Modelling the habitat suitability of the Thylacine

Jessie C. Buettel[1], Damien A. Fordham[2], Sean Haythorne[1,2], Stuart C. Brown[1,3], Barry W. Brook[2]

[1]School of Natural Sciences and ARC Centre of Excellence for Australian Biodiversity and Heritage, University of Tasmania, Hobart, Tasmania 7001, Australia

[2]The Environment Institute and School of Biological Sciences, University of Adelaide, South Australia 5005, Australia.

[3]GLOBE Institute, University of Copenhagen, 1350, Copenhagen K, Denmark

Materials and Methods

Thylacine occurrence records

We collated occurrence records of *Thylacinus cynocephalus* from two sources:

a) Bounty records for the period 1888-1912 archived by the Tasmanian Government (Lands Tasmania). We geolocated harvest records based on the place of harvest in the bounty record.
b) Tasmanian Thylacine Sighting Records Database: which includes the location of verified kills or captures as well as expert sightings for the years (1913-1936) (Brook *et al.* 2021).

To reduce spatial autocorrelation, we removed all duplicates and thinned the occurrence data, ensuring a minimum distance of 4 km between observations, while at the same time retaining adequate sample size. This was done using the R function `spthin` (Aiello-Lammens *et al.* 2015), and resulted in total of 310 spatially independent records, which are available here: https://doi.org/10.25909/14751741.v1

Climate and environmental data

We accessed climatic data from the Biodiversity and Climate change Virtual Laboratory (BCCVL; bccvl.org.au) for Tasmania at a 30arcsec (~1 km) resolution. There are 19 possible climatic variables (30-year averages focused on 1990) available through BCCVL. We narrowed our selection of climate variables to six: mean diurnal range, isothermality, mean temperature in the wettest quarter, mean temperature in the driest quarter, seasonality in precipitation, and mean precipitation in the warmest quarter. This was done based on those variables that i) were identified as being potentially important correlates of Thylacine

occurrence; and ii) had correlation coefficients that did not exceed 0.7, and, therefore, were independent and not collinearly related (Dormann *et al.* 2013). We did not correct for climatic change during the 20th century prior to 1990, because it has been small in Tasmania, particularly during the first half of the 20th century (Scharples 2011).

Candidate environmental variables deemed to be potentially important correlates of Thylacine occurrence were: pre-European vegetation, distance to water, elevation, topographic roughness, and land-use. Spatial data on pre-European vegetation came from the National Vegetation Information System (NVIS; https://www.environment.gov.au/land/native-vegetation/national-vegetation-information-system), which we reclassified to four macro-vegetation groups: rainforest, tall eucalypt forest, woodland, and shrubland/grassland. We accessed surface water information from Geoscience Australia (https://www.ga.gov.au/scientific-topics/national-location-information/national-surface-water-information) and calculated the distance of each cell to the nearest lake or river. Elevation at 30arcsec was collected from the Australian Bureau of Agricultural and Resource Economics and Sciences (ABARES, agriculture.gov.au). Topographic roughness was calculated from NASA Shuttle Radar Topography Mission (SRTM) global elevation data (https://www2.jpl.nasa.gov/srtm/) using a 250m analysis window. Land-use data for Tasmania came from the Australian Bureau of Agricultural and Resource Economics and Sciences (ABARES; agriculture.gov.au). Land-use in 1990 (the oldest data available) was grouped into four predominant types: conservation/nature reserves, modified native, plantation and farming/urban. All environmental data was resampled to a 30arcsec resolution. Correlation coefficients between environmental variables was < 0.7.

Pseudoabsences

We generated pseudoabsences using a climatically- and geographically-stratified approach, also known as a 'two-step-pseudoabsence selection' (Senay, Worner & Ikeda 2013). To do this, a principal co-ordinate analysis (PCA) was used to differentiate between climatically suitable and unsuitable areas. Pseudoabsences were randomly sampled from climatically dissimilar areas at a ratio of 1:1 to the presence points (Barbet-Massin *et al.* 2012). The minimum distance between pseudoabsence points, and pseudoabsence and presence points, was 8 km.

Projecting habitat suitability

To account for inter-model variability in projections of probability of Thylacine occurrence (a proxy of habitat suitability; Guisan & Zimmermann 2000) we fit an ensemble of species distribution models (SDM) (Araújo & New 2007), using the R package `sdm` (Naimi & Araujo 2016). For the ensemble, we used five different model algorithms to represent the breadth of techniques available for modelling species distributions: generalised linear model (GLM), Multivariate Adaptive Regression Splines (MARS), Random Forest classification (RF), Radial Basis Function (RBF; artificial neural network) and Flexible Discriminant Analysis (FDA). We assessed the accuracy of each algorithm using the Receiver Operating Characteristic (ROC) curve (Allouche, Tsoar & Kadmon 2006). We used repeated k-fold cross validation (k = 10) to evaluate model performance and accuracy, and to tune model parameters (for each of the five algorithms) until they maximised the Area Under the Curve (AUC: Table 1). AUC values were also used for model refinement (variable selection/rejection), where only predictor values that provided the most accurate and parsimonious predictions were retained (Table 2). An ensemble projection of probability of occurrence for the Thylacine was derived by calculating a multi-model (unweighted) average of all five model projections (Marmion *et al.* 2009). This spatial layer of habitat suitability is available within poems (`poems::thylacine_hs_raster`, Haythorne *et al.* (2021)).

Table 1: Cross validation scores for Thylacine species distribution models.

Model	Predictor	AUC
GLM	6	0.78
MARS	9	0.79
RF	11	0.81
RBF	11	0.78
FDA	6	0.78

Area Under the Curve (AUC) scores for five SDM algorithms based on 10-fold cross validation: generalised linear model (GLM), Multivariate Adaptive Regression Splines (MARS), Random Forest classification (RF), Radial Basis Function (RBF; artificial neural network) and Flexible Discriminant Analysis (FDA). Predictor shows the number of predictors needed to maximise AUC.

Table 2: Climate and environmental predictors in different SDMs.

Predictors	GLM	MARS	RF	RBF	FDA
meanDiurnal		✓	✓	✓	
isothermal	✓	✓	✓	✓	
meanTwetQ		✓	✓	✓	
meanTdryQ	✓	✓	✓	✓	✓
precipseas	✓	✓	✓	✓	✓
precipwarmQ	✓	✓	✓	✓	✓
elevation		✓	✓	✓	✓
distFW	✓		✓	✓	✓
roughness			✓	✓	
vegetation		✓	✓	✓	
land-use	✓	✓	✓	✓	✓

Predictor variables in different SDM algorithms used to estimate probability of occurrence for Thylacine: mean diurnal range (meanDiurnal), isothermality (isothermal), mean temperature in the wettest quarter (meanTwetQ), mean temperature in the driest quarter (meanTdryQ), seasonality in precipitation (precipseas), mean precipitation in the warmest quarter (precipwarmQ), elevation, distance to freshwater (distFW), topographic roughness, pre-European vegetation type and land-use. Ticks show variables selected based on cross validation. SDM algorithms are described in Table 1.

References

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.

Allouche, O., Tsoar, A. & Kadmon, R. (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology,* **43,** 1223-1232.

Araújo, M.B. & New, M. (2007) Ensemble forecasting of species distributions. *Trends in Ecology & Evolution,* **22,** 42-47.

Barbet-Massin, M., Jiguet, F., Albert, C.H. & Thuiller, W. (2012) Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution,* **3,** 327-338.

Brook, B.W., Sleightholme, S.R., Campbell, C.R., Jarić, I. & Buettel, J.C. (2021) Extinction of the Thylacine. *bioRxiv***,** 2021.2001.2018.427214.

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.

Guisan, A. & Zimmermann, N.E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling,* **135,** 147-186.

Haythorne, S., Fordham, D.A., Brown, S.C., Buettel, J.C. & Brook, B.W. (2021) poems: Pattern-Oriented Ensemble Modeling System. R package version 1.0.1 https://cran.r-project.org/package=poems.

Marmion, M., Parviainen, M., Luoto, M., Heikkinen, R.K. & Thuiller, W. (2009) Evaluation of consensus methods in predictive species distribution modelling. *Diversity and Distributions,* **15,** 59-69.

Naimi, B. & Araujo, M.B. (2016) sdm: a reproducible and extensible R platform for species distribution modelling. *Ecography,* **39,** 368-375.

Scharples, C. (2011) Potential Climate Change Impacts on Geodiversity in the Tasmanian Wilderness World Heritage Area: A Management Response Position Paper. *Nature Conservation Report Series 11/04*. Hobart.

Senay, S.D., Worner, S.P. & Ikeda, T. (2013) Novel Three-Step Pseudo-Absence Selection Technique for Improved Species Distribution Modelling. *PLOS ONE,* **8,** e71218.