



Does GARP really fail miserably? A response to Stockman *et al.* (2006)

Kristina M. McNyset* and Jason K. Blackburn

World Health Organization Collaborating
Center for Remote Sensing and GIS for Public
Health, Department of Geography and
Anthropology, Louisiana State University,
Baton Rouge, Louisiana 70803, USA

ABSTRACT

Stockman *et al.* (2006) found that ecological niche models built using DesktopGARP ‘failed miserably’ to predict trapdoor spider (genus *Promyrmekiaphila*) distributions in California. This apparent failure of GARP (Genetic Algorithm for Rule-Set Production) was actually a failure of the authors’ methods, that is, attempting to build ecological niche models using single data points. In this paper, we present a re-analysis of their original data using standard methods with the data appropriately partitioned into training/testing subsets. This re-evaluation generated accurate distributional predictions that we contrast with theirs. We address the consequences of model-building using single data points and the need for a foundational understanding of the principles of ecological niche modelling.

*Correspondence: Kristina M. McNyset, USEPA
— ORD/WED, 200 SW 35th St. Corvallis, OR
97333, USA; E-mail: mcnyset.kristina@epa.gov

Keywords

GARP, predictive ecological niche modelling.

INTRODUCTION

In their recent paper, Stockman *et al.* (2006) utilize the DesktopGARP (DG) ecological niche-modelling approach to predict potential distributions of trapdoor spiders (genus *Promyrmekiaphila*) in north-central California. They found that GARP (Genetic Algorithm for Rule-Set Production) models ‘failed miserably’ in their ability to accurately predict *Promyrmekiaphila* presence or absence and concluded from this failure that DG should not be used to predict distributions of non-vagile species. We contend that it was the authors’ apparent ignorance of not only the use of DG as a modelling tool, but also of the principles of ecological niche modelling that led to the generation of poorly predicting models, and not a failure of DG. They used what we would consider highly unusual, and indefensible, methods to generate their models using DG. Consequently, the conclusions they draw and comparisons they make are invalid and promote the misuse of a readily available modelling tool. To illustrate this, we utilized their data points in a re-analysis using standard methods and produced results that will be contrasted with theirs.

Complete descriptions of appropriate methods using DG are available elsewhere (Peterson & Cohoon, 1999; Peterson, 2001; Anderson *et al.*, 2002, 2003; Wiley *et al.*, 2003; among others); however, a brief overview here is appropriate. If the default settings are left unchanged, DG will subset individual taxon data into 50/50 train/test subsets for each model generated. As there is a stochastic element to the GARP model-building process, there

is no unique solution and a number of models are usually generated from which a best-subset is chosen (Anderson *et al.*, 2003). If the ‘Best Subsets Selection Parameters’ option is selected, DG will generate models until it finds 20 with no more than 10% ‘extrinsic’ omission, or until it reaches the maximum number of runs specified (also defaulted to 20, though usually changed by the user to a more appropriate number (such as 200, for example), allowing DG to generate as many models as necessary). This extrinsic omission is calculated on the 50% testing subset. Once 20 low-omission models are found, a median commission percentage is calculated across those 20 models. The 10 (50%) low-omission models that have individual commission percentages closest to that median are set aside as the best subset. All of these numbers and thresholds may be changed by the user.

Stockman *et al.* (2006) gathered 42 occurrence records for *Promyrmekiaphila* from museum and private collection records. They then used these data points *individually* to build 20 models each using DG (for a total of 840), used the best subsets procedure to choose 10 of those models per data point, then summed the remaining 420 models to generate their final ‘best’ subset. We know of no literature supporting the construction of ecological niche models using a single data point. Indeed, there is a substantial literature investigating the minimum and optimal number of data points necessary to build predictive models. Across modelling techniques, estimates range as high as 250 points necessary to achieve accuracy, for example, using generalized additive model (GAM) or generalized linear model (GLM) (Pearce & Ferrier, 2000). However, Stockwell & Peterson (2002) found a

rapid increase in accuracy using DG as the number of points included in model-building climbed towards 20, with diminishing returns thereafter until near-maximal accuracies were reached with 50 points.

Beyond the supporting literature, we object on first principles to building ecological niche models based on single data points. In this kind of analysis, an ecological niche is generally conceptualized as the combination of ecological parameters within which a species can maintain populations without immigration (Grinnell, 1917), which can be understood as the 'mean phenotype of a population' (Holt & Gaines, 1992). From a practical perspective, we assume that the occurrence points included in niche-modelling analyses represent a sufficient sample of the environmental space occupied by a species. Species tend to occur within ranges of environmental variables (e.g. between -2 °C and 8 °C minimum temperature in January), not at single values. It is impossible for a single point to represent these ranges. This violation of first principles is even more egregious in the case of Stockman *et al.* (2006) as their data represent a genus, not just a single species.

Two of the rule types included in a GARP rule-set represent ranges (range rules and negated range rules; Stockwell & Peters, 1999). It is not surprising, given the constraints Stockman *et al.* (2006) put on the data, that DG was unable to generate accurate predictive models. We reran Stockman *et al.*'s (2006) analyses, mimicking their DG input settings with the exception of including all 42 data points. We used the same environmental coverages they included, though resampled to a spatial resolution of 0.01 degrees (~ 1 km \times 1 km pixel size). We felt this was the best compromise between the relatively coarse nature of some of the coverages, given that source data for the vegetation and soil classes were 1 : 2,000,000 scale vector maps, and the rest of the variables. The 30-m resampling in Stockman *et al.* (2006) is inappropriately fine-scale. We left the training/test proportion at the default 50%, which meant that for each model-generating iteration, DG randomly subset the data into 21 training and 21 testing subsets. This allowed use of a testing subset to evaluate model quality at each iteration, generating model-quality metrics, including the 'extrinsic' omission measure, used in the best subsets selection step, a procedure rendered meaningless in Stockman *et al.*'s (2006) process, as it is impossible to subset a single data point. We generated 840 models and let DG select a 420-model best subset. Though this is a large number of best-subset models, we wanted to be able to make a direct comparison to Stockman *et al.*'s (2006) results, though upon evaluation, the distributional prediction from this best-subset were not substantially different from predictions made from analyses we ran choosing a 10-model best-subset. The resulting predicted distribution is shown in Fig. 1.

Our results stand in stark contrast to those found by Stockman *et al.* (2006). Using the 80 occurrence points from collections made by the authors in 2005 as an independent validation set, our models were quite accurate. It is not clear how Stockman *et al.* (2006) calculated omission, so we are reporting two omission values. We found 0% total omission for the model set (meaning that all points were predicted by at least one model), and 9% average omission, calculated as the average omission for all the data points across all models. Stockman *et al.* (2006) also used

those 80 points to generate another best-model set, again using each data point individually. We also used those 80 points in a 50/50 training/test subset to generate 1600 models, 800 of which were chosen for the best subset. Those results are shown in Fig. 2. This time we used the original 42 museum records as independent validation data and found 0% total omission and 1% average omission. The resulting model sets from both analyses are very similar.

Although the predicted distributions we generated using appropriate methods are more accurate than those generated by Stockman *et al.* (2006), this does not mean that these are the *best* models possible. Other aspects of Stockman *et al.*'s (2006) methods may be impinging on model quality. For example, we used the same environmental coverages they used for comparative purposes even though it contains two categorical variables (soil texture type and vegetation class). Stockman *et al.* (2006) make the point that variables must be, 'prudently chosen'. We agree, and in general, prefer to include continuous measures of variables rather than categorical classes for a number of reasons. Categorical variables usually involve some post-processing of underlying continuous data, and cut-offs between categories can be arbitrarily decided, and including categorical variables with many categories can lead to their inclusion in misleading range rules. It may be that a different environmental coverage set would improve model quality (Elith *et al.*, 2006). However, our results sufficiently illustrate the critical flaw in their methods — building models using single presence points.

Stockman *et al.* (2006) did not report omission and commission for their BIOCLIM and GLM results, so it is difficult for us to compare our GARP results to those models, though visually the GARP models appear to be more accurate. Additionally, it is difficult to compare results as the authors opted to develop three different models using three different combinations of environmental coverages. As Stockman *et al.* (2006) pointed out, this lack of control between results makes it difficult to determine if it was any particular modelling approach or combination of environmental variables that lead to relative differences in inter-approach predictive power. While this may not totally negate their comparisons, it certainly confounds interpretation of model quality from one approach to another. Also, they used all the available data points to build their BIOCLIM and GLM models, making an invalid comparison to GARP models built on single points.

Because there is a lack of consistency in both what environmental variables and how many data points are included in each different modelling approach, any cross-approach model comparison is inappropriate. It is worth noting that while Stockman *et al.* (2006) applied the kappa statistic, the χ^2 statistic, and the area under the curve (AUC) from a receiver operating characteristic analysis, multiple authors have reported that kappa and χ^2 are sensitive to sample size (Fielding & Bell, 1997; Anderson *et al.*, 2003) and that χ^2 can have high significance in models with unacceptable omission rates (Anderson *et al.*, 2003), and the authors did not acknowledge these potential errors in their presentation of either kappa and χ^2 .

In general, we welcome more analyses along the lines of what Stockman *et al.* (2006) purported to be. That is, it is important to

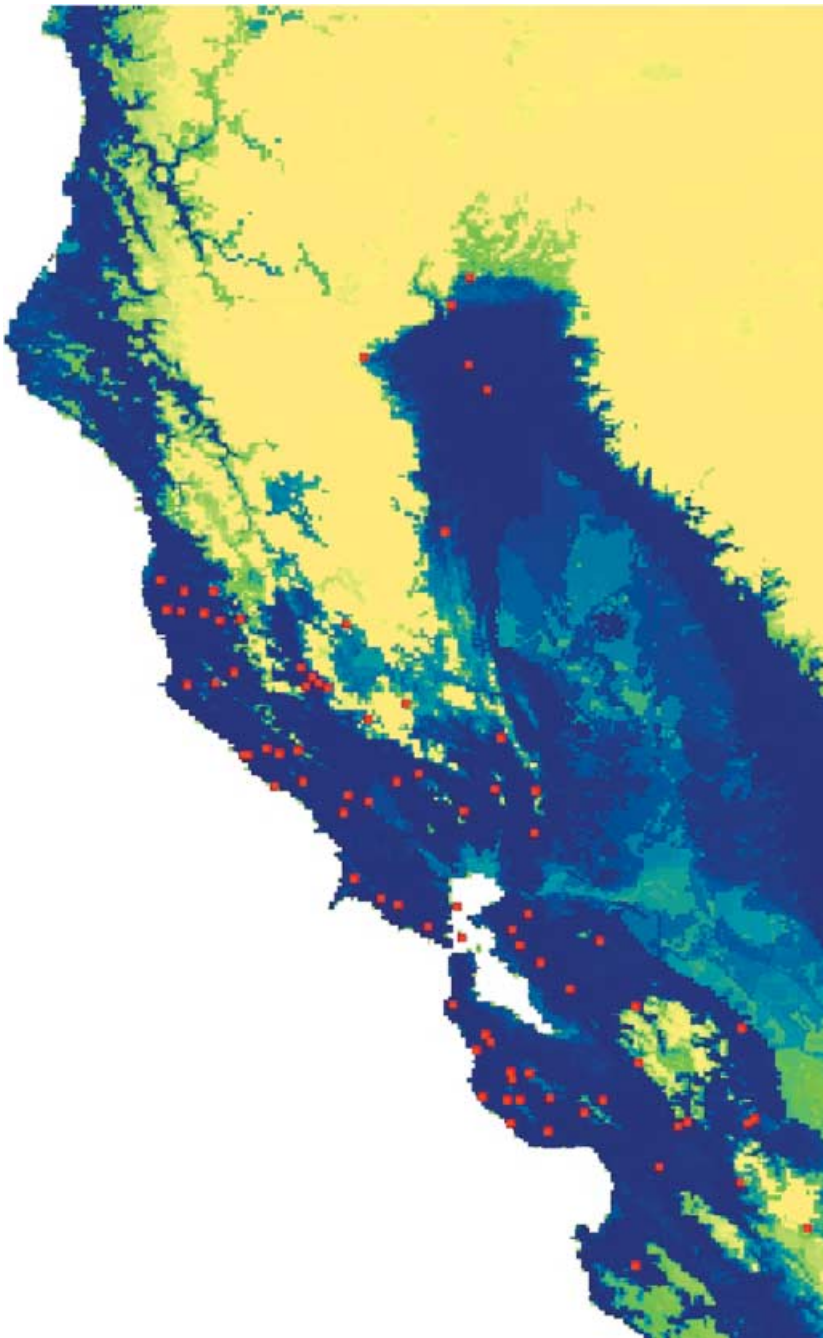


Figure 1 Distributional predictions of *Promyrmekiaphila* by 420 GARP models built using 42 museum collection records. The colour gradient from green to dark blue indicates increasing model-agreement of predicted presence. Red squares represent 80 new collection records made by Stockman *et al.* (2006) from March–June 2005 used as an independent validation data set.

see more field tests of predictive models. Also, the question of whether different modelling approaches are equally effective in predicting distributions of non-vagile species deserves consideration, though our results would indicate that for these spiders, GARP models are quite accurate. Additionally, Para-Olea *et al.* (2005) and Adjemian *et al.* (2006) found GARP models to be accurate when predicting distributions of other low-vagility species (salamanders and several flea species, respectively), though more tests of these methods with different organisms would be informative. However, Stockman *et al.*'s (2006) study is not a fair test of these methods, and only serves to cloud the literature. We

contend that Stockman *et al.*'s (2006) statement that DG is a 'black box' is more indicative of their misuse of the method and misconceptions about the methodology, rather than an accurate statement about DG. While genetic algorithms, by design, tend to produce complex rule-sets, these are not impossible to interpret (Rice *et al.*, 2003), nor are appropriate accuracy metrics difficult to derive and interpret. We hope that in the future, researchers will be aware of the points raised here when conducting predictive analyses, and cautious of literature that lacks a solid foundation in the fundamental principles of ecological niche modelling.

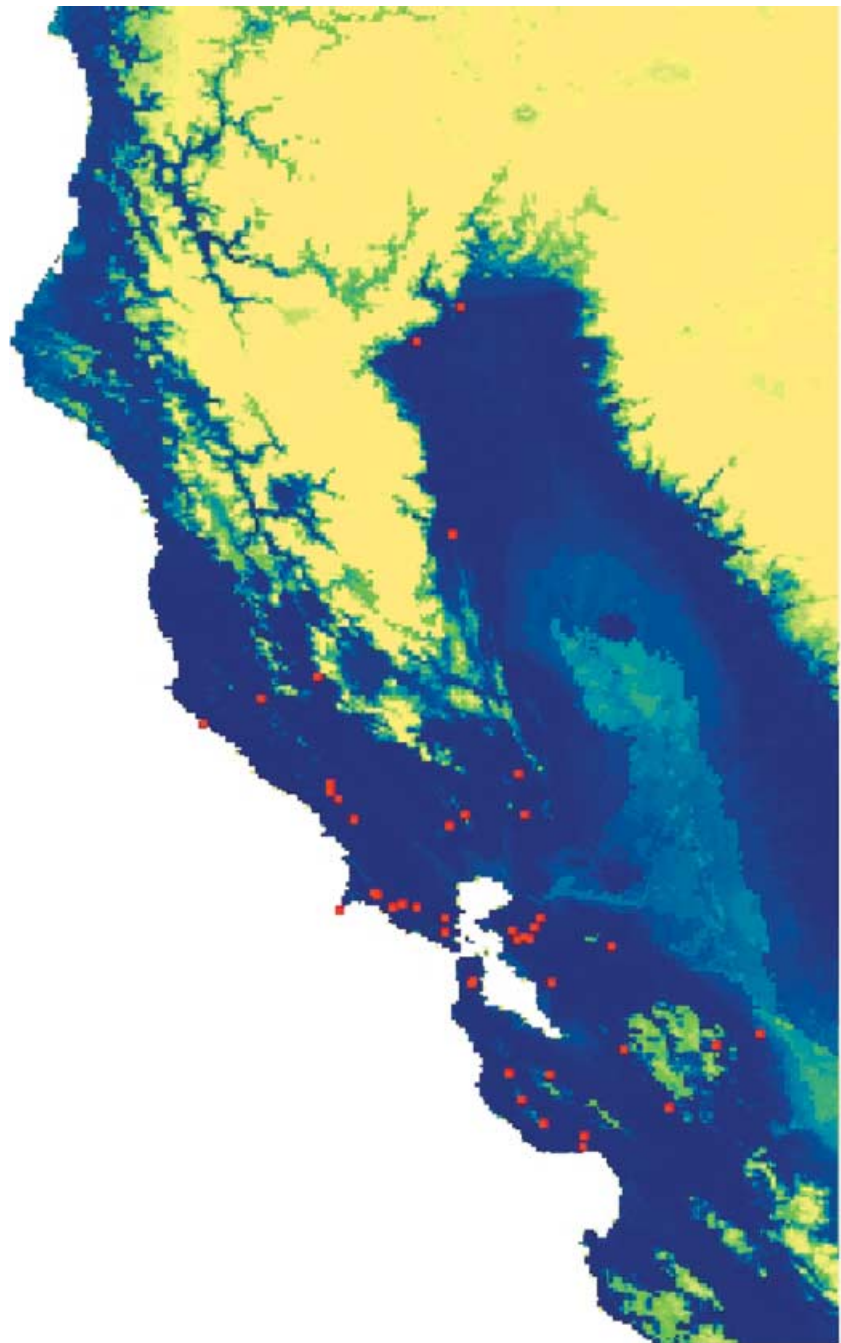


Figure 2 Distributional predictions of *Promyrmekiaphila* by 800 GARP models built using 80 collection localities (collected by Stockman *et al.* (2006) in March–June 2005). The colour gradient from green to dark blue indicates increasing model-agreement of predicted presence. Red squares represent 42 museum collection records used as an independent validation data set.

REFERENCES

- Adjemian, J.C.Z., Girvetz, E.H., Beckett, L. & Foley, J.E. (2006) Analysis of Genetic Algorithm for Rule-Set Production (GARP) modeling approach for predicting distributions of fleas implicated as vectors of plague, *Yersinia pestis*, in California. *Journal of Medical Entomology*, **43**, 93–103.
- Anderson, R.P., Gomez-Laverde, M. & Peterson, A.T. (2002) Geographical distributions of spiny pocket mice in South America: insights from predictive models. *Global Ecology and Biogeography*, **11**, 131–141.
- Anderson, R.P., Lew, D. & Peterson, A.T. (2003) Evaluating predictive models of species' distributions: criteria for selecting optimal models. *Ecological Modelling*, **162**, 211–232.
- Elith, J. & Graham, C.H. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.
- Fielding, A.H. & Bell, J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24**, 38–49.
- Grinnell, J. (1917) Field tests of theories concerning distributional control. *The American Naturalist*, **51**, 115–128.

- Holt, R.D. & Gaines, M.S. (1992) Analysis of adaptation in heterogeneous landscapes: implications for the evolution of fundamental niches. *Evolutionary Ecology*, **6**, 433–447.
- Para-Olea, G., Martínez-Meyer, E. & Pérez-Ponce de León, G. (2005) Forecasting climate change effects on salamander distributions in the Highlands of Central Mexico.
- Pearce, J. & Ferrier, S. (2000) An evaluation of alternative algorithms for fitting species distribution models using logistic regression. *Ecological Modelling*, **128**, 127–147.
- Peterson, A.T. (2001) Predicting species' geographic distributions based on ecological niche modeling. *Condor*, **103**, 599–605.
- Peterson, A.T. & Cohoon, K.P. (1999) Sensitivity of distributional prediction algorithms to geographic completeness. *Ecological Modelling*, **117**, 159–164.
- Rice, N.H., Martínez-Meyer & Enrique, Peterson, A.T. (2003) Ecological niche differentiation in the *Aphelocoma* jays: a phylogenetic perspective. *Biological Journal of the Linnean Society*, **80**, 369–383.
- Stockman, A.K., Beamer, D.A. & Bond, D.A. (2006) An evaluation of a GARP model as an approach to predicting the spatial distribution of non-vagile invertebrate species. *Diversity and Distributions*, **12**, 81–89.
- Stockwell, D.R.B. & Peters, D. (1999) The GARP modeling system: problems and solutions to automated spatial predictions. *International Journal of Geographical Information Science*, **13**, 143–158.
- Stockwell, D.R.B. & Peterson, A.T. (2002) Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, **148**, 1–13.
- Wiley, E.O., McNyset, K.M., Peterson, A.T., Robins, C.R. & Stewart, A.M. (2003) Niche modeling and geographic range predictions in the marine environment using a machine-learning algorithm. *Oceanography*, **16**, 120–127.