Nature Conservation 18: 27–46 (2017)
doi: 10.3897/natureconservation.18.12235
http://natureconservation.pensoft.net

RESEARCH ARTICLE

A peer-reviewed open-access journal
Nature Conservation
Launched to accelerate biodiversity conservation

Applying surrogate species presences to correct sample bias in species distribution models: a case study using the Pilbara population of the Northern Quoll

Shaun W. Molloy[1], Robert A. Davis[1], Judy A. Dunlop[2], Eddie J.B. van Etten[1]

1 *School of Science, Edith Cowan University, 270 Joondalup Drive, Joondalup, Western 5 Australia, 6027*
2 *Department of Parks and Wildlife, Science and Conservation Division, Dick Perry Ave, Kensington Western Australia, 6151*

Corresponding author: *Shaun W. Molloy* (shaunecologist@gmail.com)

Academic editor: *B. Gruber* | Received 12 February 2017 | Accepted 19 April 2017 | Published 9 May 2017

http://zoobank.org/6DC8ABF9-C4E8-499B-A7E0-4BC23BC35CD4

Citation: Molloy SW, Davis RA, Dunlop JA, van Etten EJB (2017) Applying surrogate species presences to correct sample bias in species distribution models: a case study using the Pilbara population of the Northern Quoll. Nature Conservation 18: 25–46. https://doi.org/10.3897/natureconservation.18.12235

Abstract

The management of populations of threatened species requires the capacity to identify areas of high habitat value. We developed a high resolution species distribution model (SDM) for the endangered Pilbara northern quoll *Dasyurus hallucatus*, population using MaxEnt software and a combined suite of bioclimatic and landscape variables. Once common throughout much of northern Australia, this marsupial carnivore has recently declined throughout much of its former range and is listed as endangered by the IUCN. Other than the potential threats presented by climate change, and the invasive cane toad *Rhinella marina* (which has not yet arrived in the Pilbara). The Pilbara population is also impacted by introduced predators, pastoral and mining activities. To account for sample bias resulting from targeted surveys unevenly spread through the region, a pseudo-absence bias layer was developed from presence records of other critical weight-range non-volant mammals. The resulting model was then tested using the biomod2 package which produces ensemble models from individual models created with different algorithms. This ensemble model supported the distribution determined by the bias compensated MaxEnt model with a covariance of of 86% between models with both models largely identifying the same areas as high priority habitat. The primary product of this exercise is a high resolution SDM which corroborates and elaborates on our understanding of the ecology and habitat preferences of the Pilbara Northern Quoll population thereby improving our capacity to manage this population in the face of future threats.

Keywords
Northern Quoll, Pilbara, MaxEnt, biomod2, Sample Bias, Cane Toad, Threatened Species

Introduction

Species distribution models (SDMs) use environmental data from known locations of a species to predict places where that species could potentially occur within landscapes or regions (Booth et al. 2014). SDMs have been used to identify critical habitats for species with greatly reduced distributions (Hamilton et al. 2015; Jetz and Freckleton 2015; Manthey et al. 2015), provide potential translocation sites for species based on known habitat requirements (Adhikari et al. 2012) and predict the movement of invasive species across landscapes under different scenarios (Kearney et al. 2008; Elith et al. 2010). Spatially explicit probability of presence, or prediction of occurrence maps, generated using SDM algorithms, have been used to inform conservation planning and habitat management at both coarse and fine scales. Consequently, they can guide or prioritise future survey efforts and aid in assessing the conservation status of target species. SDMs relate known occurrences of a species with various environmental variables and predict a probability that a species will occur in areas where no data on its occurrence is available. Thus, they help to identify potentially suitable habitat (Elith and Leathwick 2009; Guisan and Thuiller 2005).

The accuracy of a SDM depends on such factors as the quality and appropriateness (in regard to sample size and representativeness) of the presence and/or absence data for the target species or community, the expertise of the modeller, the selection of an appropriate modelling tool (or software package), the selection of an appropriate suite of predictive/independent variables, the quality of the variable data used, and an acknowledgement of the strengths and limitations of the SDM (Elith et al. 2010; Guillera-Arroita et al. 2015). However, where there are shortfalls in appropriate data, modellers often compensate by focussing their efforts on the development and evaluation of novel methods to improve the performance of their models, and thus, to better predict the environmental suitability for species in applied studies (Barbosa and Schneck 2015; Guillera-Arroita et al. 2015).

In this study we set up a SDM to determine the potential distribution (PD) of the northern quoll *Dasyurus hallucatus* population found in the Pilbara biogeographic region of Western Australia (Thackway and Cresswell 1997). This is a unique population of a threatened species for which conservation management to date has been limited by significant knowledge gaps (Cramer et al. 2015). Limited and largely opportunistic sampling has resulted in constricted and potentially biased presence data and this in turn has resulted in problems in determining appropriate predictive or independent variables. Furthermore, the literature associated with these SDMs shows no evidence that the models have been tested for covariance and sample bias as described by Hijmans (2012) and Fithian et al. (2015), nor have differences between these SDMs been evaluated or explained.

Northern quolls are a suitable subject for distribution modelling as they have a strong habitat affiliation with complex rocky areas, often in close association with permanent water (Begg 1981; Braithwaite and Griffiths 1994; Oakwood 1997; Pollock 1999; Schmitt et al. 1989). In the Pilbara, quoll habitat affiliation aligns with mesas or ranges which are often the focus of iron-ore extraction (channel-iron deposits and banded iron stone formations) and granite outcrops which are often quarried for road and rail bed materials (Ramanaidou and Morris 2010).

Once widely distributed from the Pilbara region of Western Australian (WA) across northern Australia to southern Queensland (Figure 1), the mainland distribution of the northern quoll has now contracted to several disjunct populations (Burbidge et al. 2009; Oakwood 2008). This collapse has largely been linked to invasion by the toxic cane toad *Rhinella marina* (Braithwaite and Griffiths 1994; Doody et al. 2015; How et al. 2009; Oakwood 2004; Woinarski 2010). Other impacts currently causing rapid and severe declines in northern Australia's critical weight range mammal fauna (i.e. terrestrial species within the weight range 35 – 4200 g are considered to be particularly vulnerable to introduced predators) are also likely to also be impacting on the carnivorous northern quoll (Burbidge and McKenzie 1989; Cramer et al. 2016). These include altered fire regimes, the grazing impacts of introduced herbivores, climate change and both enhanced mortality and competition for resources (including prey animals) from introduced predators, in particular the feral cat *Felis catus* (Burbidge et al. 2009; Cook 2010; Woinarski et al. 2015). As a consequence of all these recent declines, the northern quoll is listed as endangered under both the Commonwealth's *Environment Protection and Biodiversity Conservation Act 1999* (*EPBC Act 1999*) and the Western Australian *Biodiversity Conservation Act 2016*.

The main WA populations of northern quoll occur in two discrete mainland regions, the Kimberley and Pilbara, separated by the arid Great Sandy Desert. Both mitochondrial DNA sequences and nuclear microsatellite loci reveal clear differentiation between Kimberley and Pilbara populations and a greater distinction between these populations than those in the Northern Territory and Queensland (Spencer et al. 2013; Westerman and Woolley 2016). These WA populations also differ from those remaining in Queensland and the Northern Territory in regard to both genetic structure and demographic parameters and represent the last intact populations in Australia that have not experienced major declines subsequent to the introduction of the cane toad and consequently display the highest levels of genetic integrity (How et al. 2009; Spencer et al. 2013; Spencer 2010).

Given that the Pilbara population of the northern quoll is genetically and demographically distinct from all other populations, retains its pre-European genetic diversity, is currently outside of the cane toad's distribution, and has much of its habitat still intact, this population has been assigned a high conservation, research and management priority (Cramer et al. 2015).

The available presence data for this population is clustered around areas of development interest to the mining industry, or where targeted surveys have occurred. This

begged the question: was this an example of sample bias or a true representation of northern quoll distribution? Sample bias, where sampling has not been uniform over the project area, e.g. where only easily accessed areas, or known populations have been sampled, has the potential to distort a SDM (Phillips et al. 2009). Lacking true absence data for this exercise and being aware of the limited capacity of pseudo-absences to compensate for high levels of sample bias (Phillips et al. 2009), we sought to find a method to eliminate or minimise sample bias in our SDM.

Our objective was to construct a high resolution SDM for the Pilbara northern quoll by applying an innovative form of bias compensation to a well proven modelling method, MaxEnt, and testing this SDM with an ensemble model.

Methods

Study area

The area modelled for this exercise is the Pilbara biogeographic region (Fig. 1) (Thackway and Cresswell 1997). This selection encompasses all the known occurrences of the unique Pilbara population at the time of the study, and satisfies the requirements and priorities of this project's stakeholders and funding bodies.

Presence data

All presence data, both for northern quoll and surrogate species, was supplied by the West Australian Department of Parks and Wildlife NatureMap database (Naturemap 2015), and comprised of (~2000) records for this species within the Pilbara up to, and inclusive of, 2014. These species records are both targeted and opportunistic, sourced from museum, and Parks and Wildlife fauna surveys as well as compulsory returns from biological consultants and industry. The NatureMap threatened species database is continually being verified and updated by cross-referencing from grey literature, fauna returns and reporting required under the Western Australian *Biodiversity Conservation Act 2016.*

Variable selection

To obtain optimum efficiency, minimize multicollinearity and prevent overfitting, the suite of variables used should be kept compact (preferably ≤10 in number) and be comprised of those variables best able to define the PD of the target species or community (Beaumont et al. 2005; Elith et al. 2011; Hijmans 2012; Van Gils et al. 2012). To accomplish this, we reviewed the literature on the Northern Quoll in general and the Pilbara population in particular, to identify independent variables likely to be in-

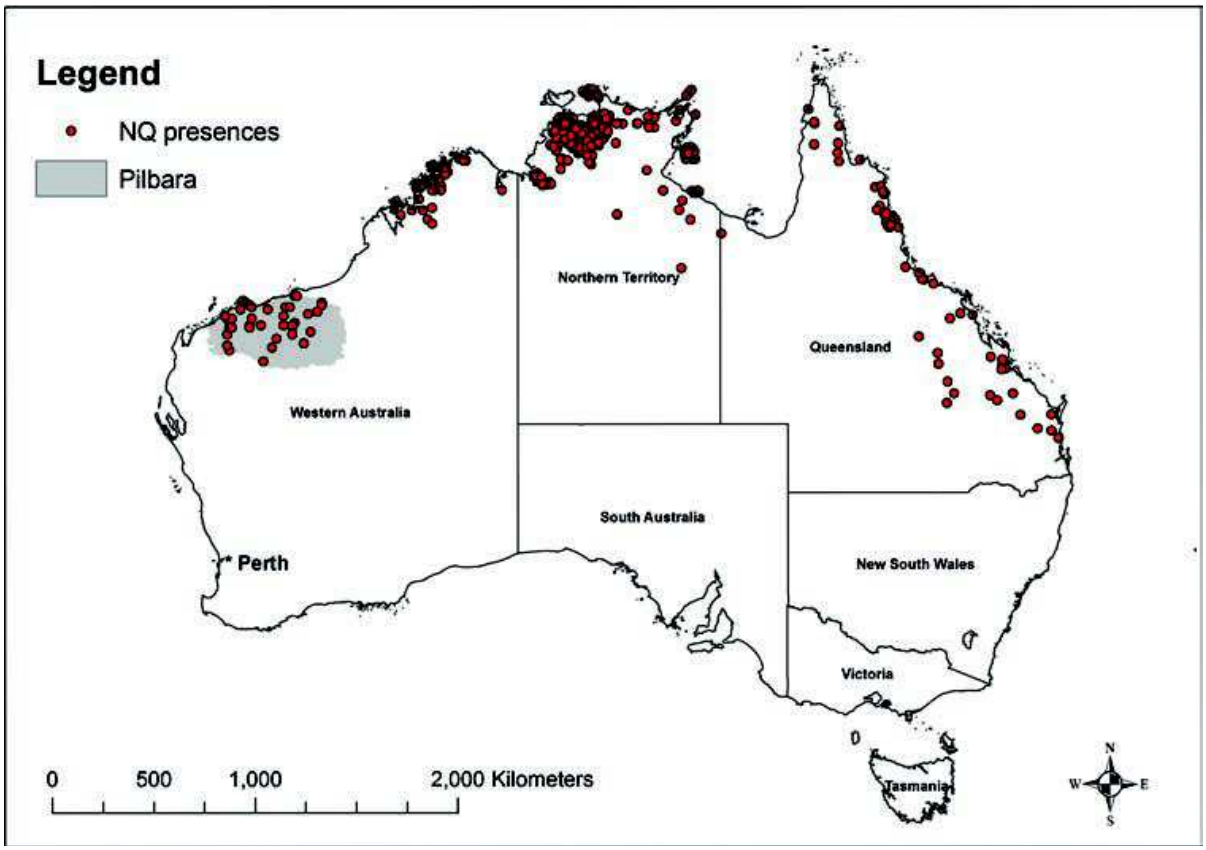


Figure 1. Northern Quoll presences with Pilbara bioregion shaded. Presence data for this example was sourced from Atlas of Living Australia (2015).

fluential or linked to its distribution and therefore suitable for producing a SDM (see Suppl. material 1). All variable data sets were downloaded at, or converted to, a pixel resolution of 30 seconds (~1km^2) using the WGS 1984 datum and clipped to the Pilbara bioregion. Map projection (WGS 84) was consistent in all data sets, which were aligned by importing raster data onto a common grid and converting the outputs to ASCII grid file format using ArcGIS 10.4.

To avoid using unnecessary time and resources in identifying an appropriate suite of predictive variables, a two-stage process was adopted. The first stage used a series of statistical tests (described below) to halve the number of potential variables so as to quickly and effectively limit multicollinearity in the bioclimatic variables and to remove those variables for which their contribution to the model was low or counter-productive. The second stage was to use a stepwise elimination process to identify a final suite of predictive variables suitable for use in all further modelling.

Firstly, to reduce multicollinearity between scalar variables, we calculated both the Pearson and Spearman rank correlation coefficient between each pair of variables using the northern quoll presence data. This was done using the pairs function in the 'psych' R package (Revelle 2014). For each pair of highly correlated variables ($r > 0.70$), we selected only the single variable deemed to be the most relevant for identifying northern quoll presences based on ecological relevance and expert opinion (Phillips and Dudík 2008). Categorical variables were tested for association with each other using Pearson's chi-square

test for significance. Similarly, relationships between categorical and scalar variables were tested using linear models with a 0.05 level of significance (Agresti and Kateri 2011).

The final cut was undertaken through a step-wise elimination process using MaxEnt (Phillips et al. 2006) looking at both the contribution of each remaining variable and the consequences of its omission. This was done by building a MaxEnt SDM using all remaining predictive variables against the Pilbara northern quoll presence-only point data set and then iteratively removing the worst performing variable in regard to variable contribution and jack-knife tests. The model was then re-run and if sufficiently robust, the process was repeated. The model at the start of this process was robust, with acceptable Area Under Curve (AUC) and regularised training gain values, so if the revised model was not sufficiently robust the variable was reintroduced to the model and the process repeated with the second worst performing variable deleted. This process was repeated until we had a minimum number of variables capable of producing and SDM with a minimum AUC value of 0.9. This minimum threshold value was set to ensure a statistically very strong model (Elith et al. 2011).

MaxEnt modelling

For our primary modelling tool we selected MaxEnt for its capacity to produce effective SDMs using presence-only data (Booth et al. 2014; Yackulic et al. 2013). MaxEnt, or maximum entropy, modelling is a machine learning modelling tool which seeks to estimate a target probability distribution by finding the probability distribution of maximum entropy, i.e. where variable parameters are closest to homogenous, subject to the limitations of the data used (Guillera-Arroita et al. 2015). This is achieved by applying the predictive or independent variables against training data which are a subset of presences randomly selected and assumed to be representative for the modelled distribution (Radosavljevic and Anderson 2014).

Some limitations have been recognised with MaxEnt, notably a tendency for it to underperform where there is a biased sample, poorly chosen predictive variables or inadequate testing of results (Bystriakova et al. 2012; Elith et al. 2011; Kramer-Schadt et al. 2013). However, where these limitations are addressed, it remains a well-supported modelling tool because it is relatively easy to use and has a capacity to link fine-scale bioclimatic data to species distributions to produce accurate probability-based outputs suitable for informing conservation management actions (Guillera-Arroita et al. 2015; Phillips and Dudík 2008; Syfert et al. 2013; Williams et al. 2012). Consequently, we constructed our MaxEnt SDM with the following criteria:

- Withholding a random 30% of presences for testing purposes over 10 bootstrapped repetitions (Barbet-Massin et al. 2012).
- Combining all presences within a pixel (~1km^2) as a single presence. This resulted in a reduction in the number of presences from 1984 to 324.

Bias compensation

To compensate for our limited presence data we followed Fithian et al. (2015) in constructing a bias layer using substitute species which we applied only to the MaxEnt model. Therefore, we generated a bias grid by substituting records of other non-volant critical weight range (CWR) mammals of 35–4200g (Burbidge and McKenzie 1989), obtained throughout the Pilbara in lieu of northern quoll presence/absence data. These data were used as a presence/absence point data set that would reflect a sampling effort likely to detect northern quoll. The reasoning behind this is: 1) CWR presence records reflect sampling effort; 2) sampling for non-volant CWR mammals would most likely result in northern quoll presence records (e.g., capture, sighting, tracks, scats, or other physical evidence such as remains) if indeed they were present; 3) therefore, presence records for non-volant CWR mammals are suitable for use as pseudo-absence data; and 4) a point density analysis of non-volant CWR mammal presences for the whole of the Pilbara would indicate the degree of bias present in the northern quoll presence records and could therefore be used as a bias grid in the MaxEnt model.

To construct this bias grid, presence records for all non-volant CWR mammals (including northern quoll) in the Pilbara were gathered from the Department of Parks and Wildlife Nature Map data base (Department of Parks and Wildlife 2007-) and categorised into northern quoll presences and pseudo-absences. We note that, although this was a separate data set to that of the original presence data set, many presences were replicated. All records were then used to conduct a Point Density Analysis (PDA) using the Point Density function of the Spatial Analyst toolbox in ArcGIS 10.3. This function counted the total number of records for each cell within a 44 cell radius (the default radius). The resulting shapefile was then directly incorporated as a bias grid into the MaxEnt SDM.

Testing the SDM with an ensemble package

As the above SDM was compiled using just one modelling tool and as different algorithms and methodologies can yield very different and often contradictory results, we opted to test the rigour of the preferred (MaxEnt) SDM using ensemble modelling techniques. This involved compiling a suite of different algorithms to construct multiple SDMs for the target species within a single platform and then combining these SDMs to produce a single ensemble, or composite, SDM (Crimmins et al. 2013; Grenouillet et al. 2011). This approach enabled us to compare the MaxEnt SDM with individual and ensemble model outputs and differences between modelling algorithms to be compared.

The ensemble modelling was undertaken using the biomod2 package in the R platform (Thuiller et al. 2013). This package allowed the use of the same variable, presence and pseudo-absence data used to develop the preferred MaxEnt SDM.

We selected the five best performing modelling algorithms for our ensemble model. These were Generalised Linear Model (GLM), Generalised Additive Model (GAM), Generalised Boosted Model (GBM), Flexible Discriminant Analysis (FDA) and Multiple Adaptive Regression Splines (MARS). In running these a random 30% of presences would be used to calibrate the model and 70% of presences could be withheld for testing. This process was then repeated 10 times to add rigour to the results. Unlike MaxEnt, the biomod2 package does not provide an option to use a bias layer to compensate for sample bias. Therefore we applied the surrogate presence data set, from which we constructed the bias layer, as a substitute for true absences in our model imputs.

All outputs of all algorithms were evaluated with a True Skill Statistic (TSS), and Receiver Operator Characteristic (ROC - a test comparable with the MaxEnt's AUC statistic) and combined. A weighting was given to each algorithm based on ROC performance and all model outputs were combined to produce a weighted mean SDM which we used as our biomod2 output.

Comparisons between the MaxEnt and biomod2 SDM were again made using the pairs function in the psych R package. This was done by compiling a point data set of 10,000 random points across the study area. This point data set was then used to extract values from both SDMS and the two resulting data sets compared through the pairs analysis.

The individual modelling packages used, their results and the results of the ensemble modelling process are given in Suppl. material 2.

Results

Variable selection

From the broad suite of variables tested (Suppl. material 1) we derrived a final suite of seven variables with acceptable levels of covariance for use in all models (Figure 2). Spearman rank correlation coefficients are well under the 0.7 threshold. Although Pearson correlation coefficients were also under this value the Spearman value is used as the most appropriate given that not all distributions appear normal. The MaxEnt analysis provided the contribution and importance values for each variable (Table 1).

MaxEnt SDM

The bias file (Figure 3) demonstrates that sampling has not been uniform with most sampling being undertaken in the north east of the region in areas subject to relatively heavy mining activity. On further examination more instensive sampling along infrastructure corridors, e.g railways, major roads and powerline routes, became obvious.

The MaxEnt SDM (Figure 4a) appears to be a robust model with a high average AUC value of 89.5 (the full MaxEnt readout including AUC plots, variable responses,

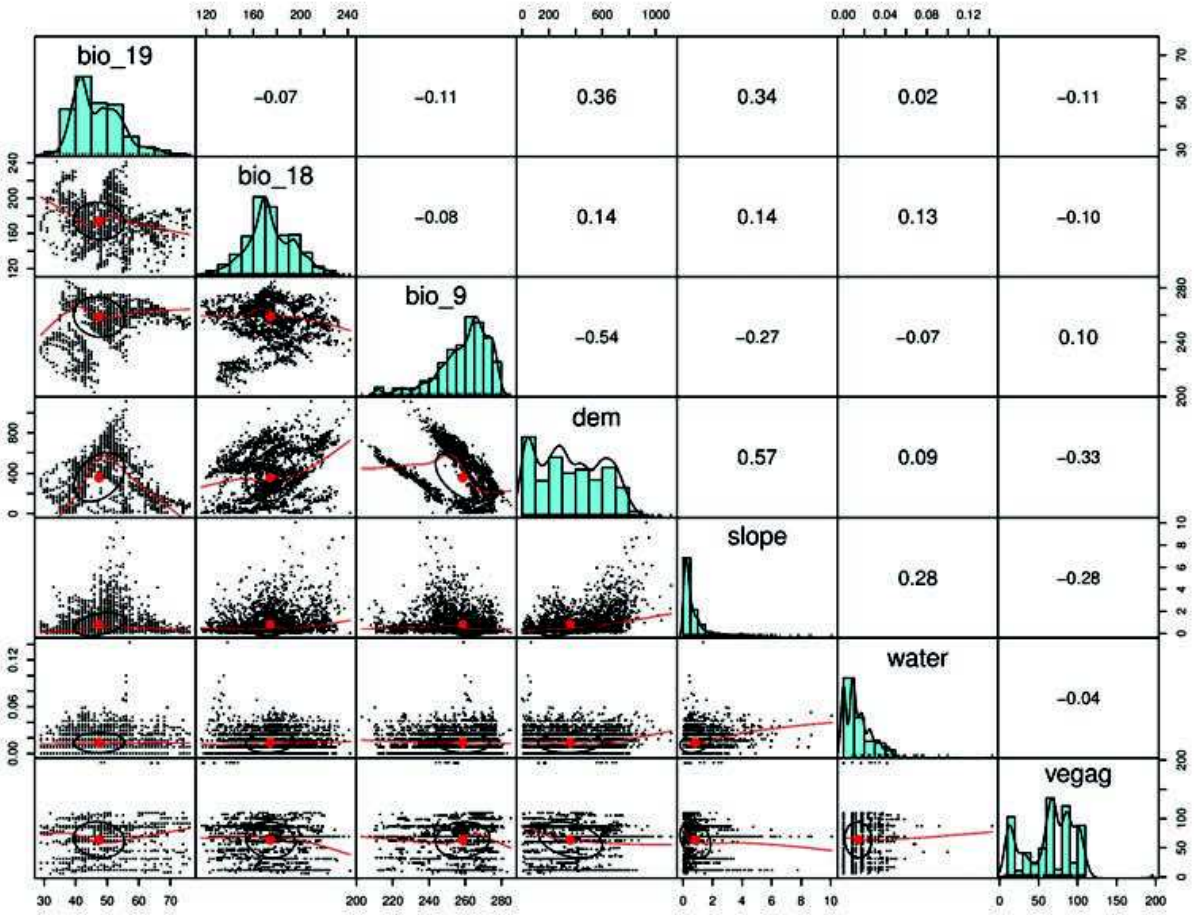


Figure 2. Pairs analysis of predictive variables against northern quoll presences. Diagonal=variable name and histogram, left of diagonal= scatter plots and trend lines and right of diagonal gives Spearman rank correlation coeffient. Axis figure represent point values corresponding variables.

Table 1. Final suite of variables with % contribution and permutation importance as determined through step-wise MaxEnt analyses. All contribution and importance values reflect positive relationships to northern quoll presence. Source data for all variables is available in Suppl. material 1.

Variable	% Contribution	Permutation Importance
Vegag= Department of Agriculture and Food Western Australia Vegetation Mapping (Rangelands)	35	15.5
DEM = Digital Elevation Model	26.2	37.4
BIO18 = Precipitation of Warmest Quarter	15.1	16.3
Slope= Terrain slope raster produced from the DEM	11.5	14
BIO9 = Mean Temperature of Driest Quarter	4.7	3.3
BIO19 = Precipitation of Coldest Quarter	4.1	9.3
Water = Euclidean Distance to Water Courses	3.4	4.2

sensitivity, threshold diagnostic data is given in Suppl. material 2). This model identifies a high probability of occurrence for many areas already known to be northern quoll habitat such as the rocky habitats on the western edge of the Hamersley Ranges,

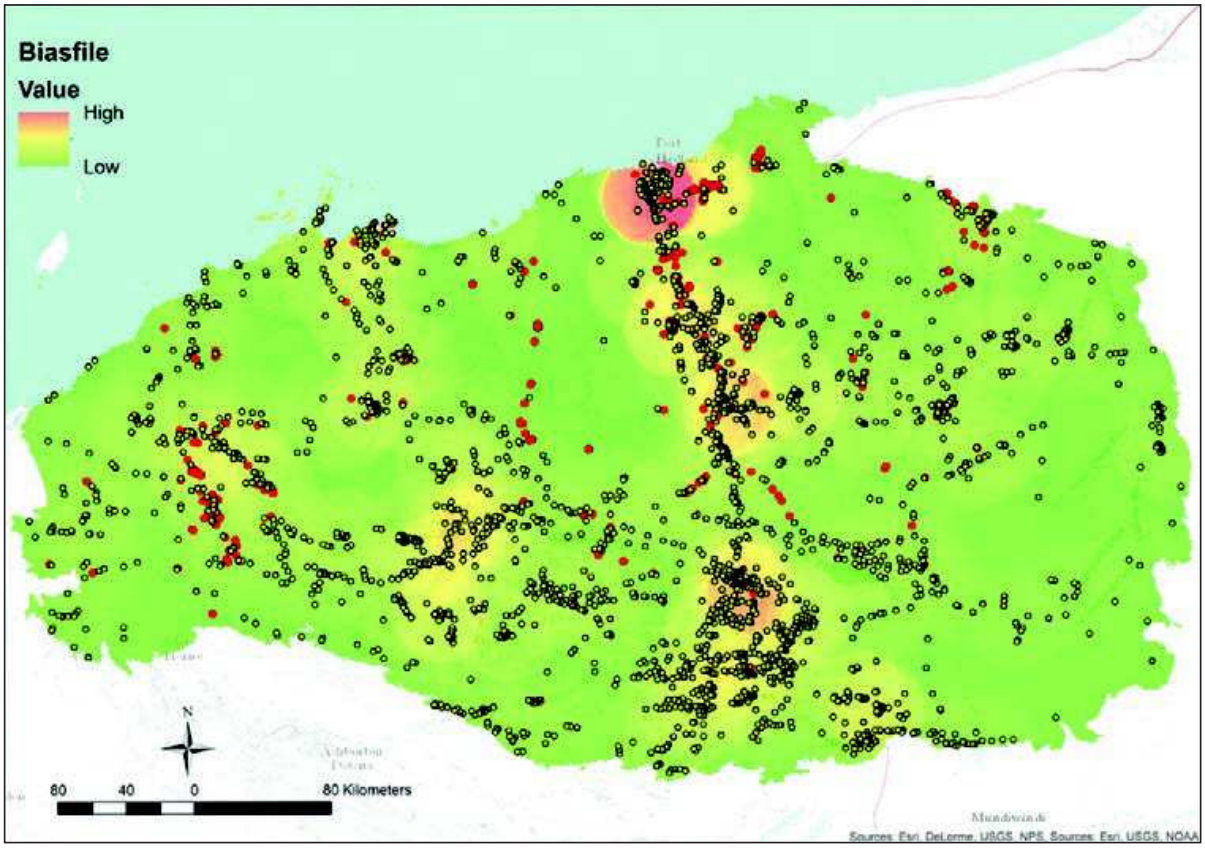


Figure 3. Bias grid GIS file created from pseudo absence data (presence records for critical weight range mammals). Red dots are Northern Quoll presences and yellow dots are other non-volant CWR species.

the rugged Chichester Ranges and the granite outcrops of the Abydos Plains (a map of these areas is given in Suppl. material 1). However, it projects beyond known presences to predict a low probability of occurrence in the Fortescue catchment, the sandy coastal regions of the Pilbara and in the southern areas of the Hamersley Ranges, and to predict a high potential for northern quoll habitat in many areas where this species has not been previously identified, particularly in the central west and far eastern parts of the region. When compared to the MaxEnt SDM, constructed without the use of the bias layer (Figure 4b), it appears that the use of the bias layer extends predicted habitat further beyond those highly sampled areas where northern quoll are frequently encountered identifying further areas, where this species has not been, or has rarely been, previously encountered.

Comparison with the biomod 2 ensemble model

The full outputs for the biomod2 modelling process are given in Suppl. material 3. This is a robust model with ROC values for individual SDMs varying between 0.88 and 0.97. To facillitate a visual comparison, the ensemble model has been projected at the same extent, symbology and resolution as the MaxEnt model (Figure 5). A comparison between this (emsemble) SDM and the MaxEnt model (Figure 4) shows

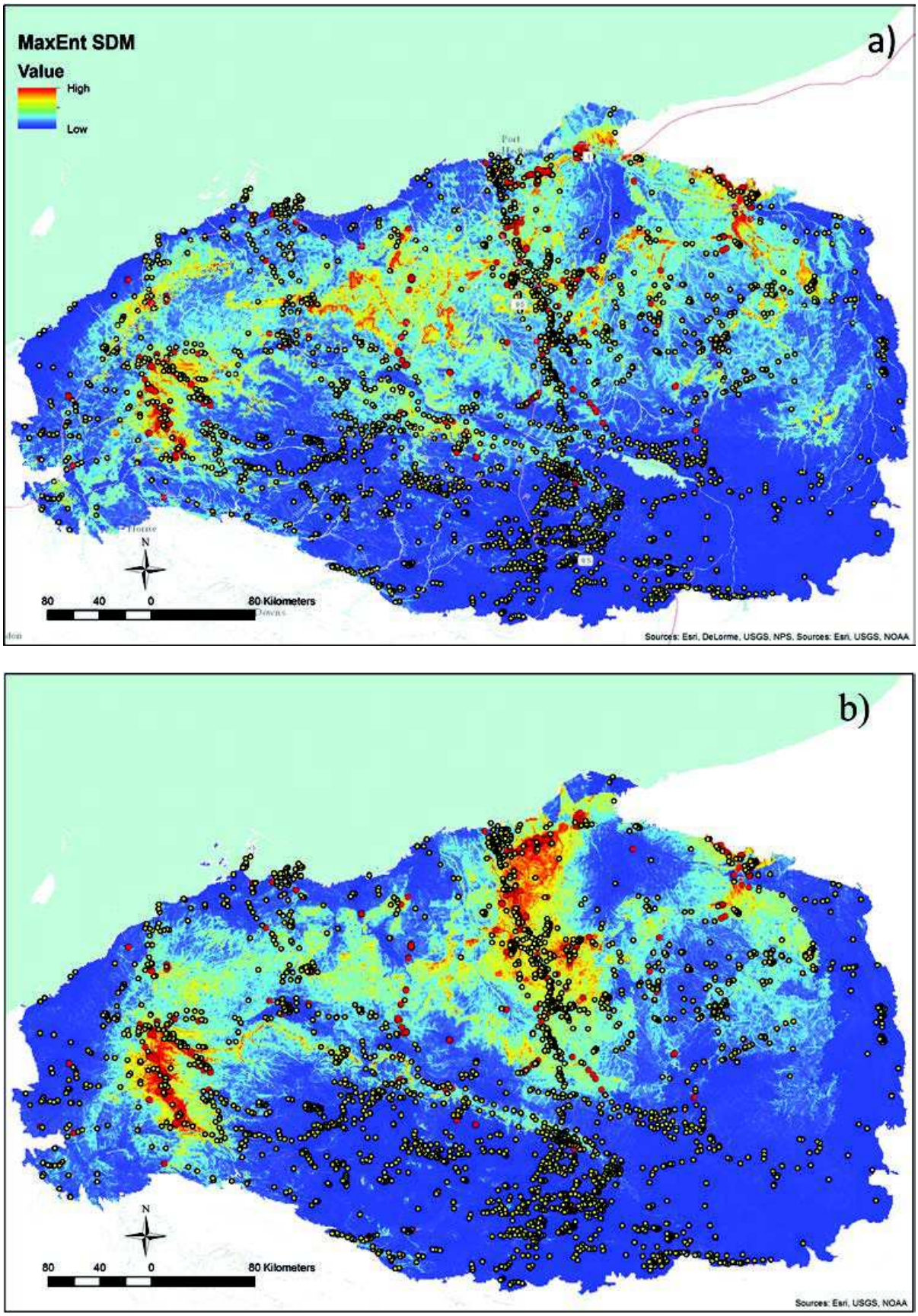


Figure 4. a MaxEnt SDM constructed with Bias Grid (Figure 3) and the final variable suite (Table 1) **b** MaxEnt SDM constructed without bias grid. Red dots indicate northern quoll presences and yellow, surrogate presences.

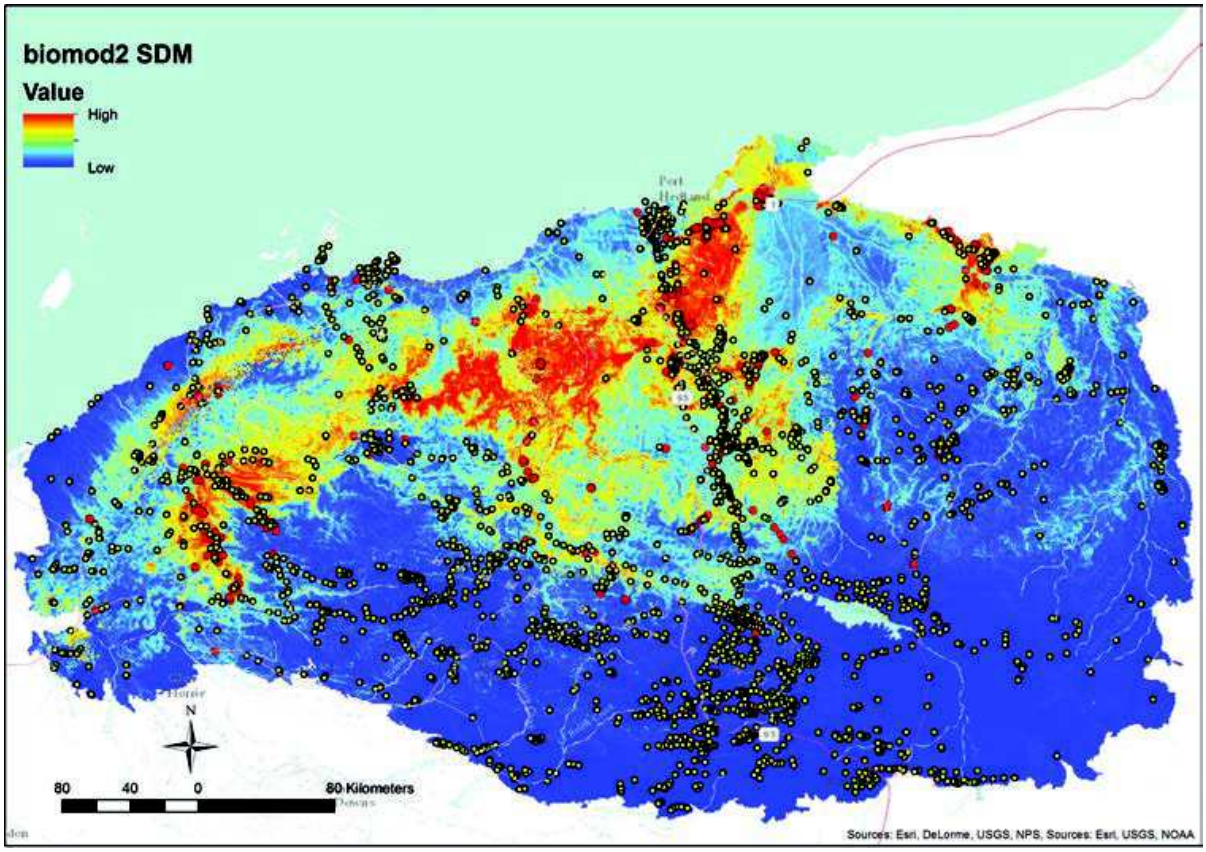


Figure 5. Weighted mean biomod2 (ensemble model) projected in ArcGIS 10.3 as per the MaxEnt model in Figure 3.

a strong general similarity between both models in terms of the areas selected as likely habitat, with the main differences being variations in habitat value of the areas selected rather than which areas were actually selected. The pairs analysis (Figure 6) shows a strong similarity between these SDMs with an overall Spearman correlation coefficient of 0.86. Histograms show very similar distributions across the full range of both models and like the MaxEnt SDM the biomod2 ensemble model also highlights similar, but little sampled, areas as potential northern quoll habitat.

Discussion

The independent variables that we used were suitably diverse in nature with an acceptable level of covariance within the variables used in the final selection. The use of the jack-knife test to determine the final suite showed that the removal of any of the variables selected in the final suite would have compromised what was a strong model. The literature informs us that the number of variables used was appropriate for a modelling exercise of this nature (Phillips and Dudík 2008; Stockwell and Peterson 2002). Furthermore, the fact that all runs of all models applied these variables differently and, in most instances, allotted acceptable levels of significance to all or most of these variables (while consistently producing high test values), supports this finding. It should

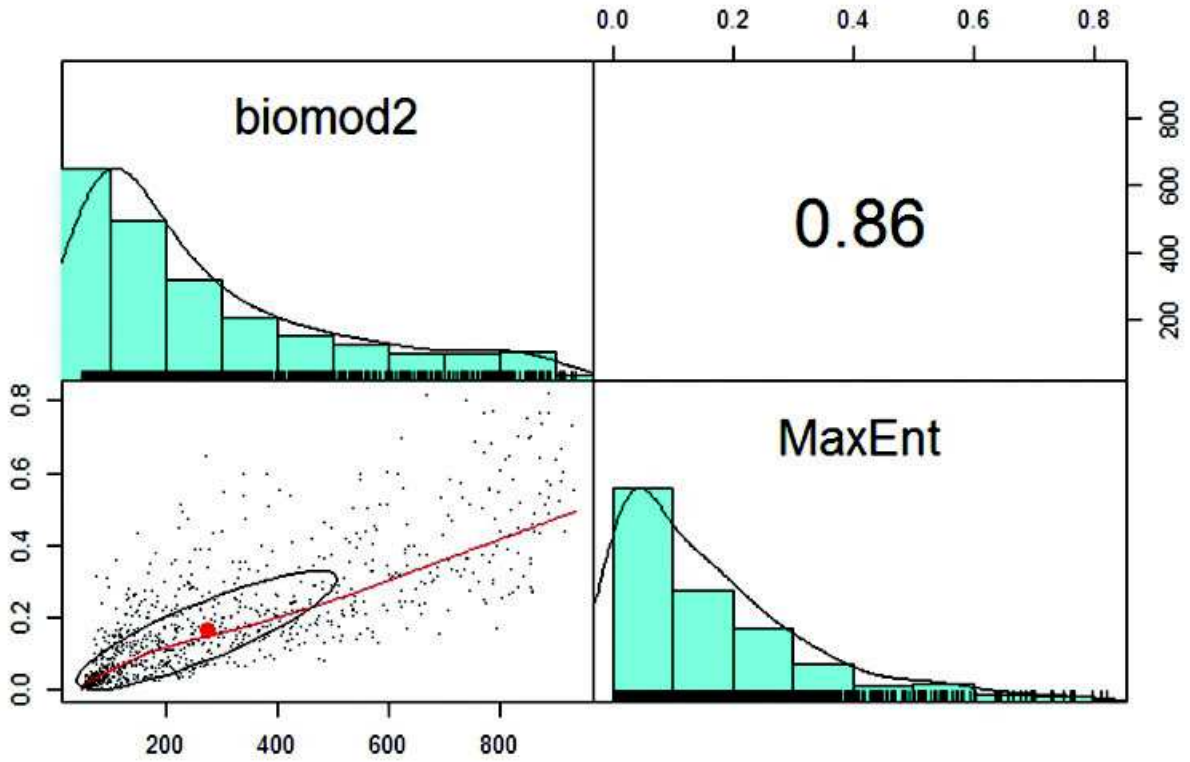


Figure 6. Pairs analysis between MaxEnt and biomod2 SDMs using 10,000 random points.

be acknowledged that the bias grid was derived from pseudo-absence data for which suitability was determined through a series of assumptions. Consideration of sampling bias was necessary because of a potential shortfall in suitable presence/absence data and the fact that most of the targeted sampling for the northern quoll has been undertaken over a relatively limited area and timeframe. Specifically, many surveys have been conducted in mining areas and associated infrastructure (either historical, current or proposed) which tend to confine the survey effort to particular types of geomorphology. Surveys of non-volant mammals conducted in the region tend to be for environmental impact assessments associated with mining or part of comprehensive regional surveys (e.g. Bamford Consulting Ecologists (2013); Biota Environmental Sciences (2012); Eco Logical (2012)) or part of comprehensive regional surveys (McKenzie et al. 2009) and hence absence of northern quoll in these surveys could be reasonably assumed to be true absences.

As demonstrated (Table 1, Figure 2, Suppl. materials 2 and 3) northern quolls were found to conform strongly to ecological habitat associated with the vegetation, and slope, topography of the rocky areas of the Pilbara bioregion. Primary areas of northern quoll occupation in the Pilbara included the western edge of the Hamersley Ranges, in the granite outcrops of the Abydos Plain, and in the more rugged areas of the Chichester Ranges. Our models identified a low likelihood of occurrence in the Fortescue River floodplain and its upper catchment, the sandy coastal regions of the Pilbara and in the central and southern parts of the Hamersley Ranges. We also identified several areas with a high probablity of presence which have not, as yet, been

adequately sampled and which have not been identified as habitat in previous modelling efforts. This is particuarly so in the eastern Pilbara IBRA region. Here, areas are given a high probability of northern quoll although few to no records of the species exist. This area and others identified as of high probability would be logical areas for further survey effort (as well as providing a ready field-based validation of the model), and demonstrate the utility value of SDMs.

The conservation of a threatened species requires good information about population locations and ecological requirements within its geographic range, particularly when threatening processes are ongoing. Applying appropriately selected surrogate species data to both the MaxEnt and biomod2 software packages has enabled us to develop SDMs which identify remarkably similar core areas of likely quoll habitat, as well as less optimal habitat that may only be occupied in favourable conditions. These apparently less favourable habitats may however be of high conservation value as current information suggests that all Pilbara northern quoll populations are genetically linked, and high level of dispersal occurs between geographically distant populations (Spencer et al 2013, Woolley 2015). Consequently, smaller populations of northern quolls in less-preferred habitat may be important in maintaining gene-flow throughout the region.

The SDM pairs analysis (Figure 6) shows a very strong overall correlation between the preferred MaxEnt SDM and the ensemble model with minor differences in predicted habitat being more about conflicts of scale rather than the actual areas nominated. The correlation tests also indicate a strong similarity between the preferred MaxEnt SDM and the ensemble model overall, and a very strong similarity between these models in identifying areas of high habitat value, i.e. when comparing the highest quartiles. This supports our earlier observation that the difference between the two models is largely one of resolution rather than different predictions.

In comparison with previous SDMs on this northern quoll population (Biologic 2012; CliMAS 2014; Eco Logical 2012) our SDM (particularly the MaxEnt model) picks up many of the same areas identified as having a high habitat value, but with an apparently greater level of definition. This is particuarly so in those areas which have not been heavily and specifically sampled for northern quoll. In short, there is a tendency for these models not to project far beyond those areas where northern quoll are known to exist. As such these models tend to tell us little more than what we already know, thereby limiting their value for conservation planning. This could also be said for the trial MaxEnt and biomod2 SDMs constructed without the use of surrogate species, either in the form of a bias file or as pseudo-absences. In both cases, areas of high quality habtat appeared to be much more limited to areas already known to be habitat and less likely to projet into areas where northern quoll records are either absent or less frequent.

In this study we have demonstrated a methodology capable of addressing three of the more common problems associated with SDMs, specifically how to: 1) address bias in a high resolution SDM over a large and diverse landscape with a limited, and potentially biased, presence only data set; 2) selecting an appropriate suite of predictive variables for the construction of such a model; and 3) establish a means by which the suitability and outputs of a modelling tool can be verified. We have developed an

innovative approach to constructing an SDM by pre-emptively identifying problems likely to arise due to data limitations and addressing these issues by reviewing the options available and selecting a combination of responses which minimise bias effects and meet the needs and constraints of the modeller.

We note that a true comparison between a model with randomly selected psuedo-absences and a bias-corrected model using a bias layer created from surrogate presences, remains the preferred way to demonstrate the usefulness of this form of bias compensation. However, in the absence of broad-scale sampling and ground truthing to test truth versus prediction (as is proposed) the demonstrated approach remains the most feasible form of bias compensation under the given circumstances.

Being aware of the resource and skill limitations preventing many conservation managers from constructing SDMs, this methodology was deliberately selected to meet these limitations. All data used was freely available and all software used was freeware, other than the use of a commonly used GIS package that could also be substituted with freeware. Skill levels were limited to what the authors considered average for an ecological project team, i.e. no modelling, statistical, GIS or programming specialists were required for this study.

Conclusion

In this exercise we produced a SDM that predicted areas where new populations and sub-populations of the northern quoll might be found outside of areas currently known to be habitat. By identifying areas of high habitat value, this SDM facilitates the identification and conservation of high priority habitat areas, potential translocation sites and potential movement corridors. As a consequence of having identified a sound suite of predictive variables, our understanding of the habitat requirements of the Pilbara population of the northern quoll has been increased. Finally, by comparing the distributions identified through this exercise with proposed mining and infrastructure projects in the environmental impact assessment process, this SDM can be used to minimise impacts on this unique and important northern quoll population.

The ensemble modelling process validates our choice of the MaxEnt model with bias file as the preferred SDM. However, this is a desktop exercise derived from a relatively small and uniform sample. This model should be validated and refined through on ground sampling and research. Our study exemplifies a preferred practice in the use of SDMs by corroborating our findings with a control, in this case one derived through an ensemble modelling approach. We commend this practice to modellers and caution against outcomes derived from only a single modelling approach. Our study has revealed the most comprehensive known refined distribution map for the endangered Northern Quoll. It has confirmed their reliance on rocky upland habitats and their limited distribution within the Pilbara region. These outcomes will be of great importance to land managers when considering the impacts of planned developments within the region.

References

Adhikari D, Barik S, Upadhaya K (2012) Habitat distribution modelling for reintroduction of Ilex khasiana Purk., a critically endangered tree species of northeastern India. Ecological Engineering 40: 37–43. https://doi.org/10.1016/j.ecoleng.2011.12.004

Agresti A, Kateri M (2011) Categorical data analysis. Springer Berlin Heidelberg, 206–208. https://doi.org/10.1007/978-3-642-04898-2_161

Atlas of Living Australia (2015) Atlas of Living Australia. http://www.ala.org.au/ [accessed 9/06/2015]

Bamford Consulting Ecologists (2013) Assessment of fauna values, Bonnie East, Warrigal North and Coongan. Report prepared for BC Iron Nullagine. Bamford Consulting Ecologists, Perth, WA.

Barbet-Massin M, Jiguet Fdr, Albert CcHln, Thuiller W (2012) Selecting pseudo-absences for species distribution models: how, where and how many? Methods in Ecology and Evolution 3: 327–338. https://doi.org/10.1111/j.2041-210X.2011.00172.x

Barbosa FG, Schneck F (2015) Characteristics of the top-cited papers in species distribution predictive models. Ecological Modelling 313: 77–83. http://dx.doi.org/10.1016/j.ecolmodel.2015.06.014

Beaumont LJ, Hughes L, Poulsen M (2005) Predicting species distributions: use of climatic parameters in BIOCLIM and its impact on predictions of species' current and future distributions. Ecological Modelling 186: 251–270. https://doi.org/10.1016/j.ecolmodel.2005.01.030

Begg RJ (1981) The Small Mammals of Little Nourlangie Rock, N.T III. Ecology of *Dasyurus hallucatus*, the Northern Quoll (*Marsupialia* : *Dasyuridae*). Wildlife Research 8: 73–85. https://doi.org/10.1071/WR9810073

Biologic (2012) Habitat Modelling for Selected Species of Conservation Significance in the Pilbara. Prepared for BHP Billiton Iron Ore Pty Ltd,. Biologic, Perth, WA.

Biota Environmental Sciences (2012) Fauna Habitats and Fauna Assemblage of the Proposed FMG Stage B Rail Corridor and Mindy Mindy, Christmas Creek, Mt Lewin and Mt Nicholas Mine Areas. Report prepared for Fortescue Metals Group. Biota Environmental Sciences Pty Ltd, Perth WA.

Booth TH, Nix HA, Busby JR, Hutchinson MF (2014) BIOCLIM: the first species distribution modelling package, its early applications and relevance to most current MAXENT studies. Diversity and Distributions 20: 1–9. https://doi.org/10.1111/ddi.12144

Braithwaite RW, Griffiths AD (1994) Demographic variation and range contraction in the northern quoll, *Dasyurus hallucatus* (Marsupialia: Dasyuridae). Wildlife Research 21: 203–217. https://doi.org/10.1071/wr9940203

Burbidge AA, McKenzie NL (1989) Patterns in the modern decline of western Australia's vertebrate fauna: Causes and conservation implications. Biological Conservation 50: 143–198. https://doi.org/10.1016/0006-3207(89)90009-8

Burbidge AA, McKenzie NL, Brennan KEC, Woinarski JCZ, Dickman CR, Baynes A, Gordon G, Menkhorst PW, Robinson AC (2009) Conservation status and biogeography of Australia's terrestrial mammals. Australian Journal of Zoology 56: 411–422. http://dx.doi.org/10.1071/ZO08027

Bystriakova N, Peregrym M, Erkens RHJ, Bezsmertna O, Schneider H (2012) Sampling bias in geographic and environmental space and its effect on the predictive power of species distribution models. Systematics and Biodiversity 10: 305–315. https://doi.org/10.1080/14772000.2012.705357

CliMAS (2014) Climate Change and Biodiversity in Australia. http://climas.hpc.jcu.edu.au/ [accessed 10/06/2015]

Cook A (2010) Habitat use and home-range of the Northern Quoll: effects of fire. University of Western Australia.

Cramer VA, Barnett B, Cook A, Davis R, Dunlop J, Ellis R, Morris K, van Leewen S (2015) Research priorities for the conservation and management of the northern quoll (Dasyurus hallucatus) in the Pilbara region of Western Australia. Submitted to Australian Journal of Mammalogy: 26.

Cramer VA, Dunlop J, Davis R, Ellis R, Barnett B, Cook A, Morris K, van Leeuwen S (2016) Research priorities for the northern quoll (Dasyurus hallucatus) in the Pilbara region of Western Australia. Australian Mammalogy. https://doi.org/10.1071/AM15005

Crimmins SM, Dobrowski SZ, Mynsberge AR (2013) Evaluating ensemble forecasts of plant species distributions under climate change. Ecological Modelling 266: 126–130. https://doi.org/10.1016/j.ecolmodel.2013.07.006

NatureMap (2015) Mapping Western Australia's Biodiversity. http://naturemap.dpaw.wa.gov.au/ [accessed 10/06/2015]

Doody JS, Soanes R, Castellano CM, Rhind D, Green B, McHenry C, Clulow S (2015) Invasive Toads Shift Predator-prey Densities in Animal Communities by Removing Top Predators. Ecology 96(9): 2544–2554. https://doi.org/10.1890/14-1332.1

Eco Logical (2012) Predictive species habitat modelling for four species across the Pilbara IBRA. Prepared for BHP Billiton Iron Ore Pty Ltd. ed. Biologic, Perth.

Elith J, Kearney M, Phillips S (2010) The art of modelling range-shifting species. Methods in Ecology and Evolution 1: 330–342. https://doi.org/10.1111/j.2041-210X.2010.00036.x

Elith J, Leathwick JR (2009) Species distribution models: ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics 40: 677. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Elith J, Phillips SJ, Hastie T, Dudík M, Chee YE, Yates CJ (2011) A statistical explanation of MaxEnt for ecologists. Diversity and Distributions 17: 43–57. https://doi.org/10.1111/j.1472-4642.2010.00725.x

Fithian W, Elith J, Hastie T, Keith DA (2015) Bias correction in species distribution models: pooling survey and collection data for multiple species. Methods in Ecology and Evolution 6: 424–438. https://doi.org/10.1111/2041-210X.12242

Grenouillet G, Buisson L, Casajus N, Lek S (2011) Ensemble modelling of species distribution: the effects of geographical and environmental ranges. Ecography 34: 9–17. https://doi.org/10.1111/j.1600-0587.2010.06152.x

Guillera-Arroita G, Lahoz-Monfort JJ, Elith J, Gordon A, Kujala H, Lentini PE, McCarthy MA, Tingley R, Wintle BA (2015) Is my species distribution model fit for purpose? Matching data and models to applications. Global Ecology and Biogeography 24: 276–292. https://doi.org/10.1111/geb.12268

Guisan A, Thuiller W (2005) Predicting species distribution: offering more than simple habitat models. Ecology letters 8: 993–1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x

Hamilton SH, Pollino CA, Jakeman AJ (2015) Habitat suitability modelling of rare species using Bayesian networks: Model evaluation under limited data. Ecological Modelling 299: 64–78. https://doi.org/10.1016/j.ecolmodel.2014.12.004

Hijmans RJ (2012) Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. Ecology 93: 679–688. https://doi.org/10.1890/11-0826.1

How RA, Spencer PBS, Schmitt LH (2009) Island populations have high conservation value for northern Australia's top marsupial predator ahead of a threatening process. Journal of Zoology 278: 206. https://doi.org/10.1111/j.1469-7998.2009.00569.x

Jetz W, Freckleton RP (2015) Towards a general framework for predicting threat status of data-deficient species from phylogenetic, spatial and environmental information. Philosophical Transactions of the Royal Society of London B: Biological Sciences 370: 20140016. https://doi.org/10.1098/rstb.2014.0016

Kramer-Schadt S, Niedballa J, Pilgrim JD, Schröder B, Lindenborn J, Reinfelder V, Stillfried M, Heckmann I, Scharf AK, Augeri DM (2013) The importance of correcting for sampling bias in MaxEnt species distribution models. Diversity and Distributions 19: 1366–1379. https://doi.org/10.1111/ddi.12096

Manthey JD, Campbell LP, Saupe EE, Soberón J, Hensz CM, Myers CE, Owens HL, Ingenloff K, Peterson AT, Barve N (2015) A test of niche centrality as a determinant of population trends and conservation status in threatened and endangered North American birds.

McKenzie N, Van Leeuwen S, Pinder A (2009) Introduction to the Pilbara biodiversity survey, 2002–2007. Records of the Western Australian Museum, Supplement 78: 3–89. https://doi.org/10.18195/issn.0313-122x.78(1).2009.003-089

Oakwood M (1997) The ecology of the northern quoll, Dasyurus hallucatus. Canberra: ANU.

Oakwood M (2004) The effect of cane toads on a marsupialcarnivore, the northern quoll, Dasyurus hallucatus. Envirotek: Ecological Research, Survey and Education.

Oakwood M (2008) Northern Quoll. In: Van Dyck S, Strahan R (Eds) The Mammals of Australia. Reed New Holland, Sydney, 57–59.

Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. Ecological Modelling 190: 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

Phillips SJ, Dudík M (2008) Modeling of Species Distributions with Maxent: New Extensions and a Comprehensive Evaluation. Ecography 31: 161–175. https://doi.org/10.1111/j.0906-7590.2008.5203.x

Phillips SJ, Dudík M, Elith J, Graham CH, Lehmann A, Leathwick J, Ferrier S (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications 19: 181–197. https://doi.org/10.1890/07-2153.1

Pollock AB (1999) Notes on status, distribution and diet of Northern Quoll Dasyurus hallucatus in the Mackay-Bowen area, mideastern Queensland. Australian Zoologist 31: 388–395. https://doi.org/10.7882/AZ.1999.040

Radosavljevic A, Anderson RP (2014) Making better Maxent models of species distributions: complexity, overfitting and evaluation. Journal of Biogeography 41: 629–643. https://doi.org/10.1111/jbi.12227

Ramanaidou E, Morris R (2010) A synopsis of the channel iron deposits of the Hamersley Province, Western Australia. Applied Earth Science: Transactions of the Institutions of Mining and Metallurgy: Section B 119: 56–59. https://doi.org/10.1179/037174510x12853354810624

Revelle W (2014) psych: Procedures for psychological, psychometric, and personality research. Northwestern University, Evanston, Illinois: 165.

Schmitt LH, Bradley AJ, Kemper CM, Kitchener DJ, Humphreys WF, How RA (1989) Ecology and physiology of the northern quoll, *Dasyurus hallucatus* (Marsupialia, Dasyuridae), at Mitchell Plateau, Kimberley, Western Australia. Journal of Zoology 217: 539–558. https://doi.org/10.1111/j.1469-7998.1989.tb02510.x

Spencer P, How R, Hillyer M, Cook A, Morris K, Stevenson C, Umbrello L (2013) Genetic analysis of northern quolls from the Pilbara Region of Western Australia. Year one – final report. Unpublished report prepared for Department of Sustainability, Environment, Water, Population and Communities, Canberra and Department of Parks and Wildlife. Perth Western Australia.

Spencer PBS, How RA, Schmitt LH (2010) The Northern Quoll Population on Koolan Island: Molecular and Demographic Analysis. Version 4 ed. Mount Gibson Iron Limited, Perth, WA.

Stockwell DR, Peterson AT (2002) Effects of sample size on accuracy of species distribution models. Ecological Modelling 148: 1–13. https://doi.org/10.1016/S0304-3800(01)00388-X

Syfert MM, Smith MJ, Coomes DA (2013) The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. PloS one 8: e55158. https://doi.org/10.1371/journal.pone.0055158

Thackway R, Cresswell I (1997) A bioregional framework for planning the national system of protected areas in Australia. Natural Areas Journal 17: 241–247.

Thuiller W, Georges D, Engler R (2013) biomod2: ensemble platform for species distribution modeling. R package version 3.0. 3. http. https://cran.r-project.org/web/packages/biomod2

Van Gils H, Conti F, Ciaschetti G, Westinga E (2012) Fine resolution distribution modelling of endemics in Majella National Park, Central Italy. Plant Biosystems-An International Journal Dealing with all Aspects of Plant Biology 146: 276–287. https://doi.org/10.1080/11263504.2012.685194

Westerman M, Woolley PA (2016) Comment on a research note reporting new populations of the northern quoll in Western Australia. Australian Mammalogy 38: 124–126. https://doi.org/10.1071/AM15024

Williams KJ, Belbin L, Austin MP, Stein JL, Ferrier S (2012) Which environmental variables should I use in my biodiversity model? International Journal of Geographical Information Science 26: 2009–2047. https://doi.org/10.1080/13658816.2012.698015

Woinarski JCZ (2010) Monitoring indicates rapid and severe decline of native small mammals in Kakadu National Park, northern Australia. Wildlife Research 37: 116. https://doi.org/10.1071/WR09125

Woinarski JCZ, Burbidge AA, Harrison PL (2015) Ongoing unraveling of a continental fauna: Decline and extinction of Australian mammals since European settlement. Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.1417301112

Yackulic CB, Chandler R, Zipkin EF, Royle JA, Nichols JD, Campbell Grant EH, Veran S (2013) Presence-only modelling using MAXENT: when can we trust the inferences? Methods in Ecology and Evolution 4: 236–243. https://doi.org/10.1111/2041-210x.12004

Supplementary material 1

GIS data sets used in variable assessments and map of Pilbara vegetation systems
Authors: Shaun W. Molloy, Robert A. Davis, Judy A. Dunlop, Eddie J.B. van Etten
Data type: distribution data

Supplementary material 2

Full readout for the MaxEnt northern quoll SDM
Authors: Shaun W. Molloy, Robert A. Davis, Judy A. Dunlop, Eddie J.B. van Etten
Data type: statistical data

Supplementary material 3

Weighted mean SDMs for individual algorithms and evaluation statistics (biomod2)
Authors: Shaun W. Molloy, Robert A. Davis, Judy A. Dunlop, Eddie J.B. van Etten
Data type: statistical data

Applying surrogate species presences to correct sample bias in species distribution models; a case study using the Pilbara population of the Northern Quoll.

Supplementary data file 1:

GIS data sets used in variable assessments and map of Pilbara vegetation systems.

Table 1. GIS data sets used in variable assessments. Bold indicates data-base and URL and italics indicate derived data.

WorldClim (http://www.worldclim.org/)
BIO1 = Annual Mean Temperature
BIO2 = Mean Diurnal Range (Mean of monthly (max temp - min temp))
BIO3 = Isothermality (BIO2/BIO7) (* 100)
BIO4 = Temperature Seasonality (standard deviation *100)
BIO5 = Max Temperature of Warmest Month
BIO6 = Min Temperature of Coldest Month
BIO7 = Temperature Annual Range (BIO5-BIO6)
BIO8 = Mean Temperature of Wettest Quarter
BIO9 = Mean Temperature of Driest Quarter
BIO10 = Mean Temperature of Warmest Quarter
BIO11 = Mean Temperature of Coldest Quarter
BIO12 = Annual Precipitation
BIO13 = Precipitation of Wettest Month
BIO14 = Precipitation of Driest Month
BIO15 = Precipitation Seasonality (Coefficient of Variation)
BIO16 = Precipitation of Wettest Quarter
BIO17 = Precipitation of Driest Quarter
BIO18 = Precipitation of Warmest Quarter
BIO19 = Precipitation of Coldest Quarter
Climond (https://www.climond.org/)
BIO34=Mean moisture index of warmest quarter
Climate Change in Australia (CSIRO) (http://www.climatechangeinaustralia.gov.au/en/)
Relative Humidity Wettest Quarter
ESRI Online Databases (http://www.esri.com/software/arcgis/arcgisonline/arcgis-open-data)
Background mapping, i.e. topographic imagery, boundaries and

placenames
Landgate (https://www2.landgate.wa.gov.au/bmvf/app/waatlas/)
Western Australian Fire Frequency
Pastoral Property (vesting)
Modis (https://earthdata.nasa.gov/)
Fire Scar
Modis Burndate
Noramalised Digital Vegetation Index (NDVI)
Near Infrared Spectrography (NIRS)
Geoscience Australia (http://www.ga.gov.au/search/index.html#/)
Total Magnetic Intensity
Gravity Anomaly
Land Tenure
Digital Elevation Model
Slope
Ruggedness
Water Courses
Euclidean Distance to Water Courses
Water Bodies
Euclidean Distance to Water Bodies
Geology
Land cover
Hydrology of Australia
Soils Mapping of Australia
Naturemap (http://naturemap.dpaw.wa.gov.au/)
NQ Presences CWR Terrestrial Mammal Presences
NQ Absences
Unites States Geological Survey (USGS) (http://earthexplorer.usgs.gov/)
Landsat Mosaic
NDVI Colourised
Department of Agriculture and Food WA (DAFWA) (https://www.agric.wa.gov.au/land-use-planning/maps-and-data)
Veg= Beards Vegetation Associations

Vegag= Department of Agriculture and Food Western Australia Vegetation Mapping (Rangelands)

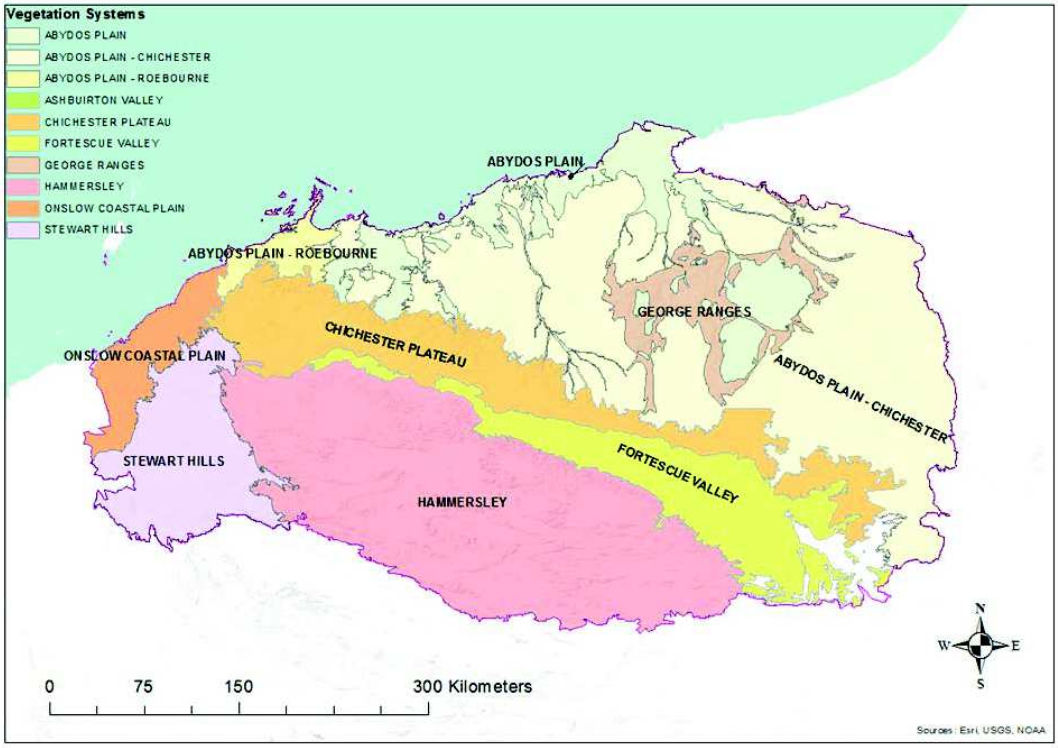


Figure 1. Pilbara vegetation systems. Taken from the Pilbara Pre-European Vegetation data set, Courtesy of the Western Australian Department of Parks and Wildlife

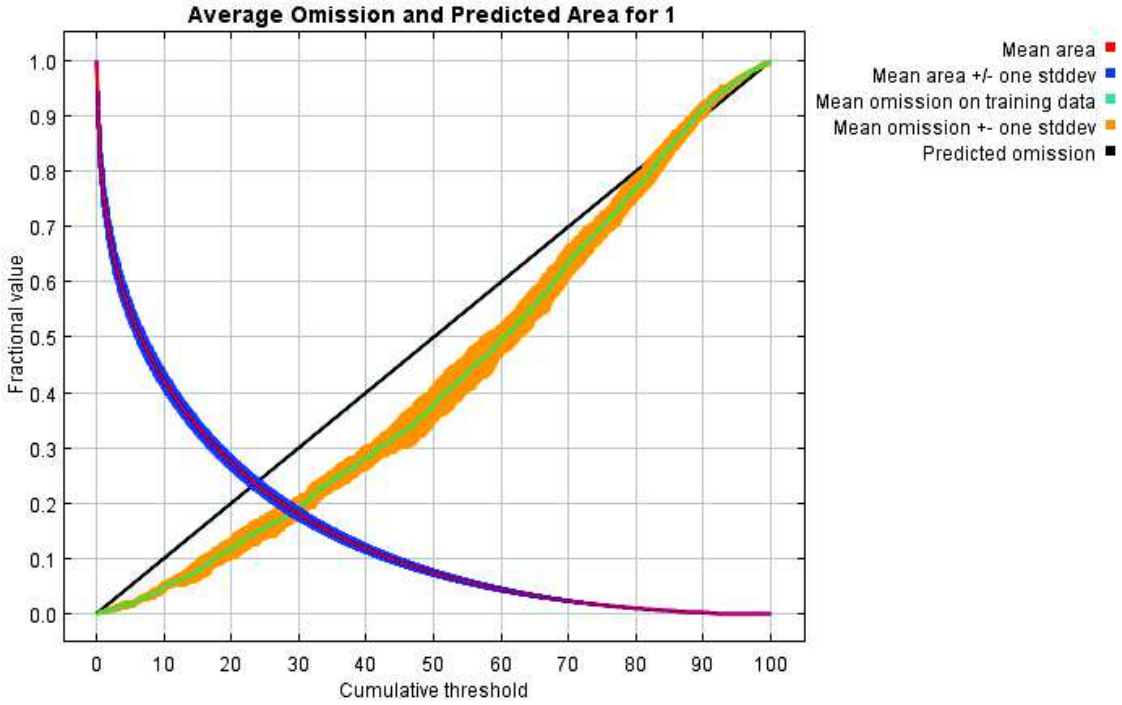
Applying surrogate species presences to correct sample bias in species distribution models; a case study using the Pilbara population of the Northern Quoll.

Supplementary data file 2:

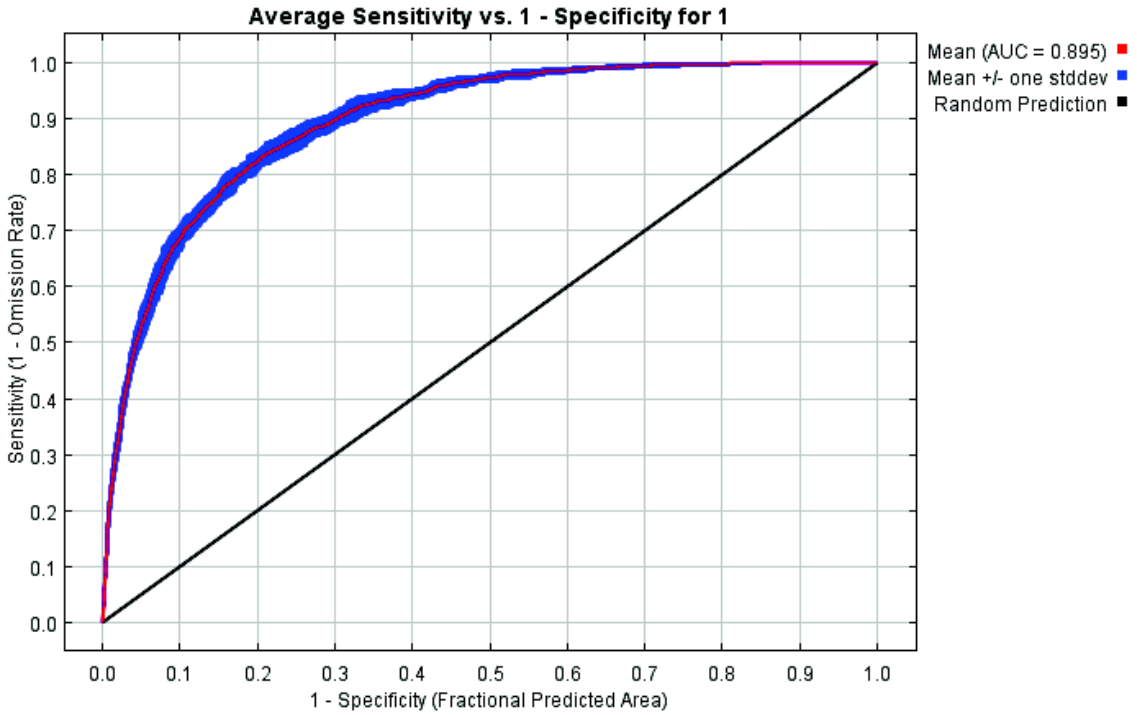
Full readout for the MaxEnt northern quoll SDM.

Analysis of omission/commission

The following picture shows the training omission rate and predicted area as a function of the cumulative threshold, averaged over the replicate runs.

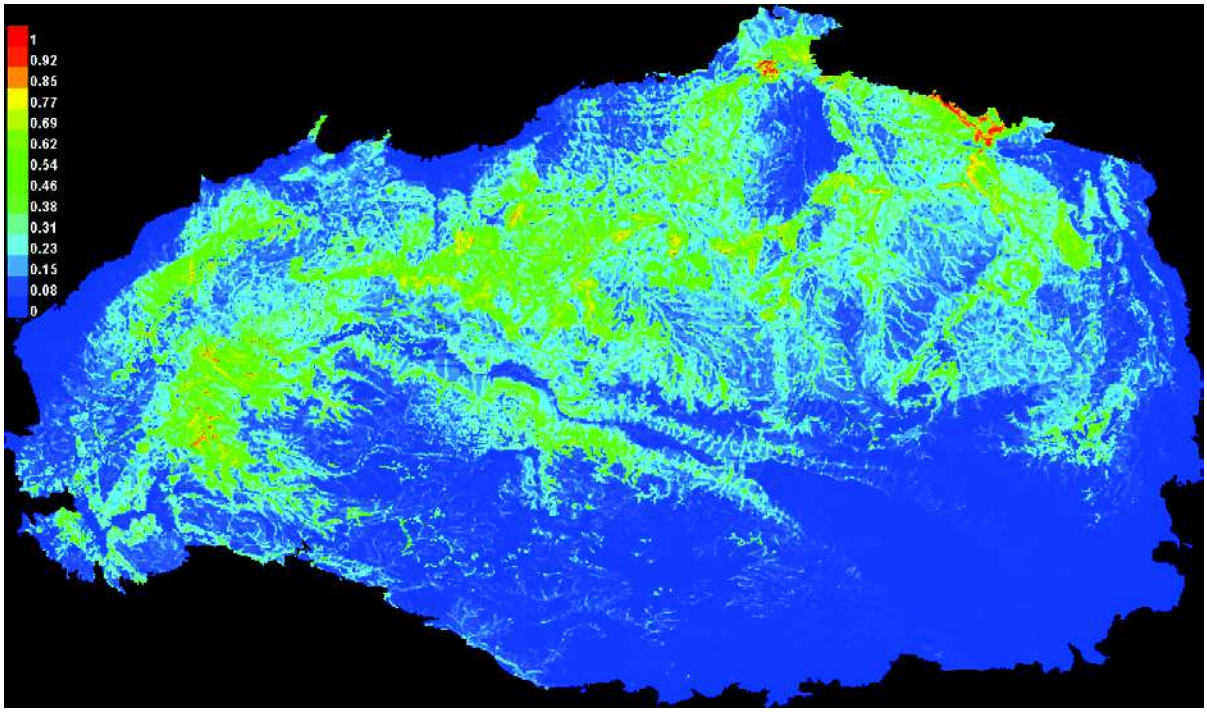


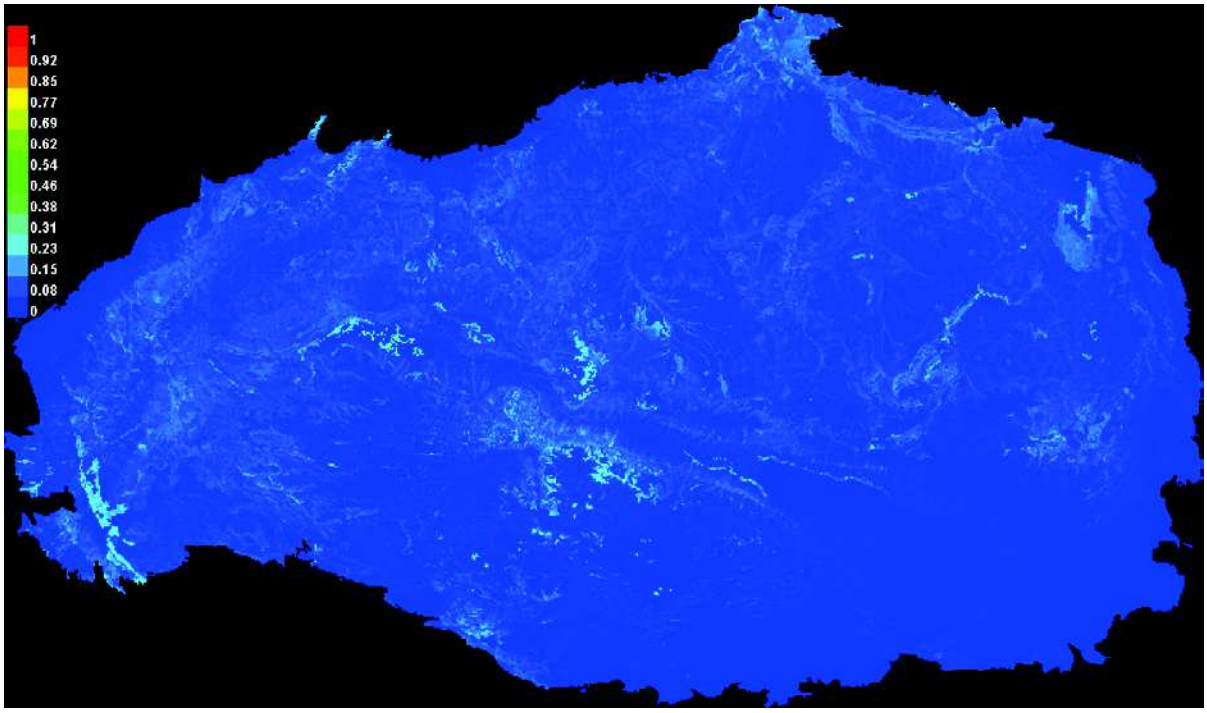
The next picture is the receiver operating characteristic (ROC) curve for the same data, again averaged over the replicate runs. Note that the specificity is defined using predicted area, rather than true commission (see the paper by Phillips, Anderson and Schapire cited on the help page for discussion of what this means). The average training AUC for the replicate runs is 0.895, and the standard deviation is 0.005.



Pictures of the model

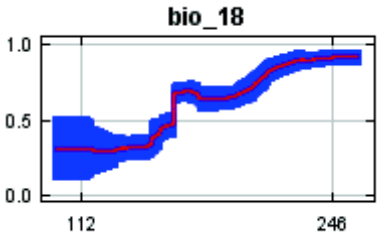
The following two pictures show the point-wise mean and standard deviation of the 10 output grids. Other available summary grids are min, max and median.

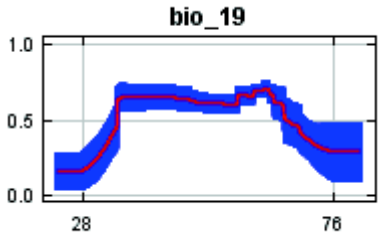


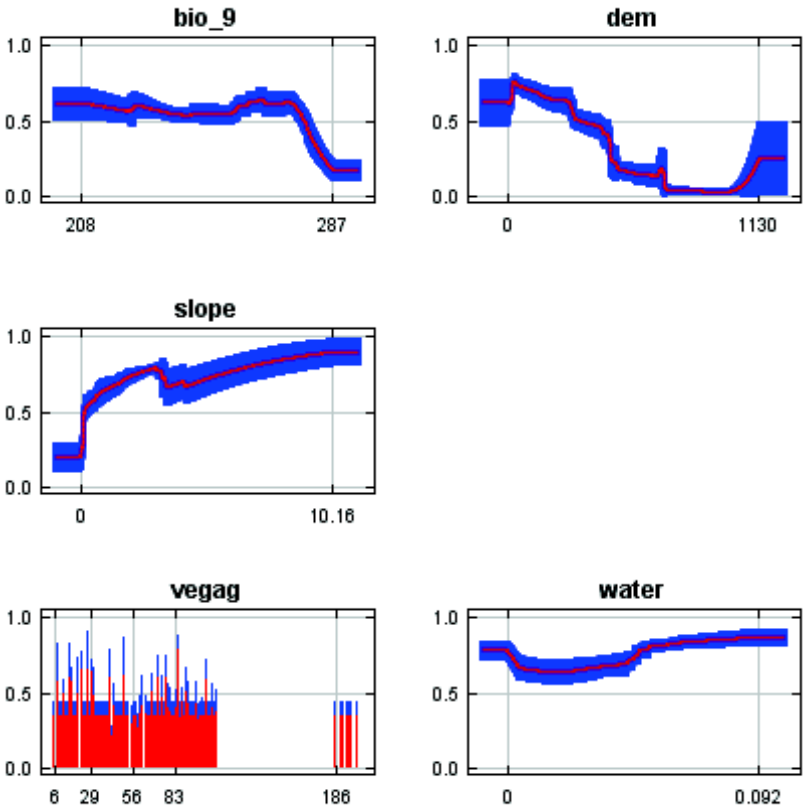


Response curves

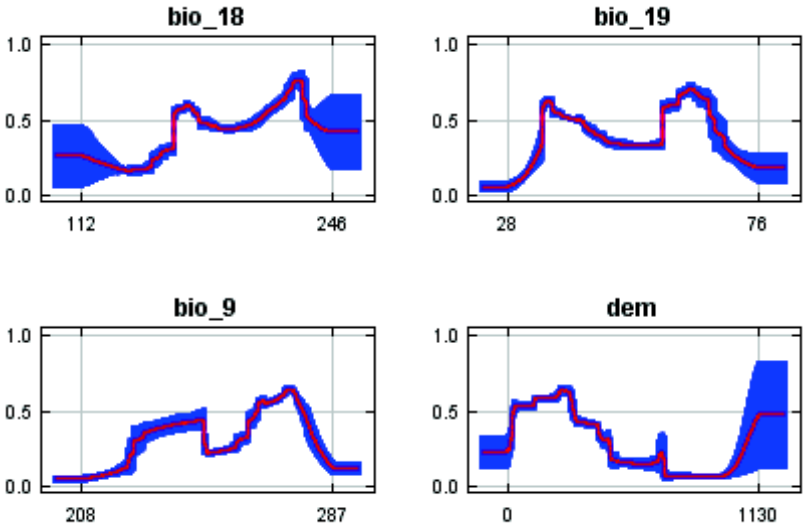
These curves show how each environmental variable affects the Maxent prediction. The curves show how the logistic prediction changes as each environmental variable is varied, keeping all other environmental variables at their average sample value. Click on a response curve to see a larger version. Note that the curves can be hard to interpret if you have strongly correlated variables, as the model may depend on the correlations in ways that are not evident in the curves. In other words, the curves show the marginal effect of changing exactly one variable, whereas the model may take advantage of sets of variables changing together. The curves show the mean response of the 10 replicate Maxent runs (red) and and the mean +/- one standard deviation (blue, two shades for categorical variables).

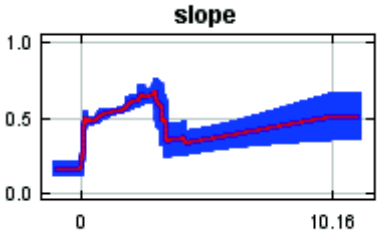


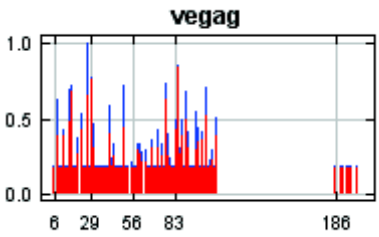


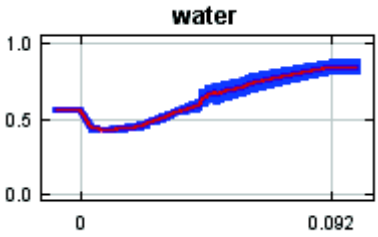


In contrast to the above marginal response curves, each of the following curves represents a different model, namely, a Maxent model created using only the corresponding variable. These plots reflect the dependence of predicted suitability both on the selected variable and on dependencies induced by correlations between the selected variable and other variables. They may be easier to interpret if there are strong correlations between variables.









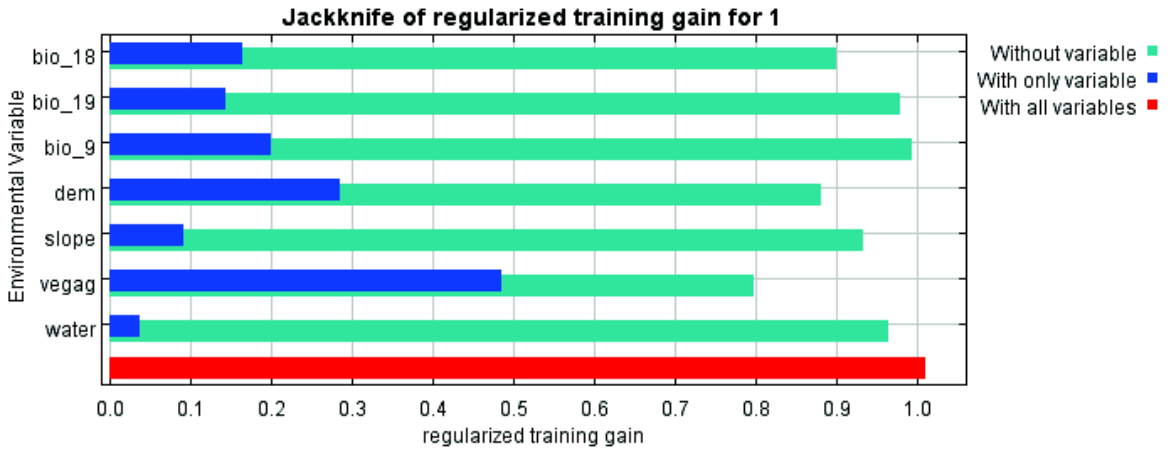
Analysis of variable contributions

The following table gives estimates of relative contributions of the environmental variables to the Maxent model. To determine the first estimate, in each iteration of the training algorithm, the increase in regularized gain is added to the contribution of the corresponding variable, or subtracted from it if the change to the absolute value of lambda is negative. For the second estimate, for each environmental variable in turn, the values of that variable on training presence and background data are randomly permuted. The model is reevaluated on the permuted data, and the resulting drop in training AUC is shown in the table, normalized to percentages. As with the variable jackknife, variable contributions should be interpreted with caution when the predictor variables are correlated. Values shown are averages over replicate runs.

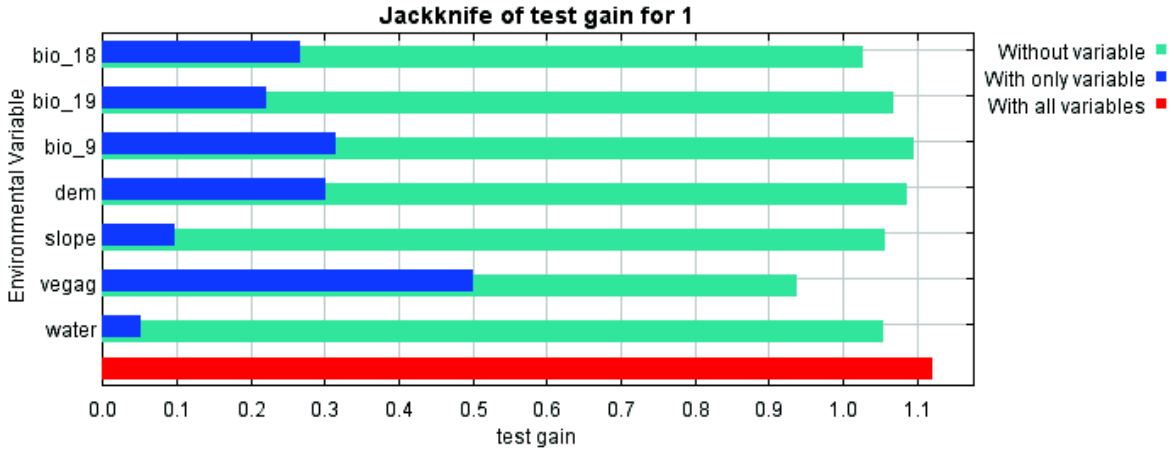
Variable	Percent contribution	Permutation importance
vegag	35	15.5
dem	26.2	37.4
bio_18	15.1	16.3
slope	11.5	14
bio_9	4.7	3.3
bio_19	4.1	9.3
water	3.4	4.2

The following picture shows the results of the jackknife test of variable importance. The environmental variable with highest gain when used in isolation is vegag, which therefore appears to

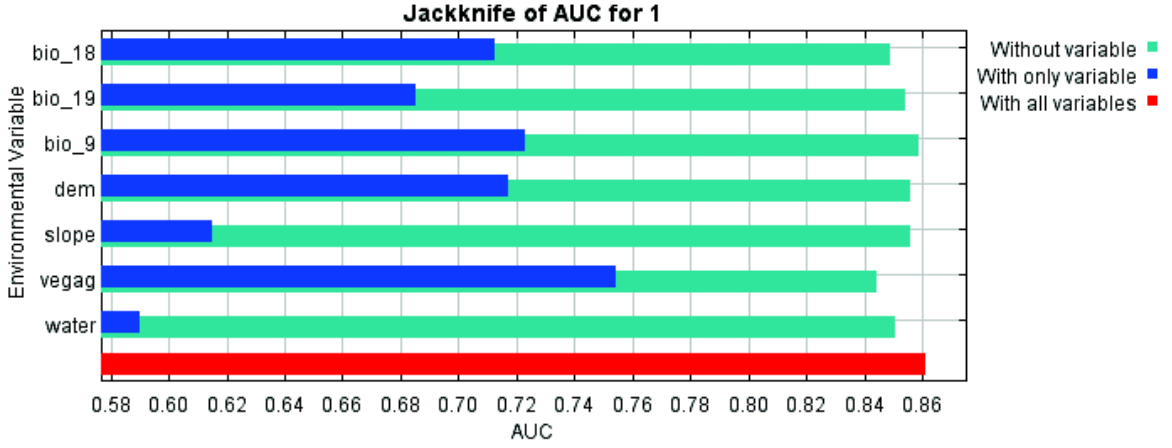
have the most useful information by itself. The environmental variable that decreases the gain the most when it is omitted is vegag, which therefore appears to have the most information that isn't present in the other variables. Values shown are averages over replicate runs.



The next picture shows the same jackknife test, using test gain instead of training gain. Note that conclusions about which variables are most important can change, now that we're looking at test data.



Lastly, we have the same jackknife test, using AUC on test data.



Applying surrogate species presences to correct sample bias in species distribution models; a case study using the Pilbara population of the Northern Quoll.

Supplementary data file 3:

biomod2: Weighted mean SDMs for individual algorithms and evaluation statistics.

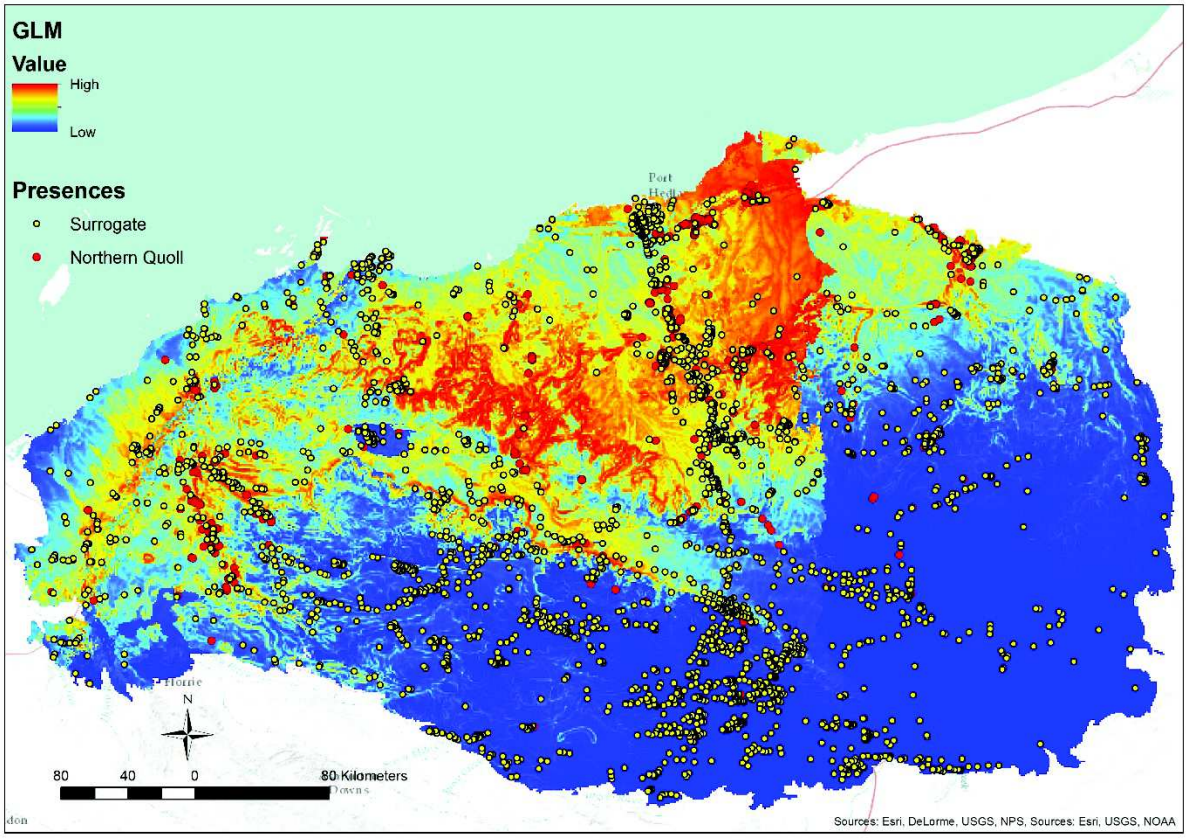


Figure 1. Weighted mean SDM for the Generalised Linear Model (GLM) algorithm.

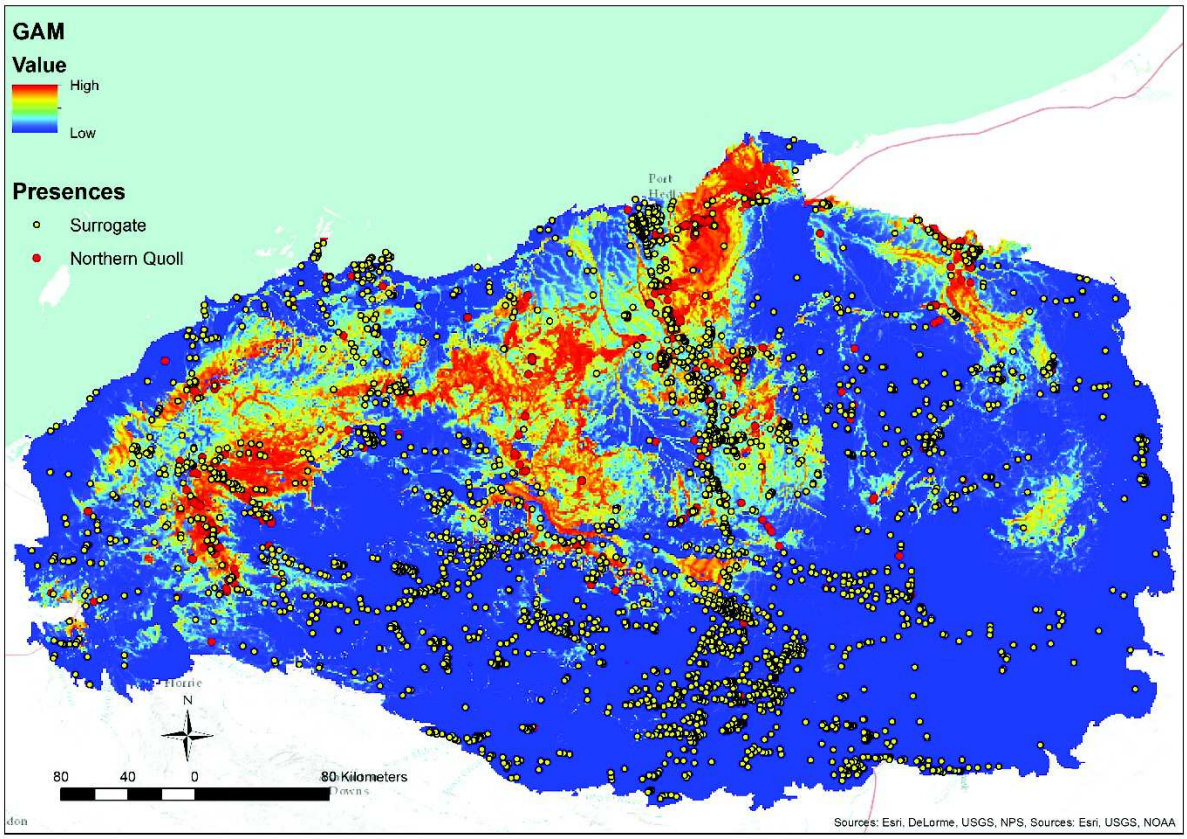


Figure 2. Weighted mean SDM for the Generalised Additive Model (GAM) algorithm.

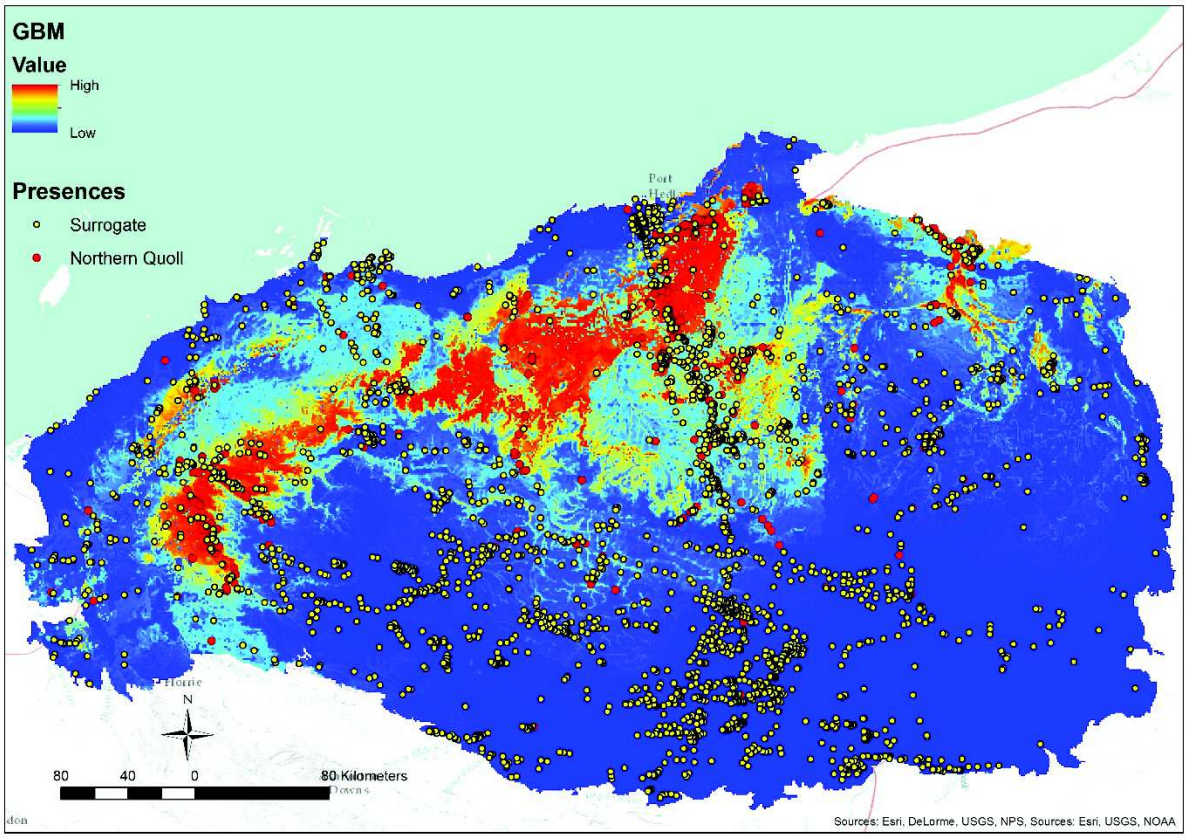


Figure 3. Weighted mean SDM for the Generalised Boosted Model (GBM) algorithm.

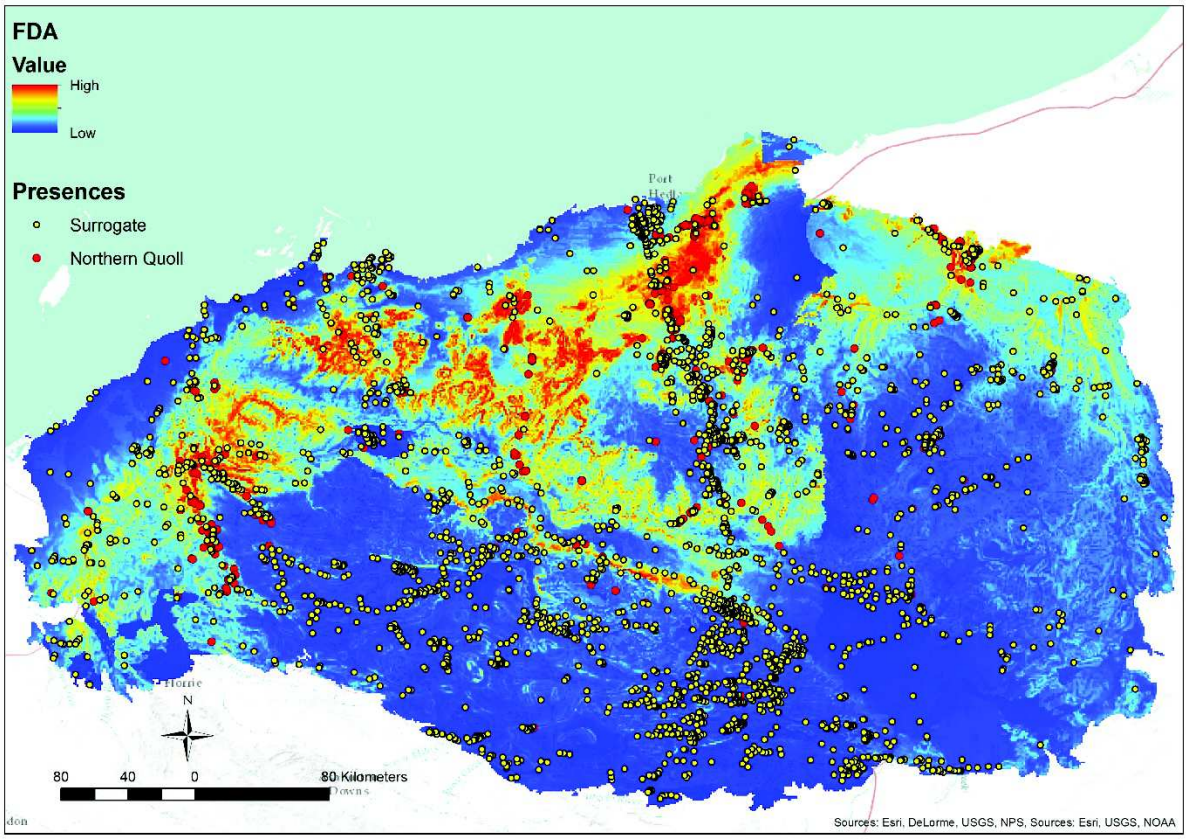


Figure 4. Weighted mean SDM for the Flexible Discriminate Analysis (FDA) algorithm.

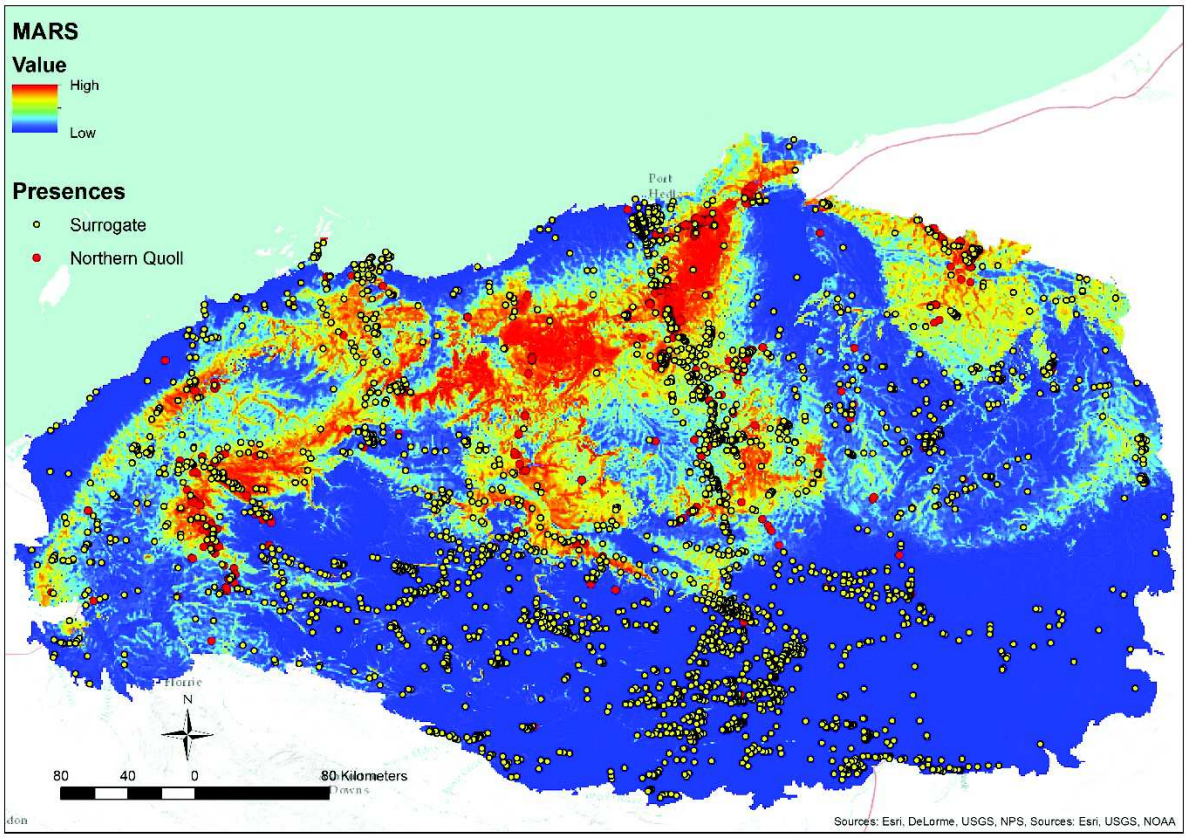


Figure 5. Weighted mean SDM for the Multiple Adaptive Regression Splines (MARS) algorithm.

Table 1, Evaluation scores for each algorithm by run using both ROC and TSS.

Score by Reciever Operator Characteristic (ROC)										
	RUN1	RUN2	RUN3	RUN4	RUN5	RUN6	RUN7	RUN8	RUN9	RUN10
GLM	0.827	0.820	0.841	0.828	0.837	0.818	0.823	0.819	0.823	0.837
GAM	0.920	0.903	0.904	0.908	0.922	0.901	0.911	0.900	0.917	0.911
GBM	0.933	0.917	0.934	0.930	0.937	0.920	0.925	0.915	0.934	0.927
FDA	0.867	0.863	0.874	0.863	0.869	0.833	0.854	0.841	0.867	0.876
MARS	0.900	0.886	0.906	0.895	0.913	0.891	0.884	0.887	0.899	0.892
Score by True Skill Statistic (TSS)										
	RUN1	RUN2	RUN3	RUN4	RUN5	RUN6	RUN7	RUN8	RUN9	RUN10
GLM	0.536	0.530	0.547	0.546	0.558	0.529	0.538	0.522	0.541	0.552
GAM	0.725	0.699	0.661	0.704	0.719	0.664	0.681	0.661	0.680	0.682
GBM	0.725	0.723	0.735	0.732	0.751	0.706	0.708	0.697	0.726	0.724
FDA	0.576	0.627	0.633	0.593	0.633	0.560	0.579	0.552	0.616	0.624
MARS	0.680	0.647	0.677	0.678	0.711	0.668	0.629	0.635	0.662	0.645

Table 2. Ensemble model (weighted mean) test results by run with mean values.

Run 1				
	Testing.data	Cutoff	Sensitivity	Specificity
TSS	0.71	531.00	84.50	86.16
ROC	0.92	530.50	84.50	86.16
Run 2				
	Testing.data	Cutoff	Sensitivity	Specificity
TSS	0.70	556.00	82.37	87.69
ROC	0.90	559.00	82.37	87.83
Run 3				
	Testing.data	Cutoff	Sensitivity	Specificity
TSS	0.71	556.00	84.50	86.66
ROC	0.92	654.50	79.94	91.33
Run 4				
	Testing.data	Cutoff	Sensitivity	Specificity
TSS	0.72	525.00	86.02	85.49
ROC	0.91	526.50	86.02	85.62
Run 5				
	Testing.data	Cutoff	Sensitivity	Specificity
TSS	0.74	585.00	84.50	89.93
ROC	0.92	592.50	84.50	90.19
Run 6				
	Testing.data	Cutoff	Sensitivity	Specificity
TSS	0.69	520.00	85.11	84.19
ROC	0.90	523.50	85.11	84.42
Run 7				

	Testing.data	Cutoff	Sensitivity	Specificity
TSS	0.68	621.00	78.12	89.93
ROC	0.91	629.50	78.12	90.23
Run 8				
	Testing.data	Cutoff	Sensitivity	Specificity
TSS	0.67	534.00	79.94	87.03
ROC	0.90	533.50	79.94	87.03
Run 9				
	Testing.data	Cutoff	Sensitivity	Specificity
TSS	0.70	567.00	82.37	87.63
ROC	0.92	585.50	81.76	88.46
Run 10				
	Testing.data	Cutoff	Sensitivity	Specificity
TSS	0.70	567.00	82.68	87.23
ROC	0.91	572.00	82.68	87.36
Mean				
	Testing.data	Cutoff	Sensitivity	Specificity
TSS	0.70	556.20	83.01	87.19
ROC	0.91	570.70	82.49	87.86

Table 3. Variable use by run/repetition and algorithm

Run 1					
	GLM	GAM	GBM	FDA	MARS
bio_19	0.006	0.080	0.002	0.090	0.000
bio_18	0.217	0.284	0.263	0.327	0.278
bio_9	0.234	0.268	0.398	0.00	0.305
dem	0.597	0.277	0.371	0.506	0.429
slope	0.182	0.164	0.189	0.144	0.210
water	0.050	0.053	0.008	0.000	0.052
vegag	0.037	0.157	0.019	0.196	0.042
Run 2					
	GLM	GAM	GBM	FDA	MARS
bio_19	0.000	0.091	0.008	0.061	0.000
bio_18	0.206	0.269	0.215	0.439	0.242
bio_9	0.285	0.261	0.396	0.000	0.315
dem	0.619	0.296	0.340	0.488	0.468
slope	0.145	0.155	0.172	0.140	0.216
water	0.036	0.041	0.005	0.000	0.044
vegag	0.044	0.173	0.029	0.192	0.039
Run 3					
	GLM	GAM	GBM	FDA	MARS
bio_19	0.010	0.075	0.004	0.104	0.000
bio_18	0.224	0.307	0.256	0.063	0.336
bio_9	0.239	0.327	0.419	0.550	0.271

dem	0.625	0.237	0.334	0.298	0.388
slope	0.136	0.076	0.195	0.179	0.195
water	0.051	0.044	0.007	0.043	0.057
vegag	0.043	0.215	0.017	0.101	0.000
Run 4					
	GLM	GAM	GBM	FDA	MARS
bio_19	0.009	0.089	0.004	0.098	0.000
bio_18	0.236	0.304	0.269	0.069	0.319
bio_9	0.246	0.310	0.396	0.496	0.230
dem	0.601	0.264	0.348	0.371	0.436
slope	0.183	0.124	0.223	0.232	0.289
water	0.031	0.039	0.004	0.026	0.042
vegag	0.038	0.179	0.023	0.020	0.000
Run 5					
	GLM	GAM	GBM	FDA	MARS
bio_19	0.014	0.073	0.002	0.049	0.000
bio_18	0.213	0.292	0.231	0.041	0.264
bio_9	0.235	0.303	0.410	0.402	0.310
dem	0.631	0.287	0.361	0.471	0.423
slope	0.161	0.090	0.189	0.114	0.208
water	0.028	0.035	0.003	0.000	0.032
vegag	0.041	0.184	0.025	0.069	0.027
Run 6					
	GLM	GAM	GBM	FDA	MARS
bio_19	0.010	0.085	0.002	0.039	0.000
bio_18	0.225	0.284	0.233	0.048	0.320
bio_9	0.289	0.327	0.406	0.422	0.262
dem	0.591	0.264	0.381	0.361	0.432
slope	0.177	0.111	0.23	0.127	0.243
water	0.031	0.041	0.003	0.000	0.042
vegag	0.038	0.197	0.017	0.120	0.000
Run 7					
	GLM	GAM	GBM	FDA	MARS
bio_19	0.005	0.099	0.003	0.000	0.000
bio_18	0.162	0.271	0.223	0.165	0.262
bio_9	0.226	0.288	0.404	0.359	0.300
dem	0.671	0.305	0.377	0.416	0.450
slope	0.182	0.126	0.221	0.194	0.187
water	0.019	0.029	0.002	0.000	0. 000
vegag	0.051	0.180	0.020	0.057	0.030
Run 8					
	GLM	GAM	GBM	FDA	MARS
bio_19	0.007	0.103	0.008	0.116	0.000
bio_18	0.227	0.314	0.255	0.066	0.316
bio_9	0.300	0.334	0.415	0.471	0.281

dem	0.578	0.247	0.310	0.380	0.434
slope	0.171	0.136	0.212	0.128	0.242
water	0.019	0.028	0.002	0.000	0.031
vegag	0.030	0.192	0.018	0.116	0.000
Run 9					
	GLM	GAM	GBM	FDA	MARS
bio_19	0.026	0.091	0.004	0.072	0.039
bio_18	0.227	0.308	0.234	0.235	0.317
bio_9	0.221	0.324	0.338	0.262	0.187
dem	0.649	0.280	0.426	0.562	0.492
slope	0.162	0.104	0.214	0.179	0.224
water	0.027	0.037	0.002	0.000	0.038
vegag	0.038	0.147	0.019	0.045	0.000
Run 10					
	GLM	GAM	GBM	FDA	MARS
bio_19	0.008	0.114	0.004	0.000	0.042
bio_18	0.210	0.318	0.202	0.239	0.32
bio_9	0.244	0.318	0.41	0.312	0.192
dem	0.651	0.246	0.35	0.422	0.468
slope	0.144	0.101	0.224	0.118	0.233
water	0.026	0.046	0.005	0.000	0.050
vegag	0.037	0.148	0.019	0.030	0.000