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: Species distribution model, Variable selection, Ecological model, WORLDCLIM

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).

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

The methods used for selecting covariates for SDM studies typically fall under four categories: (1) creating a set of uncorrelated covariates from a larger expert-chosen group of 15-30 covariates, using correlation analyses and/or variance inflation factors (VIF's) (Dormann et al., 2013), (2) using principal component analysis (PCA) to reduce dimensionality of a set of covariates (Júnior and Nóbrega, 2018), (3) using expert-opinion to select covariates that may be relevant based on the knowledge of the focal species biology, physiology and ecological requirements (Petitpierre et al., 2017; Scherrer and Guisan, 2019), and more recently (4) automated statistical model-based covariate selection techniques, such as '*embedded methods*', e.g. RIDGE and LASSO (Guisan et al., 2002; Guyon and Elisseeff, 2003; Saeys et al., 2007; Adde et al., 2023). For a more detailed overview of the different variable selection methods, we direct readers to several recent reviews on this topic (Fois et al., 2018; Fourcade et al., 2018; Melo-Merino et al., 2020).

While there are guidelines for how to implement these methods individually, e.g. using an $|r| > 0.70$ or a VIF $>$ 2-5 as a threshold for removing significantly correlated predictors (Dormann et al., 2013), there is currently no consensus on which method to use for a particular study (but see Adde et al. (2023)). As such, we aimed to demonstrate the variation in climatic suitability raster projections and predictive accuracy of MaxEnt models calibrated using the different covariate selection methods, and discuss the potential implications for model inference and possible management programmes based on these inferences. To demonstrate, we modelled the potential climatic suitability (i.e. the climatic similarity relative to the climate space occupied in the model calibration region) for the invasive insect pest *Diaphorina citri* Kuwayama (Asian citrus psyllid) (Hemiptera: Psyllidae), that is native to Asia but has become invasive in many non-native regions around the world, such as the USA, Brazil and Africa (**Insert citations**, **Insert a map of global distribution?**).

- Clarke: Can you please insert 1-2 sentences about the impact of ACP and its association with HLB?

2. Methods and materials

2.1. Species occurrence records

- Clarke to add information on where we sourced GPS records - link to key papers, maybe a supplementary table (if required).
- How many records do we have?

Spatial autocorrelation is an important factor that may affect model outputs. Filtering of species occurrence data may limit the inherent biases in the data and improve model quality (Veloz, 2009). To avoid pseudo-replication, only one occurrence record per 2.5 minute grid cell was used for model calibration. Species occurrence datasets were thinned using the '*spThin*' package (Aiello-Lammens et al., 2015), and spatial autocorrelation analyses were performed using the '*ecospat*' package (Di Cola et al., 2017). Eighty-eight occurrence records were retained to calibrate native-range models. Spatial thinning was also performed for the invaded range occurrences, however, there was no evidence for spatial autocorrelation. As such, no occurrence records were thinned for the invaded-range dataset.

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 minute resolution. These variables have been shown to effectively model the climatic suitability for non-native insects (e.g., Sutton and Martin (2022)).

2.3. Model calibration

MaxEnt (ver. 3.4.3) was implemented in the '*dismo*' R package (**?**).

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 (Sanín and Anderson, 2018). We defined the model background using the Koppen-Geiger climate classification system (available at: http://koeppen-geiger.vu-wien.ac.at) (Webber et al., 2011). 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).

- Need to update the below with appropriate references to 'terra' package

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 specified 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 (Phillips 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).

2.4. 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 refit 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 metricsdiscussed 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 AUCtest 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 (AUCtrain) and AUCtest [see (1) AUCtest above for details] (Warren and Seifert, 2011). Thus, higher AUCdiff 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)).

2.5. 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. 4.3.0 (Team, 2023). All values presented in text are presented as mean ± 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. Results

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_sdm

7. Acknlowedgements

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, finding, 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.

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.

Aiello-Lammens, M.E., Boria, R.A., Radosavljevic, A., Vilela, B., Anderson, R.P., 2015. spThin: An R package for spatial thinning of species occurrence records for use in ecological niche models. Ecography 38, 541–545. doi:10.1111/ecog.01132.

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.

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.

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.

Di Cola, V., Broennimann, O., Petitpierre, B., Breiner, F.T., D'Amen, M., Randin, C., Engler, R., Pottier, J., Pio, D., Dubuis, A., Pellissier, L., Mateo, R.G., Hordijk, W., Salamin, N., Guisan, A., 2017. Ecospat: An R package to support spatial analyses and modeling of species niches and distributions. Ecography 40, 774–787. doi:10.1111/ecog.02671.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J., Münkemüller, T., McClean, C., Osborne, P.E., Reineking, B., Schröder, B., Skidmore, A.K., Zurell, D., Lautenbach, S., 2013. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. Ecography 36, 27–46. doi:10.1111/j.1600-0587.2012.07348.x.

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.

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.

Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology 37, 4302–4315. doi:10.1002/joc.5086.

Fois, M., Cuena-Lombraña, A., Fenu, G., Bacchetta, G., 2018. Using species distribution models at local scale to guide the search of poorly known species: Review, methodological issues and future directions. Ecological Modelling 385, 124–132. doi:10.1016/j.ecolmodel.2018.07.018.

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.

Guisan, A., Edwards, T.C., Hastie, T., 2002. Generalized linear and generalized additive models in studies of species distributions: Setting the scene. Ecological Modelling 157, 89–100. doi:10.1016/S0304-3800(02)00204-1.

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.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. Journal of machine learning research 3, 1157–1182.

Júnior, P.D.M., Nóbrega, C.C., 2018. Evaluating collinearity effects on species distribution models: An approach based on virtual species simulation. PLOS ONE 13, e0202403. doi:10.1371/journal.pone.0202403.

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.

Melo-Merino, S.M., Reyes-Bonilla, H., Lira-Noriega, A., 2020. Ecological niche models and species distribution models in marine environments: A literature review and spatial analysis of evidence. Ecological Modelling 415, 108837. doi:10.1016/j.ecolmodel.2019.108837.

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.

Petitpierre, B., Broennimann, O., Kueffer, C., Daehler, C., Guisan, A., 2017. Selecting predictors to maximize the transferability of species distribution models: Lessons from cross-continental plant invasions. Global Ecology and Biogeography 26, 275–287. doi:10.1111/geb.12530.

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.

Saeys, Y., Inza, I., Larrañaga, P., 2007. A review of feature selection techniques in bioinformatics. Bioinformatics 23, 2507–2517. doi:10.1093/bioinformatics/btm344.

Sanín, C., Anderson, R.P., 2018. A framework for simultaneous tests of abiotic, biotic, and historical drivers of species distributions: Empirical tests for North American Wood Warblers based on climate and pollen. The American Naturalist 192, E48–E61. doi:10.1086/697537.

Scherrer, D., Guisan, A., 2019. Ecological indicator values reveal missing predictors of species distributions. Scientific Reports 9, 3061. doi:10.1038/s41598-019-39133-1.

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.

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.

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.

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.

Team, R.C., 2023. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ .

VanDerWal, J., Shoo, L.P., Graham, C., Williams, S.E., 2009. Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? Ecological Modelling 220, 589–594. doi:10.1016/j.ecolmodel.2008.11.010.

Veloz, S.D., 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. Journal of Biogeography 36, 2290–2299. doi:10.1111/j.1365-2699.2009.02174.x.

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.

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.

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.