Evaluating Ecological Niche Modelling Techniques

A. H. Dekker [a] **and J. J. L. Rowley** [b]

[a] *Dekker Consulting, P.O. Box 3925, Manuka, ACT, 2603, Australia*
[b] *Australian Museum Research Institute, 6 College St, Sydney, NSW, Australia*
Email: dekker@acm.org

Abstract: Ecological niche modelling is an important method for predicting the range of a species or genus. In this paper we examine a particularly challenging case of niche modelling where:

- the number of samples is small (in our case, 12, not including "background" datapoints);
- samples are taken not from an individual species, but from a group of related species within a genus (in our case, Vietnamese frogs from species within the *L. applebyi* group of the genus *Leptolalax*); and
- the predicted distribution is required to generalise to other species within the group, which are not part of the training dataset (because we hope to identify regions where undiscovered species may be found).

Of necessity, such modelling will be imprecise, and dependent on the choice of training samples. However, even an approximate predicted distribution will support conservation efforts and guide the search for additional specimens (including those of undiscovered species within the group).

Figure 1 maps the samples used in our study, while Figure 2 illustrates our experimental process. Figure 3 shows some of the frog species used. We conducted niche modelling using the Generalised Linear Model (GLM), Maxent, Random Forest, Domain, and Bioclim methods.

Our experiments showed that, for the small training datasets used in our study, GLM outperformed the other methods used. With appropriate parameters, the predicted distribution produced by GLM always included at least half the independent test samples. Similar performance was obtained with GLM even when the size of the training dataset was reduced to just four samples (plus "background" datapoints).

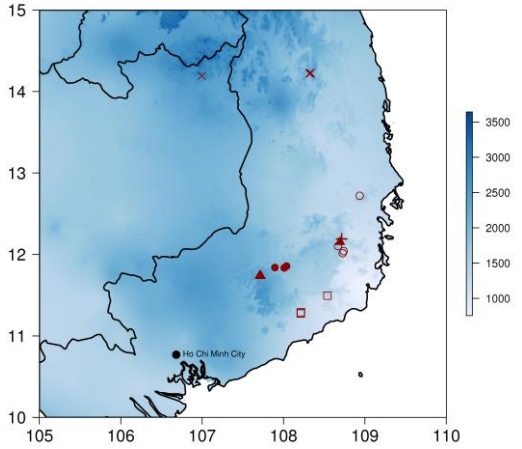


Figure 1. The 24 *Leptolalax* samples used in our study, marked on the map of central Vietnam. Map colouring indicates annual precipitation in mm, which is one of the 19 bioclimatic variables used.

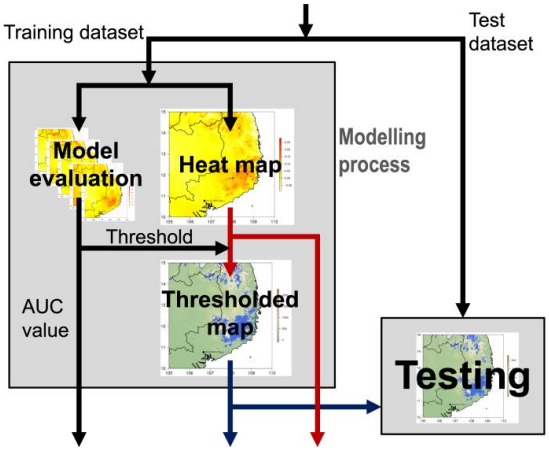


Figure 2. Our experimental process. The 24 *Leptolalax* samples are partitioned into a training dataset and an independent test dataset, which is used to assess the performance of the modelling process.

Figure 3. Some of the frogs in the study (L to R): *L. melicus*, Lineage 3, *L. bidoupensis* (photos: J. Rowley).

Keywords: *Ecological niche modelling, species distribution modelling, generalised linear model, frogs*

1. INTRODUCTION

In a world in which biodiversity loss is occurring at a rapid pace, particularly in the face of habitat loss, it is vital to prioritise limited conservation actions, especially in terms of protected areas selection. A major obstacle to this is our incomplete knowledge of biodiversity, and how it varies spatially. One method that may facilitate our spatial understanding of a poorly known region or taxonomic group is ecological niche modelling. Ecological niche modelling has been used in the past to predict the range of species or predict where currently unknown species may occur (generalising from a set of observations), thereby facilitating species discovery (Raxworthy *et al*., 2003; Rowley *et al*., 2015).

Amphibians are one of the most threatened groups of animals, with 41% of all amphibian species threatened with extinction (IUCN, 2015). In areas such as Southeast Asia, our knowledge of this group is poor, and the threats facing them are great (Rowley *et al*. 2010a). We are therefore interested in using ecological niche modelling to predict the range of amphibian species, especially in Southeast Asia.

Ecological niche modelling generally performs better with larger sample sizes (more observed locations), but for specimens in forested areas that are physically small and difficult to detect, such observations are not easy to obtain. There is therefore a need to assess how well different niche modelling techniques perform with small sample sizes. In this paper, we report experiments addressing this question.

In our experiments, we used a set of 24 observed locations of frogs from species within the *Leptolalax applebyi* group of the genus *Leptolalax* Dubois 1983, all within the Central Highlands of Vietnam and adjacent northeastern Cambodia. These locations are mapped in Figure 1 and Figure 4. Our study area is the region bounded by 105°E to 110°E longitude and 10°N to 15°N latitude.

Some of these observations reflect recognised species within the *L. applebyi* group (Rowley *et al*., 2010b; 2011), while others represent molecular lineages that may constitute new species, but have not yet been formally described. Of the 24 observations, 22 are taken from Rowley *et al*. (2015), and 2 (of *L. pyrrhops*) from Poyarkov *et al*. (2015). The 24 observations are divided into six groups of four, as shown in Figure 4. These six groups have been defined so as to minimise the species overlap between them. This is because we hope to identify regions where undiscovered species within the *L. applebyi* group may be found, and we are therefore particularly interested in whether the ecological niche modelling methods we are using can generalise from one set of species to another, closely related, set of species within the same genus.

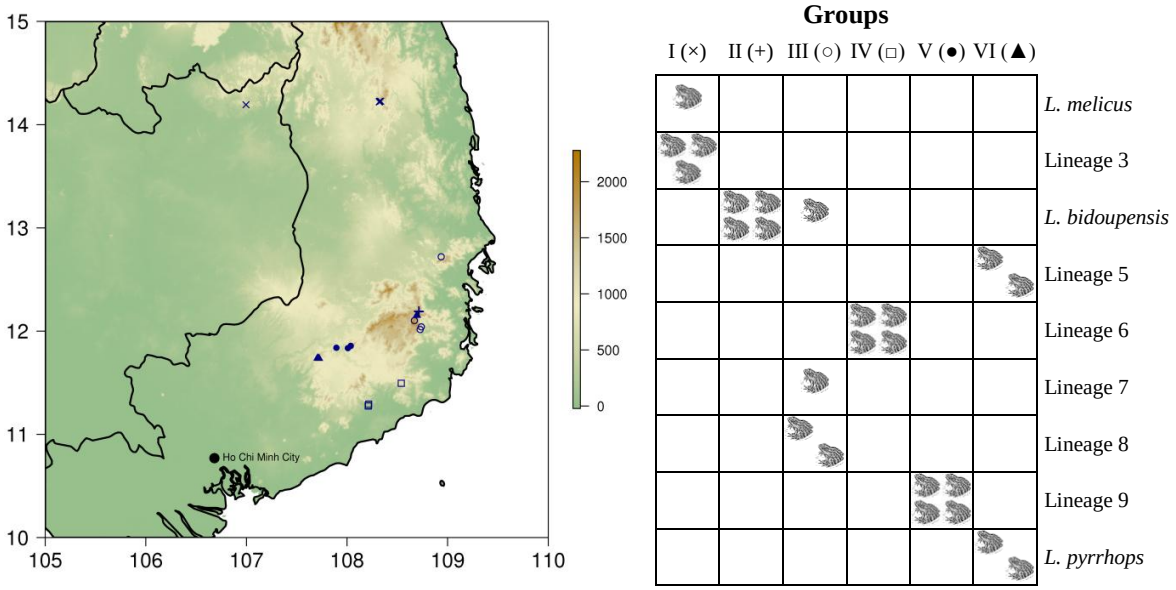


Figure 4. The 24 *Leptolalax* samples used in our study, arranged in six groups of four. Locations are marked on the map of central Vietnam. Map colouring indicates altitude in metres.

The frogs of the *L. applebyi* group (some of which are illustrated in Figure 3) display species diversity, much of which may be unrecognised. Some lineages of these frogs appear to be restricted to specific watershed basins, while others, in spite of morphological similarities, differ significantly in their mating calls. The frogs are small (with a snout-vent length of less than 30 mm), brown, and generally hidden in leaf litter on the forest floor of hilly evergreen forest, which makes finding specimens difficult (Rowley *et al*., 2010b; 2011; 2015).

2. EXPERIMENTAL PROCESS

Our experimental process is illustrated in Figure 2. Our set of 24 samples is divided into a training set of 12 (three groups) used for modelling (plus "background" datapoints), and an independent test set of 12 (three groups) used for testing. There are 20 possible training/test partitions, and our experiment used all 20.

We conducted niche modelling on each training set in order to produce a predicted range, and then used the independent test set to assess the quality of the prediction. The primary assessment of quality was the number of matches between the test set and the predicted range, i.e. the number of test samples successfully predicted. This is a challenging test because, apart from one small overlap, the training set and the test set represent different species or molecular lineages.

3. MODELLING PROCESS

Our niche modelling process is intended to replicate the actions of a researcher faced with a small set of 12 samples, from which he or she wishes to predict a range not only for the species within the set, but also for potentially undiscovered closely related species. We used the R software toolkit (version 2.15.3) with the "raster" (version 2.2-31), "randomForest" (version 4.6-7), and "dismo" (version 0.9-3) packages (Hijmans and Elith, 2015). The latter was interfaced to version 3.3.3k of the MaxEnt program.

As predictive data, we used 30-second gridded climatic data from the WorldClim database at www.worldclim.org (Hijmans *et al*., 2005). To account for correlations among the data, we performed a principal components analysis on the 19 WorldClim "bioclimatic variables" for all land grid cells in the study area. The first eight principal components were used, which together accounted for 99.98% of the variance in the data. Due to the association of *Leptolalax* species with streams in sloping terrain, we supplemented these eight climatic principal components with the slope of terrain, calculated from the WorldClim elevation data using the "terrain" function in R.

We used six methods to infer a "heat map" (spatial function predicting values in a 0/1 range) from the 12 samples and the predictive data. These were Maxent (Phillips *et al*., 2006; Elith *et al*., 2011), Random Forests (Breiman, 2001), Domain (Carpenter *et al*., 1993), Bioclim (Busby, 1991), generalised linear modelling (GLM) with a Gaussian error distribution, and GLM with a Poisson error distribution (Maindonald & Braun, 2007). We abbreviate the last two as GLM(g) and GLM(p). It should be noted that the widely used Maxent method is essentially equivalent to a version of GLM (Renner and Warton, 2013). However, Maxent appears to become overly conservative in its prediction when (as in our study) the training datasets become very small. For all but the Bioclim method, we used 1,000 randomly chosen "background" datapoints, randomly selected from the land region of the study area, but chosen to be at least 5 km from the training samples.

The "heat map" values were transformed to a predicted range by a "leave out k" (jacknifing) thresholding process, in which 20 new heat maps were constructed, each using $12-k$ randomly chosen samples, and each evaluated using the remaining k samples. This process produces an "internal" measure of model quality, the median area under the Receiver Operating Characteristic curve, or AUC value (Wisz *et al*., 2008). The median value of the maximum of the sum of the sensitivity (true positive rate) and specificity (true negative rate) was used as a threshold, as in Liu *et al*. (2005). The experiment was repeated for k = 2, 3, 4, 5, and 6.

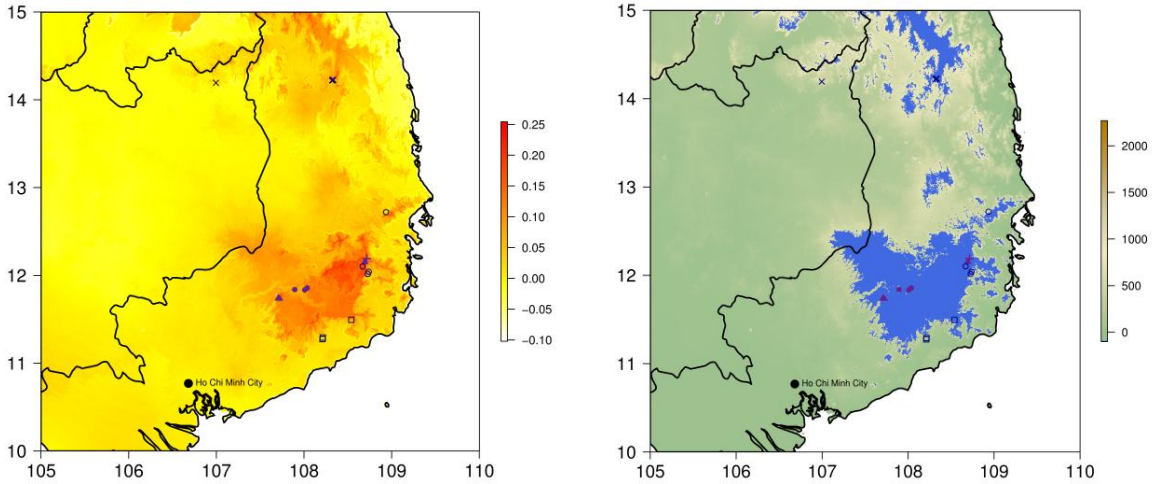


Figure 5. Worst case for GLM(g), using groups II (+), V (●), and VI (▲) for modelling (purple), and groups I (×), III (○), and IV (□) for testing (navy blue). Left: "heat map" before thresholding. Right: predicted range (blue) after thresholding with k = 4. The range's area is 23,777 km^2, and it contains 6 of the 12 test samples.

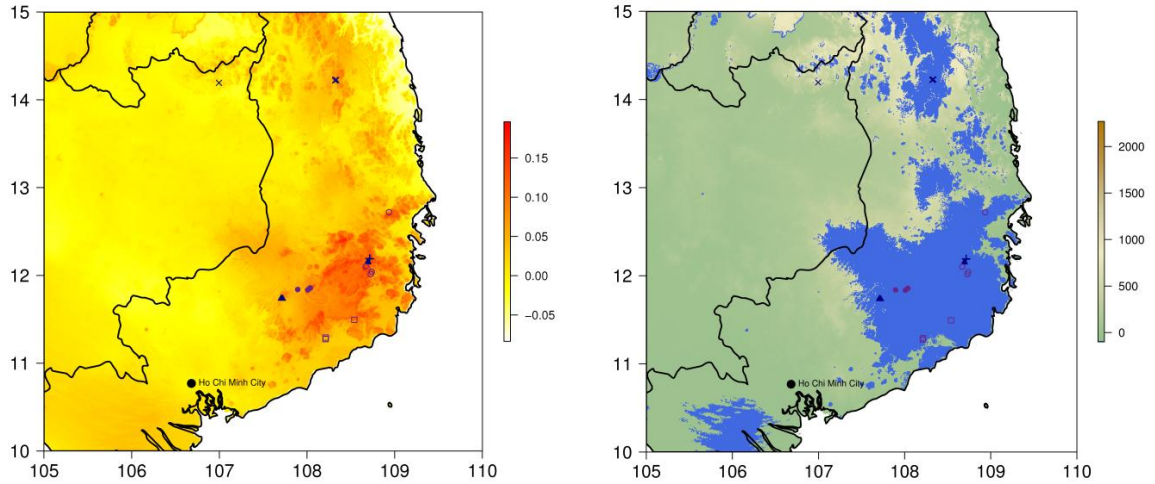
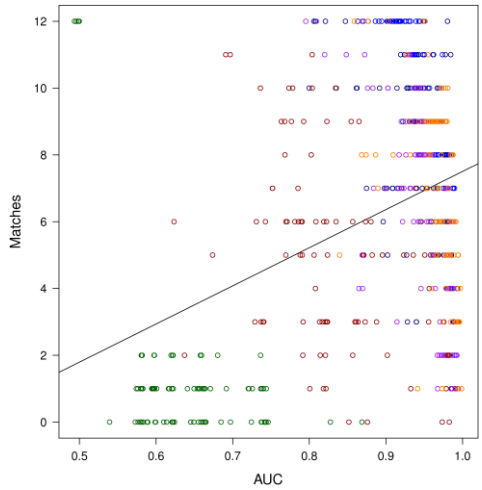
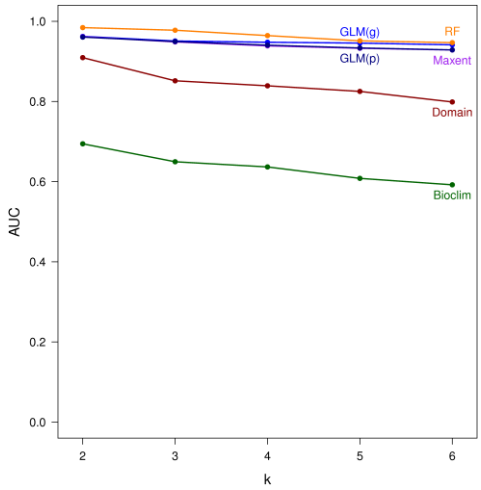


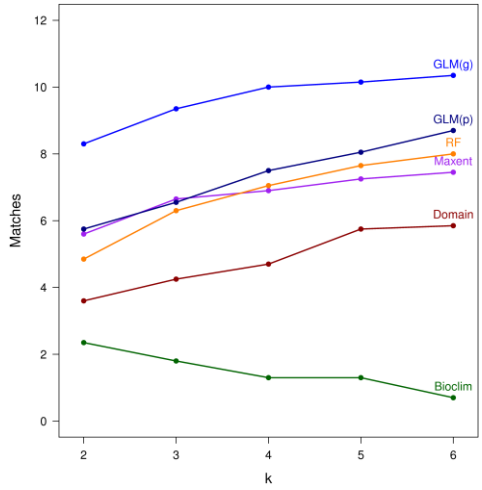
Figure 6. Best case for GLM(g), using groups III (○), IV (□), and V (●) for modelling (purple), and groups I (×), II (+), and VI (▲) for testing (navy blue), thresholded using $k = 4$. The more diverse training set has resulted in a larger predicted range than for Figure 5. This predicted range (in blue on the right) has an area of 48,956 km^2, and contains all 12 test samples.



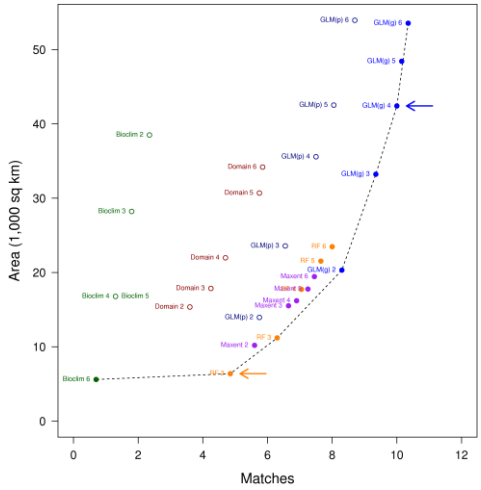
(a) Correlation between AUC values and number of matches is weak ($r = 0.39$).



(b) AUC values for different methods, averaged over all 20 train/test partitions (RF = Random Forests).



(c) Number of matches for different methods, averaged over all 20 train/test partitions.



(d) Multi-objective optimisation for no. of matches and range area. Two measures of the "best" result are the Pareto frontier (solid points) and the convex hull (dashed lines). Arrows highlight 2 possible optima.

Figure 7. Experimental results.

4. EXPERIMENTAL RESULTS

Figure 5 and Figure 6 show examples of the heat maps and thresholded maps produced by the above modelling process, generated using GLM(g) and $k = 4$, a combination which performed well, and which we recommend as the best approach to use. Out of the 20 different training/testing partitions, these figures represent the worst and best case respectively for GLM(g) and $k = 4$. Note that the heat maps for different methods are not comparable, although the thresholded maps are.

Figure 7 summarises the experimental results. Figure 7(a) shows the correlation between the AUC values (the "internal" quality measure) and the number of matches between the test set and the predicted range (the independent quality measure). This correlation was very weak ($r = 0.39$), which means that the AUC values (the "internal" quality measure) tell us little about the ability to generalise to the different species in the independent test set. The AUC values should therefore be treated with considerable caution.

Figure 7(b) shows the AUC values for different methods, averaged over all 20 train/test partitions. The GLM(g), GLM(p), Random Forest, and Maxent methods all scored highly, with mean AUC values in the ranges 0.941 to 0.962, 0.929 to 0.961, 0.947 to 0.984, and 0.928 to 0.961 respectively.

The number of matches, however, differed substantially between methods, as shown in Figure 7(c). Mean values for the GLM(g), GLM(p), Random Forest, Maxent, Domain, and Bioclim methods ranged from 8.30 to 10.35, 5.75 to 8.70, 4.85 to 8.00, 5.60 to 7.45, 3.60 to 5.85, and 0.70 to 2.35 respectively. The best-performing method was GLM(g) with $k \geq 4$. For no experimental run did this combination ever match less than 6 experimental test samples.

A potential problem with using the number of matches as a quality measure is that over-predicting the range may give an unfairly high score. We therefore considered multi-objective optimisation (Branke *et al.*, 2008) to maximise the number of matches while minimising the predicted area. Maximising the number of matches is equivalent to minimising the false negative rate, while minimising the area is equivalent to minimising the false positive rate (since "background" training points are randomly distributed). Figure 7(d) illustrates this assessment process: the best solutions are towards the lower right of the diagram. The two arrows in Figure 7(d) show two potential optima. If greater priority is given to maximising the number of matches, then GLM(g) with $k = 4$ performs best, and this is the combination we recommend for genus distribution modelling with small samples.

The reduced priority for minimising the area is justified by the fact that the "background" training points do not represent true absence data. In some cases they represent points where frogs in the *L. applebyi* group are actually found. The "false positives" given by a larger area therefore do not truly represent false positives.

If greater priority is given to minimising the predicted area, then Random Forests with $k = 2$ perform best. This more conservative prediction gives the greatest probability of finding test samples within the predicted range (i.e. the greatest number of test samples per km^2), and might therefore be useful in identifying priority search areas for new species in the genus. However, as Figure 8 shows, such prioritisation can be done even more effectively by using the heat map values produced by GLM(g).

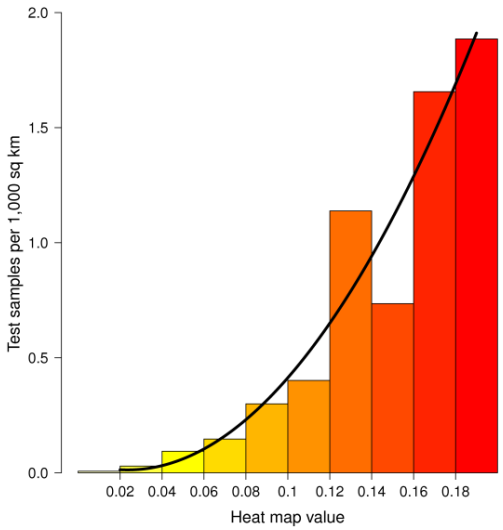


Figure 8. Heat map values calculated from the training samples with GLM(g) are a guide to the density of test samples, and can therefore be used to prioritise a search for new species within the predicted range. The density of test samples here is calculated over all 20 training/test partitions. The correlation between heat map values and the density of test samples is 0.96, fitting the curve $68.5\,v^2 - 3.2\,v + 0.05$.

5. FOUR SAMPLES?

We have shown that the GLM(g) method performs well with just 12 training samples. Can we reduce this even further, to just four? Previous work with geckos (Pearson *et al.*, 2007) had indicated that Maxent would perform reasonably well with as few as five training samples. We therefore repeated our experiment using the six individual groups I to VI as training sets (plus "background" datapoints), using $k = 2$ for each method.

The Bioclim, GLM(p), and Domain methods performed very poorly under this challenge, with the predicted area sometimes matching none of the 20 test samples, and sometimes uselessly predicting the entire study area (due to zero-variance errors). Maxent and Random Forests performed somewhat better, with 0 to 14 matches (a mean of 4.0) and 2 to 9 matches (a mean of 5.7) respectively. GLM(g) performed better still, with 9 to 15 matches (a mean of 13.2). In fact, GLM(g) appears to perform almost as well with four training samples as with 12. In addition, with these very small training sets, the superiority of GLM(g) over Maxent is even greater than when sets of 12 training samples were used. Figure 9 illustrates the worst case for GLM(g).

AUC values exaggerated the performance of Maxent in this experiment (giving a range of 0.954 to 0.999, with a mean of 0.982) compared to GLM(g), which gave a range of 0.918 to 0.999, with a mean of 0.971. Random Forests had lower AUC scores (0.721 to 0.999, with a mean of 0.906). In general, the AUC values generated by the modelling process appear to be a relatively weak guide to performance in our situation of small training sets being generalised to a set of different (though closely related) species. The selection of modelling methods based solely on AUC scores may therefore be suboptimal for our purposes.

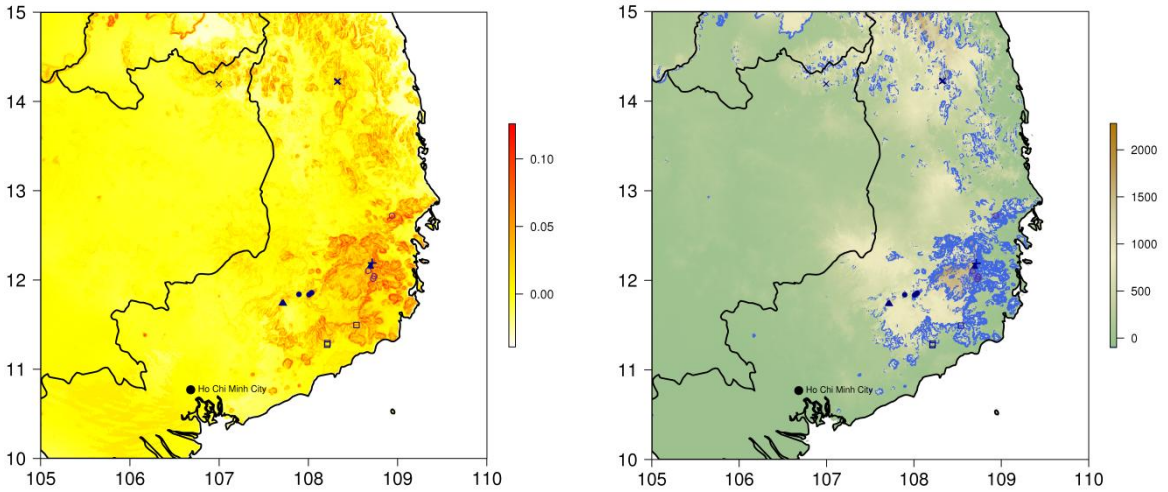


Figure 9. Worst case for GLM(g) with a single-group training set, using group III (○) for modelling (purple), and groups I (×), II (+), IV (□),V (●), and VI (▲) for testing (navy blue). The predicted range in blue on the right has an area of 14,081 km^2, and contains 9 of the 20 test samples. For comparison, the best case matches 15 out of 20 samples, and the mean case 13.2.

6. DISCUSSION AND CONCLUSIONS

We have shown that ecological niche modelling is possible even with small training datasets taken from different (but related) species. Indeed, useful results can be obtained even with just four training samples. Generalised linear modelling (GLM) with a Gaussian error distribution performs better than the Maxent, Random Forest, Bioclim, and Domain methods. Maxent, which performs very well with larger training datasets, appears to give overly conservative predictions when the number of training samples is small.

Previous work using 46 different species (Wisz *et al*., 2008) had indicated that GLM and Maxent would perform well with as few as 10 samples, with Maxent outperforming GLM. This finding was based on AUC values computed from the heat maps, using independent test data. In contrast, we have assessed the performance of different methods based on thresholded maps. This is because the modelling process we are assessing (see Figure 2) includes all the steps which an ecological modeller would go through, including thresholding (i.e. we are assessing that whole process, not just the algorithm being used). Using the same thresholding method, we found that Maxent gave overly conservative predictions of the thresholded range.

Our primary interest lies in supporting conservation efforts and guiding the search for additional specimens – not only of known species, but also of related species yet to be discovered. For these purposes an overly conservative range prediction is undesirable, and we have therefore assigned higher priority to matching the test dataset than to minimising the area of the predicted range.

While ecological niche modelling with very small training datasets will necessarily be somewhat imprecise, the quality of the predicted ranges is certainly adequate to support further research, to guide the search for additional specimens, and to identify areas where deforestation and environmental damage might be of particular concern. Since the heat map correlates with the density of test samples (and therefore, presumably, also with the true density), it can be used to prioritise search and conservation activities within the predicted range.

REFERENCES

Branke, J., Deb, K., Miettinen, K., and Słowiński, R., eds. (2008). *Multiobjective Optimization: Interactive and Evolutionary Approaches*, Lecture Notes in Computer Science 5252. Springer, Berlin.

Breiman, L. (2001), Random Forests. *Machine Learning*, 45(1), 5-32.

Busby, J.R. (1991). BIOCLIM – A bioclimate analysis and prediction system. In Margules, C.R. and Austin, M.P. (eds.), *Nature Conservation: Cost Effective Biological Surveys and Data Analysis*, 64–68. CSIRO, Melbourne.

Carpenter G., Gillison A.N., and Winter J. (1993). DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity and Conservation*, 2, 667–680.

Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E., and Yates, C.J. (2011) A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, 17, 43–57.

Franklin, J. (2010). *Mapping Species Distributions: Spatial Inference and Prediction*. Cambridge University Press.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25, 1965–1978.

Hijmans, R.J. and Elith, J. (2015). Species distribution modeling with R. R project documentation, available at cran.r-project.org/web/packages/dismo/vignettes/sdm.pdf

IUCN (2015). The IUCN Red List of Threatened Species, Version 2015.2, available at www.iucnredlist.org. Accessed 23 June 2015.

Liu, C., Berry, P.M., Dawson, T.P., and Pearson, R.G. (2005). Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, 28, 385–393.

Maindonald, J. and Braun, J. (2007). *Data Analysis and Graphics Using R: An Example-Based Approach*, 2nd edition. Cambridge University Press.

Pearson, R.G., Raxworthy, C.J., Nakamura, M., and Peterson, A.T. (2007) Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *Journal of Biogeography*, 34, 102–117.

Phillips, S.J., Anderson, R.P., and Schapire, R.E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3–4), 231–259.

Poyarkov, N.A., Rowley, J.J.L., Gogoleva, S.I., Vassilieva, A.B., Galoyan, E.A., and Orlov, N.L. (2015). A new species of *Leptolalax* (Anura: Megophryidae) from the western Langbian Plateau, southern Vietnam. *Zootaxa*, 3931(2), 221–252.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing (www.R-project.org), Vienna, Austria.

Raxworthy, C.J., Martinez-Meyer, E., Horning, N., Nussbaum, R.A., Schneider, G.E., Ortega-Huerta, M.A., and Peterson, T. (2003). Predicting distributions of known and unknown reptile species in Madagascar. *Nature*, 426, 837–841

Renner, I.W. and Warton, D.I. (2013). Equivalence of MAXENT and Poisson Point Process Models for Species Distribution Modeling in Ecology. *Biometrics*, 69, 274–281.

Rowley, J., Brown, R., Bain, R., Kusrini, M., Inger, R., Stuart, B., Wogan, G., Thy, N., Chan-ard, T., Trung, C.T., Diesmos, A., Iskandar, D.T., Lau, M., Ming, L.T., Makchai, S., Truong, N.Q., and Phimmachak, S. (2010a). Impending conservation crisis for Southeast Asian amphibians. *Biology Letters*, 6(3), 336–338.

Rowley, J.J.L., Stuart, B.L., Neang, T., and Emmett, D.A. (2010b). A new species of *Leptolalax* (Anura: Megophryidae) from northeastern Cambodia. *Zootaxa*, 2567, 57–68.

Rowley, J.J.L., Le, D.T.T., Tran, D.T.A., and Hoang, H.D. (2011). A new species of *Leptolalax* (Anura: Megophryidae) from southern Vietnam. *Zootaxa*, 2796, 15–28.

Rowley, J.J.L., Tran, D.T.A., Frankham, G.J., Dekker, A.H., Le, D.T.T., Nguyen, T.Q., Dau, V.Q., and Hoang, H.D. (2015). Undiagnosed cryptic diversity in small, microendemic frogs (*Leptolalax*) from the Central Highlands of Vietnam. *PLOS ONE*, 10(5), May 28, DOI: 10.1371/journal.pone.0128382.

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A., and NCEAS Predicting Species Distributions Working Group (2008). Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, 14, 763–773.