

Highlights

Climate predictor selection and variable reduction methods influence MaxEnt model performance and predictions of climatic suitability

Clarke van Steenderen, Guy F. Sutton

- Highlight 1
- Highlight 2
- Highlight 3

Climate predictor selection and variable reduction methods influence MaxEnt model performance and predictions of climatic suitability

Clarke van Steenderen^a, Guy F. Sutton^{a,*}

^a*Center for Biological Control, Department of Zoology and Entomology, Rhodes University, Makhanda, 6140*

Abstract

This is the abstract. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vestibulum augue turpis, dictum non malesuada a, volutpat eget velit. Nam placerat turpis purus, eu tristique ex tincidunt et. Mauris sed augue eget turpis ultrices tincidunt. Sed et mi in leo porta egestas. Aliquam non laoreet velit. Nunc quis ex vitae eros aliquet auctor nec ac libero. Duis laoreet sapien eu mi luctus, in bibendum leo molestie. Sed hendrerit diam diam, ac dapibus nisl volutpat vitae. Aliquam bibendum varius libero, eu efficitur justo rutrum at. Sed at tempus elit.

Keywords: keyword1, keyword2

1. 1. Introduction

Ecological models are important tools to aid the development and implementation of environmental policies and management programmes (Addison et al., 2013; Schuwirth et al., 2019; Sutton and Martin, 2022). These models are used for conservation planning (Guisan et al., 2013), predicting the establishment and spread of invasive species (Martin et al., 2020), implementing biological control programmes (Sutton, 2019; Mukherjee et al., 2021), and forecasting species responses to environmental change (Bocedi et al., 2014), amongst other applications. Species distribution models (SDM's) are an example of ecological models that have become increasingly popular in recent years (Elith and Leathwick, 2009). SDM's typically take the form of correlative or mechanistic models that correlate species presence/absences (or pseudo-absences) to environmental covariates to identify suitable climatic conditions for the study taxon (Elith et al., 2011). The Maximum Entropy species distribution model (hereafter 'MaxEnt') is amongst the most popular methods for climate modelling studies and has been shown to perform well compared to alternative modelling algorithms (Wisz et al., 2008; Phillips et al.,

2017). It uses maximum entropy to distinguish between environmental conditions where the focal taxon is present from environmental conditions at sites without confirmed presence records for the taxon (Elith et al., 2011).

In recent years, a number of studies have investigated and demonstrated that computational choices made during the model building process can have a significant influence on resulting model outputs and inferences drawn (Warren and Seifert, 2011; Webber et al., 2011; Shcheglovitova and Anderson, 2013; Boria et al., 2017; Sutton and Martin, 2022). Despite its importance in the model building process, covariate selection methods have received considerably little attention to date (but see Austin and Van Niel, 2011; Fourcade et al., 2018; Adde et al., 2023), whereby covariate selection refers to "identify[ing] the best subset of covariates out of a panel of many candidates, both from an ecological and statistical perspective (see Adde et al., 2023, and references therein).

Next paragraphs

- Review current methods used to select variables, e.g. R², VIF, PCA, expert opinion, and most recently, automated selection / model-based selection.
- Aims of this paper are to show how variable selection can influence model outputs and in-

*Corresponding author

Email addresses: vsteenderen@gmail.com (Clarke van Steenderen), g.sutton@ru.ac.za (Guy F. Sutton)

ferences.

2. 2. Methods and materials

2.0.1. 2.1. Species occurrence records

2.0.2. 2.2. Environmental predictors

Climate data were obtained by downloading the standard set of 19 bioclimatic variables from the WorldClim ver. 2.1 database (Fick and Hijmans, 2017) (data available at: www.worldclim.org/download.html). This dataset is representative of annual and seasonal means and variation of temperature and precipitation metrics averaged over the 1950–2000 time period (current climate) at a 2.5 min resolution. These variables have been shown to effectively model the climatic suitability for non-native insects (e.g., Trethowan et al., 2011).

2.0.3. 2.3. Model calibration

MaxEnt (ver. 3.4.3) was implemented in the ‘dismo’ R package (Hijmans et al., 2021).

Given that MaxEnt is a presence/pseudo-absence modelling algorithm, model calibration requires a user-defined geographic background to sample the climate of representative grid cells where the focal species is assumed to be absent (i.e., background points or pseudo-absences). Background definition can have a significant effect on model output (VanDerWal et al., 2009). The background should ideally represent the geographic areas available to the focal species, omitting areas where species absence is due to historical factors, dispersal constraints and/or biotic interactions (Sanin and Anderson, 2018). Following Webber et al. (2011), we defined the model background using the Koppen-Geiger climate classification system (available at: <http://koeppen-geiger.vu-wien.ac.at>). Only Koppen-Geiger climate zones that contained at least one native-range occurrence record for *D. rubiformis* in Australia were used as the background area from which background points were drawn for model calibration (Fig. 1a). The Koppen-Geiger climate zones were intersected with the occurrence records using the ‘raster’ R package (Hijmans, 2022). We randomly sampled 10 000 points (the default number used for Maxent; Merow et al. (2013)) from within this background definition using the ‘dismo’ R package (Hijmans et al., 2021).

MaxEnt models were parameterised with default settings for multiple parameters, including: convergence = 105, maximum number of iterations = 500 and prevalence = 0.5. The ‘fade by clamping’ option was selected to prevent extrapolation well outside the range of climatic values in the model training area (Philips et al., 2017). Model predictions were obtained using the ‘logistic output’ to create continuous climatic suitability raster layers scaled between 0 (climatically unsuitable) and 1 (climatically suitable).

2.0.4. 2.4. Model evaluation

Model tuning experiments were applied to the native-range MaxEnt models to derive within-sample evaluation metrics to guide the selection of optimal MaxEnt parameter configurations (feature classes and regularisation multipliers). Optimised parameter configurations would then be used to re-fit the MaxEnt models before being projected into a novel geographic region and making projections of climatic suitability for *D. rubiformis*. Model tuning was performed by building MaxEnt models with varying (1) feature class combinations (H = Hinge only, L = Linear only, LQ = Linear and Quadratic and LQH = Linear, Quadratic and Hinge features) and (2) regularisation multipliers (1:8). In total, 32 MaxEnt models were specified. Native-range model performance and optimal parameter configurations were assessed using 4-fold spatial block cross validation using ‘ENMeval’ (Kass et al., 2021).

Optimal parameter configurations were assessed using multiple metrics that reflect different aspects of model performance. Four metrics were calculated, including: (1) discriminatory ability (AUCtest), (2) overfitting (AUCdiff), (3) omission rates (OR10), and (4) overall parsimony (AICc). The use of AUC analyses for assessing the fit of MaxEnt models has been criticised for a variety of reasons (see Lobo et al., 2008; Peterson et al., 2008). However, AUC metrics are arguably the most widely used metrics to evaluate MaxEnt model performance, and as such, we believe it is important to include them in our evaluation, and contrast the results obtained using AUC versus other metrics.

We specified five final MaxEnt models, four models calibrated with FC and RM values that optimised model performance based on the metrics discussed below, and a MaxEnt model calibrated with default FC and RM values. Our intention was to compare MaxEnt model predictions and perfor

mance depending on which metric was used to select optimal parameter configurations relative to the default MaxEnt settings.

- (1) AUC test assesses the model's ability to discriminate between predicted presence at withheld portions of the data used to test the model versus pseudo-absence points. An AUC of less than 0.8 is considered a poor model, between 0.8 and 0.9 is a fair model, between 0.9 and 0.995 a good model, and > 0.995 an excellent model (Fielding and Bell, 1997). Thus, higher AUC test values indicate increased ability to discriminate between testing and background points.
- (2) AUC diff is the difference between AUC values calculated on training points only (AUC_{train}) and AUC_{test} [see (1) AUC_{test} above for details] (Warren and Seifert, 2011). Thus, higher AUC diff values indicate whether the MaxEnt model is overfit on the training data, and thus, may perform poorly when evaluated against testing points.
- (3) OR 10 is the 10% training omission rate (Boria et al., 2014). Overfit models have omission rates higher than the theoretical expectation for the threshold applied (Shcheglovitova and Anderson, 2013). As such, the OR 10 criterion selected models calibrated with MaxEnt settings which best approximated the expected 0.10 omission rate. Models with omission rates increasingly higher than the expected value were considered to have a higher degree of overfit (Boria et al., 2017).
- (4) The Akaike Information Criterion corrected for small sample sizes (AICc) criterion simultaneously scores models according to their complexity and goodness-of-fit. AICc was used as the primary evaluation metric as it is calculated using MaxEnt models built using the entire species occurrence dataset (i.e. all the occurrence points in the native-range), unlike AUC and OR10 (and numerous other metrics frequently used for model evaluation) which may be spatially biased due to the partitioning of the species occurrence dataset into training and evaluation sets (Sanin and Anderson, 2018). Optimal parameter configurations were determined by selecting model configurations which produced the lowest value for AICc (i.e., AICc=0; following Kass et al. (2021)).

2.0.5. 2.5. Model visualisation

- Need to discuss how we map the different rasters, and maybe how we quantified the difference in suitability projections between climate layers?

All modelling and statistical analyses were conducted in R ver. 4.0.3 (R Core Team, 2020). All values presented in text are presented as mean \pm standard error, unless otherwise stated. A standardised ODMAP methods protocol (Overview, Data, Model, Assessment, and Prediction) has been completed for this study and can be found in Supplementary File S1. ODMAP standardises the reporting of SDM modelling studies to improve transparency and reproducibility (Zurell et al., 2020).

3. 3. Results

4. 4. Discussion

5. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

6. Data availability

All data and code required to reproduce the analyses are available in a public GitHub repository: https://github.com/guysutton/MS_climate_variable_selection_

7. Acknowledgements

CVS and GFS acknowledge funding from the South African Working for Water (WfW) programme of the Department of Forestry, Fisheries and the Environment: Natural Resource Management Programmes (DFFE: NRMP). Funding was also provided by the South African Research Chairs Initiative of the Department of Science and Technology and the National Research Foundation (NRF) of South Africa. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and the NRF does not accept any liability in this regard.

8. Supplementary materials

References

- Adde, A., Rey, P.L., Fopp, F., Petitpierre, B., Schweiger, A.K., Broennimann, O., Lehmann, A., Zimmermann, N.E., Altermatt, F., Pellissier, L., Guisan, A., 2023. Too many candidates: Embedded covariate selection procedure for species distribution modelling with the covsel R package. *Ecological Informatics* 75, 102080. doi:[10.1016/j.ecoinf.2023.102080](https://doi.org/10.1016/j.ecoinf.2023.102080).
- Addison, P.F.E., Rumpff, L., Bau, S.S., Carey, J.M., Chee, Y.E., Jarrad, F.C., McBride, M.F., Burgman, M.A., 2013. Practical solutions for making models indispensable in conservation decision-making. *Diversity and Distributions* 19, 490–502. doi:[10.1111/ddi.12054](https://doi.org/10.1111/ddi.12054).
- Austin, M.P., Van Niel, K.P., 2011. Improving species distribution models for climate change studies: Variable selection and scale. *Journal of Biogeography* 38, 1–8. doi:[10.1111/j.1365-2699.2010.02416.x](https://doi.org/10.1111/j.1365-2699.2010.02416.x).
- Bocedi, G., Palmer, S.C., Pe'er, G., Heikkinen, R.K., Matsinos, Y.G., Watts, K., Travis, J.M., 2014. RangeShifter: A platform for modelling spatial eco-evolutionary dynamics and species' responses to environmental changes. *Methods in Ecology and Evolution* 5, 388–396. doi:[10.1111/2041-210X.12162](https://doi.org/10.1111/2041-210X.12162).
- Boria, R.A., Olson, L.E., Goodman, S.M., Anderson, R.P., 2017. A single-algorithm ensemble approach to estimating suitability and uncertainty: Cross-time projections for four Malagasy tenrecs. *Diversity and Distributions* 23, 196–208. doi:[10.1111/ddi.12510](https://doi.org/10.1111/ddi.12510).
- Elith, J., Leathwick, J., 2009. Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution and Systematics* 40, 677–697. doi:[10.1146/annurev.ecolsys.110308.120159](https://doi.org/10.1146/annurev.ecolsys.110308.120159).
- Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E., Yates, C.J., 2011. A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions* 17, 43–57. doi:[10.1111/j.1472-4642.2010.00725.x](https://doi.org/10.1111/j.1472-4642.2010.00725.x).
- Fourcade, Y., Besnard, A.G., Secondi, J., 2018. Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. *Global Ecology and Biogeography* 27, 245–256. doi:[10.1111/geb.12684](https://doi.org/10.1111/geb.12684).
- Guisan, A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I.T., Regan, T.J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., Martin, T.G., Rhodes, J.R., Maggini, R., Setterfield, S.A., Elith, J., Schwartz, M.W., Wintle, B.A., Broennimann, O., Austin, M., Ferrier, S., Kearney, M.R., Possingham, H.P., Buckley, Y.M., 2013. Predicting species distributions for conservation decisions. *Ecology Letters* 16, 1424–1435. doi:[10.1111/ele.12189](https://doi.org/10.1111/ele.12189).
- Martin, G.D., Magengelele, N.L., Paterson, I.D., Sutton, G.F., 2020. Climate modelling suggests a review of the legal status of Brazilian pepper *Schinus terebinthifolia* in South Africa is required. *South African Journal of Botany* 132, 95–102. doi:[10.1016/j.sajb.2020.04.019](https://doi.org/10.1016/j.sajb.2020.04.019).
- Mukherjee, A., Banerjee, A.K., Raghu, S., 2021. Biological control of *Parkinsonia aculeata*: Using species distribution models to refine agent surveys and releases. *Biological Control* 159, 104630. doi:[10.1016/j.biocontrol.2021.104630](https://doi.org/10.1016/j.biocontrol.2021.104630).
- Phillips, S.J., Anderson, R.P., Dudík, M., Schapire, R.E., Blair, M.E., 2017. Opening the black box: An open-source release of Maxent. *Ecography* 40, 887–893. doi:[10.1111/ecog.03049](https://doi.org/10.1111/ecog.03049).
- Schuwirth, N., Borgwardt, F., Domisch, S., Friedrichs, M., Kattwinkel, M., Kneis, D., Kuemmerlen, M., Langhans, S.D., Martínez-López, J., Vermeiren, P., 2019. How to make ecological models useful for environmental management. *Ecological Modelling* 411, 108784. doi:[10.1016/j.ecolmodel.2019.108784](https://doi.org/10.1016/j.ecolmodel.2019.108784).
- Shcheglovitova, M., Anderson, R.P., 2013. Estimating optimal complexity for ecological niche models: A jackknife approach for species with small sample sizes. *Ecological Modelling* 269, 9–17. doi:[10.1016/j.ecolmodel.2013.08.011](https://doi.org/10.1016/j.ecolmodel.2013.08.011).
- Sutton, G.F., 2019. Searching for a needle in a haystack: Where to survey for climatically-matched biological control agents for two grasses (*Sporobolus* spp.) invading Australia. *Biological Control* 129, 37–44. doi:[10.1016/j.biocontrol.2018.11.012](https://doi.org/10.1016/j.biocontrol.2018.11.012).
- Sutton, G.F., Martin, G.D., 2022. Testing MaxEnt model performance in a novel geographic region using an intentionally introduced insect. *Ecological Modelling* 473, 110139. doi:[10.1016/j.ecolmodel.2022.110139](https://doi.org/10.1016/j.ecolmodel.2022.110139).
- Warren, D.L., Seifert, S.N., 2011. Ecological niche modeling in Maxent: The importance of model complexity and the performance of model selection criteria. *Ecological Applications* 21, 335–342. doi:[10.1890/10-1171.1](https://doi.org/10.1890/10-1171.1).
- Webber, B.L., Yates, C.J., Le Maitre, D.C., Scott, J.K., Kriticos, D.J., Ota, N., McNeill, A., Le Roux, J.J., Midgley, G.F., 2011. Modelling horses for novel climate courses: Insights from projecting potential distributions of native and alien Australian acacias with correlative and mechanistic models. *Diversity and Distributions* 17, 978–1000. doi:[10.1111/j.1472-4642.2011.00811.x](https://doi.org/10.1111/j.1472-4642.2011.00811.x).
- Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A., Group, N.P.S.D.W., 2008. Effects of sample size on the performance of species distribution models. *Diversity and Distributions* 14, 763–773. doi:[10.1111/j.1472-4642.2008.00482.x](https://doi.org/10.1111/j.1472-4642.2008.00482.x).