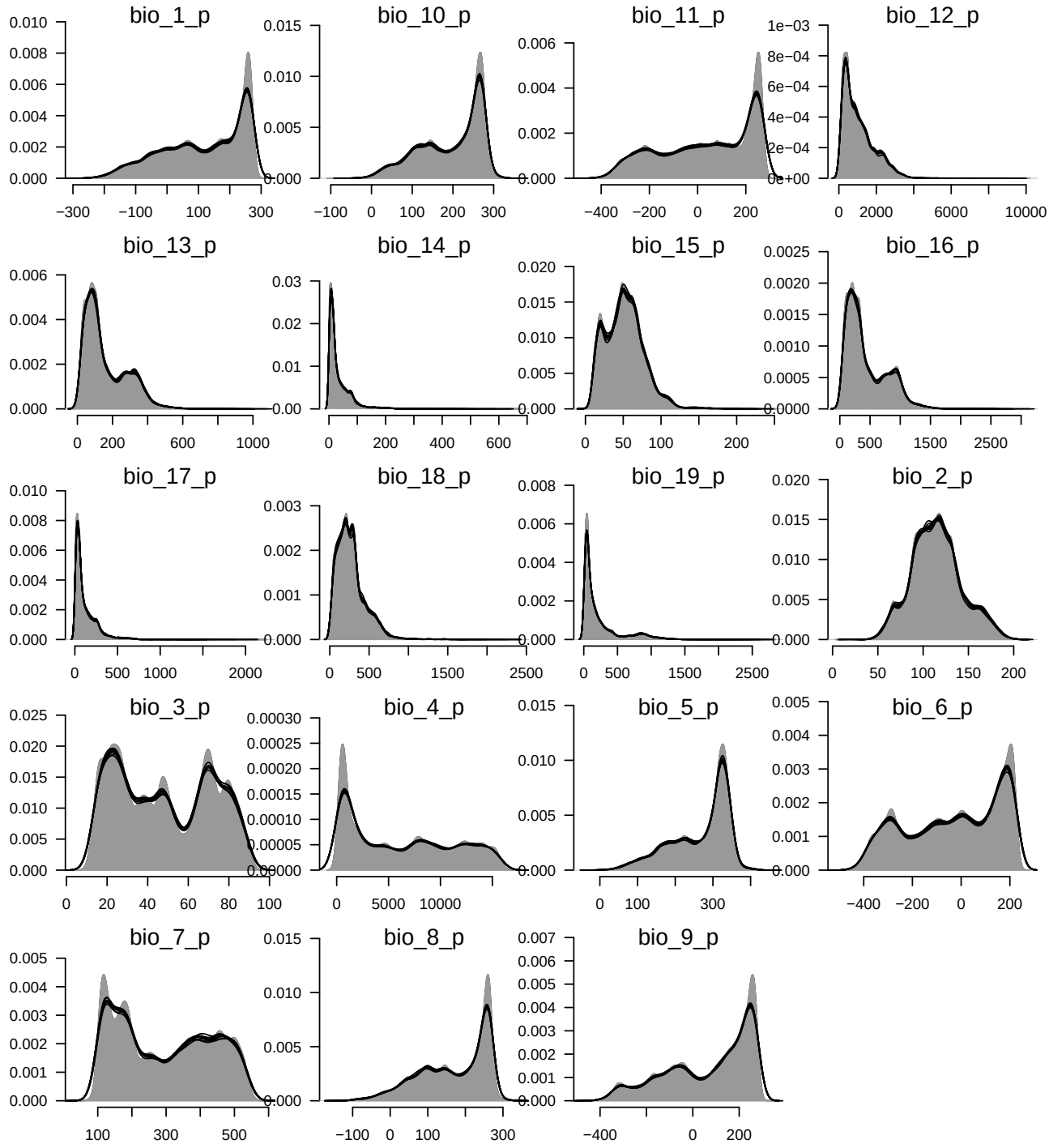

Figure A3: Comparison of full density (grey polygon) of each bioclim filter predictor compared to density esimtated from 10k background points (black lines; 10 replicate runs for different background samples). This is just a heuristic that I use to see if our background sample looks like the full background. I would've thought 10k background samples is plenty based on these, though the samples miss the extremes. And of course this is univariate and I'm not checking coverage in multidimensional environmental space. Based on the convergence of the gain (Fig. A1), evidently this is not close enough. The analogous plots for the spatial predictors look basically the same.

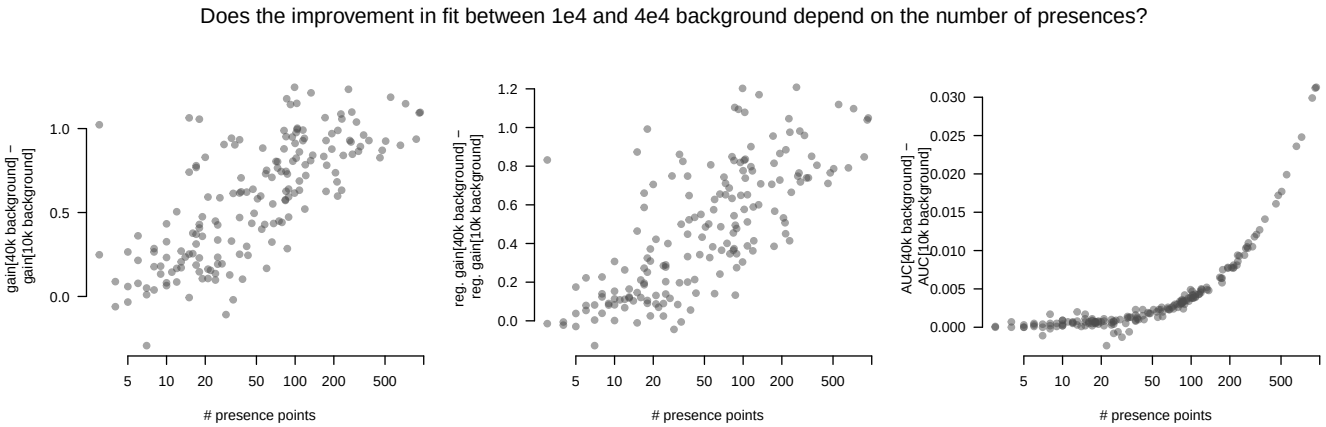

Figure A4: More presence samples suggests the need for more background points.

References

William Fithian and Trevor Hastie. Finite-sample equivalence of several statistical models for presence-only data. *The Annals of Applied Statistics*, In Press, 2013.