Species distribution modelling of *Aloidendron dichotomum* (quiver tree)

UNIVERSITY OF CAPE TOWN

DEPARTMENT OF STATISTICAL SCIENCES

Qobo Dube

supervised by:

Dr. Ian Durbach[1,2]

[1]University of Cape Town: Department of Statistical Sciences

[2]SEEC (Centre for Statistics in Ecology, Environment and Conservation)

Submitted in partial fulfilment of the degree of:

MASTER OF SCIENCE IN ADVANCED ANALYTICS AND DECISION SCIENCE

03 December 2017

Abstract

A variety of species distribution models (SDMs) were fit to data collected by a 15,000km road-side visual survey of *Aloidendron dichotomum* populations in the Northern Cape region of South Africa, and Namibia. We fit traditional presence/absence SDMs as well as SDMs on how proportions are distributed across three species stage classes (juvenile, adult, dead). Using five candidate machine learning methods and an ensemble model, we compared a number of approaches, including the role of balanced class (presence/absence) datasets in species distribution modelling. Secondary to this was whether or not the addition of species' absences, generated where the species is know not to exist have an impact on findings. The goal of the analysis was to map the distribution of *Aloidendron dichotomum* under different scenarios.

Precipitation-based variables were generally more deterministic of species presence or lack thereof. Visual interpretation of the estimated *Aloidendron dichotomum* population under current climate conditions, suggested a reasonably well fit model, having a large overlap with the sampled area. There however were some conditions estimated to be suitable for species incidence outside of the sampled range, where *Aloidendron dichotomum* are not known to occur.

Habitat suitability for juvenile individuals was largely decreasing in concentration towards Windhoek. The largest proportion of dead individuals was estimated to be on the northern edge of the Riemvasmaak Conservancy, along the South African/Namibian boarder, reaching up to a 60% composition of the population. The adult stage class maintained overall proportional dominance.

Under future climate scenarios, despite maintaining a bulk of the currently habitable conditions, a noticeable negative shift in habitat suitability for the species was observed. A temporal analysis of *Aloidendron dichotomum*'s latitudinal and longitudinal range revealed a potential south-easterly shift in suitable species conditions. Results were however met with some uncertainty as SDMs were uncovered to be extrapolating into a substantial amount of the study area.

We found that balancing response class frequencies within the data proved not to be an effective error reduction technique overall, having no considerable impact on species detection accuracy. Balancing the classes however did improve the accuracy on the presence class, at the cost of accuracy of the observed absence class.

Furthermore, overall model accuracy increased as more absences from outside the study area were added, only because these generated absences were predicted well. The resulting models had lower estimated suitability outside of the survey area and noticeably different suitability distributions within the survey area. This made the addition of the generated absences undesirable.

Results highlighted the potential vulnerability of *Aloidendron dichotomum* given the pessimistic, yet likely future climate scenarios.

Acknowledgements

I would like to express my gratitude to my supervisor Ian Durbach for the insights, assistance and ideas shared, throughout this project. Sam Jack, from the Plant Conservation Unit at the University of Cape Town provided photographs, species occurrence data, and along with Timm Hoffman were helpful collaborators. The advice and support from Vernon Visser, Res Altwegg, Bergit Erni, Theodor Stewart and the SEEC (Centre for Statistics in Ecology, Environment and Conservation) community are greatly appreciated. This project was sponsored by the National Research Foundation (NRF) and Department of Science Technology (DST). I would also like to thank my family, the Mavako clan for their encouragement and support during my studies, eDube, eNoko, Godlwayo, Mthembo, Shonge, Ntundlande. Finally, many thanks to MSc class of 2016.

Contents

List of Figures

List of Tables

INTRODUCTION

Background

Species occurrence patterns in the geographic space respond to environmental change, adjusting to follow suitable climatic conditions, or by remaining isolated in unchanged areas, potentially leading to extirpation or worse, global extinction. Species distribution models (SDMs) are used to predict or explain why a species occurs where it does in the geographic space based on known distributions of the environmental space.

As a consequence of technological advancements and increases in computational power and precision, there has been a shift in the tools and methodologies used to model species distributions. The process which previously was intensively qualitative has largely shifted towards quantitative models. The field of statistics and its varying learning methods and algorithms has been central in this shift in SDMs, as seen in a vast amount of literature (Zimmermann *et al.*, 2010; Elith and Leathwick, 2009; De'Ath, 2007; Guisan and Thuiller, 2005). SDMs aim to shed some more light on conservation biology and ecology, research areas of increasing importance because of global climate change.

Climatic factors (e.g. annual precipitation and minimum temperature of the coldest month) are used as inputs into the species distribution model. The factors are used to determine the possible presence or absence of a particular species based on observed occurrences. A major assumption of this type of analysis is that the species is at equilibrium with its current environment (Guisan and Thuiller, 2005).

The statistical learning models used are correlative in nature, and there is no guarantee whether the correlation structure between climate and species presence will remain constant across time (Dormann *et al.*, 2013).

In practice SDMs are implemented widely in ecology research areas. Some of the areas included are marine ecosystems, wildlife and for exploratory analysis (Zimmermann *et al.*, 2010). This contributes towards ensuring the continued survival of the species being modelled. With a trained statistical model, different input values can be used for the purpose of scenario planning, specifically around planning responses to possible effects of climate change.

The *Aloidendron dichotomum* (quiver tree) is a large, iconic tree species occurring in arid parts of Southern Africa and whose greatest threat is global climate change (Midgley *et al.*, 2007). Other threats include, but are not limited to, agriculture and overgrazing (Midgley *et al.*, 1997). Species presence is observed from the Western Cape in South Africa all the way into Namibia. The Karoo region, where the species occurs, is a semi-desert region recognised for its high level of biodiversity and a succulent plant community equipped to survive the arid ecosystem (Hoffman *et al.*, 2009). Observed evidence and research have proven droughts and increasing temperatures to have an effect on species prevalence (Jack *et al.*, 2016), considering the occurrence of the species in arid to semi-arid climates, where resources are scarce. The effect of climate change on the quiver tree therefore has the potential to be lasting (Jack *et al.*, 2014; Foden *et al.*, 2007).

Climate change is an issue faced by the world at large, with expected devastating biological effects. Climate change scenarios foresee an increase in the frequency of droughts in the Karoo which is a top "biodiversity hotspot" (Tadross *et al.*, 2005), potentially leading to widespread species extinction due to habitat loss. Midgley *et al.* (1997) documented the decline in the prevalence of *Aloidendron dichotomum*, a long-lived tree described as a cornerstone element of the climatic region, on account of its contribution as a moisture source for other forms of life like birds and mammals.

Research Aims

The intention of this study was to assess current knowledge of the *Aloidendron dichotomum*'s distribution by analysing the association between species occurrence patterns and climate under different scenarios, both present and future, in order to anticipate the possible effects of global climate change.

Data in this dissertation was initially collected to asses the effect of bioclimatic and geographic variables on the viability of *Aloidendron dichotomum* populations (Jack *et al.*, 2014). As a result, Jack *et al.* (2014) collected additional data, counts of the species in one of three stage classes. Individuals were classified either as one of juvenile, adult or dead stage classes (see Figure 2.1), referred to later in Section 3.1.1. This added information is potentially of interest when creating distribution models.

Distributional maps of the study area generated from SDMs were used to identify spatial patterns indicating species prevalence in its varying stage classes, or probabilistically where the stage classes were most likely to occur.

In addition to the assumption of equilibrium, SDMs specific to this study were limited by the quality of the data. *Aloidendron dichotomum* observations were an approximate count of individuals, recoded along a roadside survey, thus limiting the sampling range. Nevertheless, results reported at a species-wide and stage class level were used for recommendations, to identify any changes or improvements required in current conservation policy.

Research Objectives

The main research objective was to construct a map approximating the distribution of environmental conditions suitable *Aloidendron dichotomum* occurrence. Secondary to this, a question this study looked to answer was what the potential effect of climate change on the distribution of *Aloidendron dichotomum* in the Karoo biome is, given an overall predicted increase in temperature and drought prevalence in the region? To address this the study:

- Fit a number of machine/statistical learning models relating bioclimatic and geographic covariates to presence/absence data both for specific stage classes (juvenile, adult, dead) and live individuals (juvenile and adult).
- Fit a number of machine/statistical learning models relating bioclimatic and geographic covariates to specific stage class proportions (juvenile, adult, dead) given species presence.
- Construct maps for *Aloidendron dichotomum* distributions given the different target variables (suitability, proportions) under current and future climatic scenarios.

In the sections to follow, the study will give a review of SDM literature, addressing their benefits and limitations, and then proceed to give a description of the data and statistical learning algorithms used. Lastly, results from the statistical learning process will be reported and these findings will be used to address the current and possible future state of *Aloidendron dichotomum* distribution patterns.

LITERATURE REVIEW

Drought is a characteristic feature of much of southern Africa (Hoffman *et al.*, 2009). The Karoo, a "biodiversity hotspot", home to the *Aloidendron dichotomum* falls in this region. As a consequence of the major ecological and social impact, there has been extensive research on drought and climate change over the years, but at the time of publishing Hoffman *et al.* (2009) found little research on the impact of climate change on the succulent Karoo biome.

Aloidendron dichotomum are large trees that grow up to 9 metres, with shallow roots and succulent leaves, for rapid absorption and storage of water (Foden *et al.*, 2007). These characteristics help the trees which have a long average life span of at least 200 to 350 years (Vogel, 1974) and large geographical distribution survive droughts. Sessile species such as this one, which occur in dense populations are easier to characterise in terms of their environmental gradient profile, in contrast to mobile organisms (Elith and Leathwick, 2009), and because the tree occurs in such extreme conditions climate is bound to exert a greater influence on species presence. A sessile species is an immobile organism that is fixed in one place.

As a consequence of expected rising temperatures and longer drought episodes in a region already with minimal moisture, a shift in the distribution of the *Aloidendron dichotomum* is anticipated (Jack *et al.*, 2016; Foden *et al.*, 2007; Foden, 2002). Foden *et al.* (2007) investigated poleward species migration due to these rising levels, because of a suspicion of trailing edge (east-west) distribution extinction. Using repeat photography to support findings, evidence confirmed their suspicion of higher mortality on the trailing edge of the species' distribution, with a

decline of up to 7% per annum due to the prolonged water stress. The study further uncovered that the assumption of trailing edge extinction leading to leading edge expansion does not always hold, and was the first to do so. Analysis revealed the species' failure to expand polewards, bringing about a shrinking distribution because of dispersal lags. Climate change is therefore expected to result in a rise in mortality and population loss. Foden *et al.* (2007) went on to suggested *Aloidendron dichotomum* could make a potentially good example for climate change because of the findings, a point which has been recently contested by Jack *et al.* (2016)

Studies suggest that climate change is capable of generating high species turnover, potentially strong enough to lead to ecosystem disruption (Thuiller, 2004; Hoffman *et al.*, 2009). Species turnover is defined as the similarity between present and future distributions of species within a given area (Thuiller, 2004). A high species turnover results in a less similar distribution between the present and future. Seedling survival is a contributing factor towards turnover and is highly influenced by drought (Jordan and Nobel, 1979).

Population reproduction is a vital point alongside survival, and adult individuals can tolerate more extreme drought conditions because of greater water retention (Jack, 2012). Hoffman *et al.* (2009) investigated whether rainfall has decreased and if there has been a rise in drought incidence since 1900, and assessed its impact in the succulent Karoo. Hoffman *et al.* (2009) found no evidence of either a drop in precipitation levels or a rise in drought prevalence in the past century, but found a decrease in rain in the latter part of the 20th century. The paper concluded that flora in the Karoo is vulnerable to environmental change because of the region's moderate climatic history.

Climate change is happening faster than the rate of species adaptation evolutionary wise and this is true especially in long-lived species (Wiens *et al.*, 2009). Desert regions are therefore progressively developing into inhospitable areas

(a) *Juvenile* (b) *Dead* (c) *Adult*

Figure 2.1: *Aloidendron dichotomum* stage classes

for flora and fauna native to the region. Thus the need to model *Aloidendron dichotomum* habitat suitability. SDMs build upon existing ecological concepts and underpinnings, and require intelligent prior selection of meaningful features for superior insight (Elith and Leathwick, 2009). They are therefore a combination of expert-based and empirical modelling.

Over the years, many efforts have been made to improve SDMs. Guisan and Thuiller (2005), Zimmermann *et al.* (2010) and Elith and Leathwick (2009) document the history and status of SDMs, reviewing the achievements and addressing limitations. The purpose of a SDM is to uncover and quantify spatial relationships between a species or biota and the environment. Early ecologists only had access to local site conditions to build models, and as computer-based modelling surfaced, SDM popularity increased, because of the ease in which one could now compute complete geographical distributions (Guisan and Thuiller, 2005).

A common topic in all three of the previously mentioned papers (Zimmermann *et al.*, 2010; Elith and Leathwick, 2009; Guisan and Thuiller, 2005) is the need for

functionally relevant predictors, because of their direct impact on model output. Distribution models are more likely to perform especially well when the data is relevant and having been collected via a comprehensive survey. This results in far greater ecological insight and improved predictive power.

The geographic space is often 2 or 3 dimensional (*latitude, longitude* and *altitude*) in contrast to the environmental space which is multidimensional. As the number of dimensions increase so does the likelihood of overfitting, along with other complications such as algorithmic complexity and the need for significantly more observations. Hence the need for variable selection, a cure to the "curse of dimensionality". SDMs largely are guilty of ignoring geographic predictors (*latitude, longitude* and *altitude*), however some are geographic with the belief that this space is more important, or due to limited availability of environmental variables (Elith and Leathwick, 2009).

Two books by Margules and Austin (1990) and Verner *et al.* (1986) strongly influenced later developments in SDM theory and practice, evaluating old and new statistical methods, and interrogating sampling design among other things. From this also emerged questions about what variables are environmentally meaningful in the modelling process. Meier *et al.* (2011) points out that SDMs often relate species' distributions to abiotic predictors and assume biotic interactions have very little influence, or rather that they are only operational at finer spatial scales. Biotic factors are the reaction of a species to other living organisms in the ecosystem such as disease or predators, whereas abiotic factors are the non-living components of the ecosystem like temperature. Excluding biotic factors from the statistical learning process gives rise to a potential results bias (Araújo *et al.*, 2014).

In a prior study, Meier *et al.* (2010) used variance partitioning to determine the contribution of abiotic and biotic predictors towards variance explained by SDMs in explaining tree species' distributions. Analysis revealed there to be little overlap in the variance explained by the two classes of predictors, more especially in

optimal growing conditions. This reinforced the inclusion of biotic predictors when a suitable scale is available.

SDMs work on a snapshot of the current species' distribution, returning projections without taking past distributions into consideration. The certainty in distribution estimates is therefore dependant on the data collection process having captured the true optimal conditions in which the species occurs. This is where they often begin to deviate slightly from ecological theory, opening the approach up to some considerable uncertainties.

A crucial premise when building a SDM is that individuals are at equilibrium with their environment. Equilibrium therefore implies that all suitable habitats are occupied, however said locations may be unoccupied if they only recently became habitable, or possibly because of disturbances (Wiens *et al.*, 2009). This is a requirement for temporal and spatial projection, and because this is a single time point observation of the distribution, a pseudo-equilibrium with the environment is implied (Guisan and Thuiller, 2005). Optimal conditions for species occurrence today may not always be identical in the future, and pseudo-equilibrium rules out the possibility of a difference in the two states. As a result of this, evolution is ruled out as the effect of the past on a species' distribution is not considered, and its optimal environment is fixed.

Another example of a lack of integration of theory and practice is dispersal. This is a concept often but not always ignored in SDMs (Merow *et al.*, 2016; Guisan and Thuiller, 2005). Potential spaces will be occupied in the future under the pretence that species are able to disperse towards more suitable conditions and reproduce. In addition to their inherent dispersal capacity, the ability of individuals to disperse to suitable places is dependent on the landscape through which species must move (Wiens *et al.*, 2009). These spaces have more often than not been disrupted by natural capital consumption resulting in dispersal lags.

Due to increased efforts towards conservation as a result of global climate change, SDMs are being used more for prediction than ever before. As a consequence of the shift, accuracy of distribution models is crucial and this enables the use of "black box" methods where understanding is secondary. Models like this are very accurate but often fail to determine relationships and explain the underlying process, but the alternative is one that is too simple and leaves much unexplained (De'Ath, 2007). Prediction and explanation are two objectives that often do not go hand in hand.

Thuiller (2004) used a plant species dataset for the UK, and future climate scenarios to examine model uncertainty. Results uncovered problematic uncertainty, in that varying models with similar present day distributions may have completely different future projections. Ideally predictions must be robust and ignorant of the statistical learning method implemented. This otherwise compromises the utility of SDMs in data-driven decision making. After taking this into consideration, Elith *et al.* (2006) and Wiens *et al.* (2009) conclude there is no outright superior method. Although, both also established that machine learning algorithms perform best overall in sensitivity and specificity. Sensitivity is a measure of correct presence predictions and specificity is the proportion of correct absence predictions.

Elith *et al.* (2006) argue for the use of presence-only datasets using community models and machine learning, which were novel methods in species distribution modelling at the time of publishing, but still emphasise the importance of the collection of observed absences. A vast amount of occurrence data available from museum and herbarium collections are presence-only and there is often little knowledge on the sampling strategy, making it difficult to infer absences. Elith *et al.* (2006) found community methods to be somewhat favourable, especially in the case of rare species. Community methods make use of presence-only information of a species, and introduce presences of other community members into the model. Despite helping uncover relationships between species, community

methods along with other presence-only models will not be as well calibrated for ranking sites by their relative suitability (Elith *et al.*, 2006)

DATA AND METHODOLOGY

This section will first describe the data collection process and the sources of data involved, giving detail on the survey technique and data extraction. Moving on from there, the section shall motivate data manipulations to be done on training data and then go on to present the six candidate models and performance measures used. The modelling of the proportional distributions of stage classes will be presented and lastly, uncertainty estimation discussed.

Data Collection

This section describes the three main sources of data used to construct SDMs, and potentially helpful manipulations of this data that we assessed as part of the dissertation.

The response variables we model are spatially referenced counts of *Aloidendron dichotomum* obtained from a roadside survey, or transformations of these counts. Predictor variables are a range of bioclimatic variables extracted from the well-known BIOCLIM database, as well as a smaller number of biogeographic variables. These are the two primary data sources used to construct SDMs linking bioclimatic and biogeographic predictors to habitat suitability. In order to use the constructed SDMs to make projections of habitat suitability into the future, we need future projections of the input data i.e. future bioclimatic data, which is also available from the same BIOCLIM database.

While our main SDMs treat the data more or less "as is", we also assessed whether any improvement in model performance can be obtained by modifying the data before modelling. In particular, we assessed the effect of balancing the number of presences and absences, which has been reported to improve classification accuracy, and also of incorporating qualitative expert information by adding additional absences at points that were not explicitly sampled but that, according to experts, lie outside of the current known range of *Aloidendron dichotomum*.

Counts and stage class compositions of *Aloidendron dichotomum*

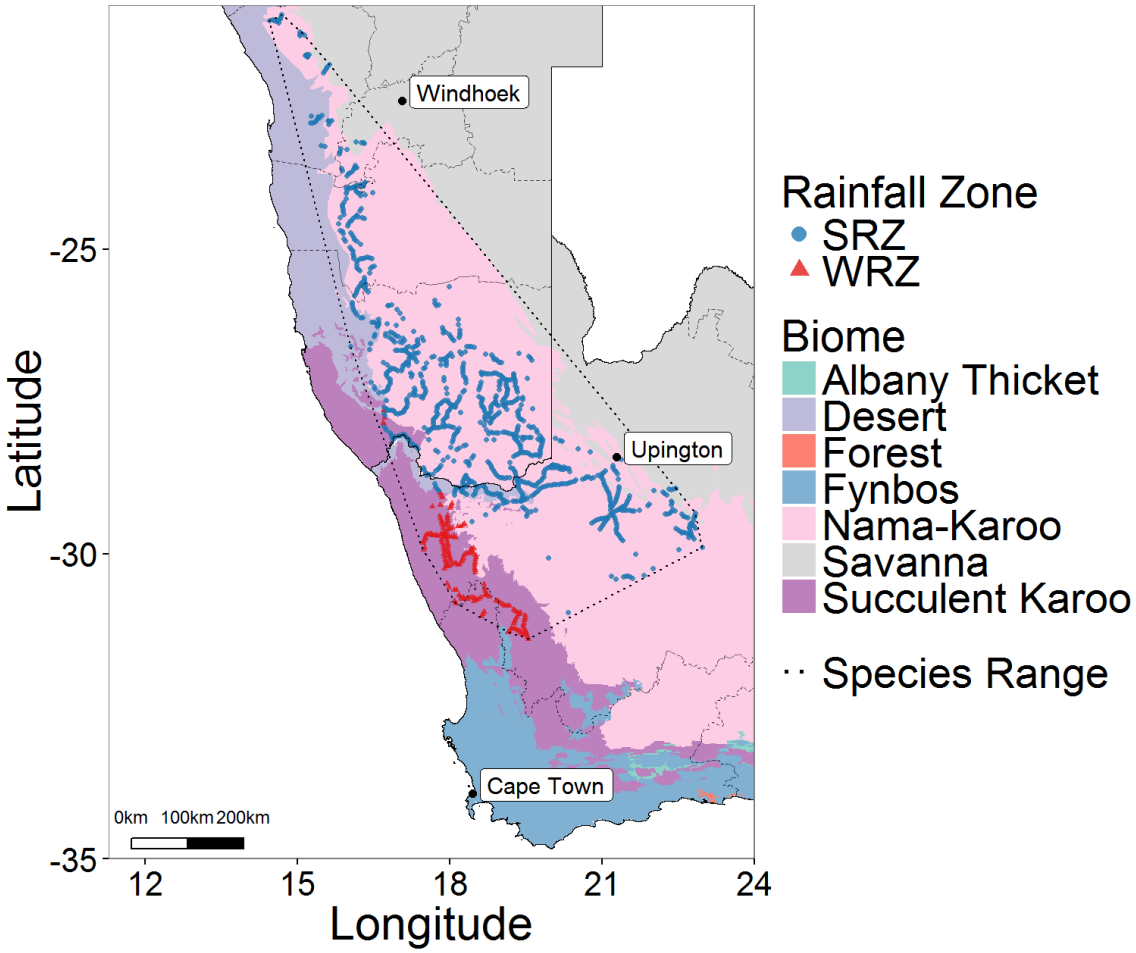


Figure 3.1: Study area indicating surveyed *Aloidendron dichotomum* presence records used in the analysis

An extensive 15000km roadside visual survey was conducted within the known range of *Aloidendron dichotomum* during 2008 and 2009. Figure 3.1 depicts the study area. Within each 5km transect, the total number of trees were estimated in a number of a pre-defined stage classes. Stage classes indicate discrete, biologically meaningful stages of tree development, and are based on a combination of size and architectural characteristics. The criteria, shown in Table 3.1, were selected by Jack *et al.* (2014) and are essentially given as fixed data for the purpose of our study. We used the coarser three-class categorisation, as also used in Jack *et al.* (2016).

Stage class	**Original stage class**	**Plant height (m)**	**Canopy Diameter (m)**	**Dichotomous branching nodes**	**Leaf rosettes**	**Reprod. moisture**
Juvenile	Juvenile	Generally <1.5	ca. 0.60	None	1	N
	Young adult	Variable, generally 2-3	1-1.5	1-3	2-8	Y
Adult	Mature adult	Variable, generally 3-5.5	2-3.5	6-10	>10	Y
	Senescent	Generally >4	2-4	8-12	>20	Y
Dead	Dead	n/a	n/a	n/a	n/a	n/a

Table 3.1: Stage class categorization criteria for *Aloidendron dichotomum* individuals (Jack *et al.*, 2014)

Trees were counted and assigned to stage classes by visual inspection with Nikon 8 x 42 binoculars. The beginning and end points of each transect were GPS marked (Garmin GPS60). Jack *et al.* (2016) reported that due to their smaller size, juvenile trees are more difficult to detect and may be under-represented in the dataset, but that they considered any bias to be independent of latitude and longitude. On the other hand, imperfect detection of fixed objects like trees, particularly in an arid landscape, might be expected to be less problematic than in

wildlife surveys. Our study does not model the detection process explicitly, and thus effectively assumes perfect detection.

Our basic data thus consist of counts of *Aloidendron dichotomum* in each of the juvenile, adult, and dead classes. Note that the dead category does not consider tree size and thus includes both dead juveniles and dead adults. Size at death is difficult to infer as trees decay over time, and we do not consider this here. The absence of trees in a transect was explicitly recorded i.e. with a zero, so that our data consists of both presences and absences.

For the purpose of constructing SDMs, we transformed the basic data into two types of response variables:

- Binary presence/absence of live trees: this was done by simply classifying the total number of juveniles plus adult trees as being equal to zero (absence) or greater than zero (presence). Note that dead trees did not contribute towards a "presence" assessment.
- Compositions across stage classes: here we divide the counts in each stage class within a transect by the total number of trees (dead or alive) in the transect.

We constructed independent SDMs for each of these two response types.

Extraction of potential bioclimatic predictors

Current bioclimatic variables for model fitting

Using the coordinate information, 16 variables from the BIOCLIM dataset (WorldClim, 2016) derived from monthly temperature and rainfall were extracted from the WorldClim database at a 30 arc-second (roughly $1km^2$) resolution. The data collected was averaged within the corresponding half-degree latitudinal bands to construct coupled climate profiles for the species (Jack *et al.*, 2016). These features are often used in species distribution modelling and represent annual trends, seasonality and extreme or limiting environmental factors (WorldClim,

2016). Variables described in Table 3.2 include mean annual precipitation and minimum temperature of the coldest month.

Bioclim	Variable name	Description
BIO1	MAT	Annual Mean Temperature
BIO4	Tseason	Temperature Seasonality (standard deviation *100)
BIO5	TmaxWmth	Max Temperature of Warmest Month
BIO6	TminCmth	Min Temperature of Coldest Month
BIO8	TwetQ	Mean Temperature of Wettest Quarter
BIO9	TdryQ	Mean Temperature of Driest Quarter
BIO10	TwarmQ	Mean Temperature of Warmest Quarter
BIO11	TcoldQ	Mean Temperature of Coldest Quarter
BIO12	MAP	Annual Precipitation
BIO13	Pwetmth	Precipitation of Wettest Month
BIO14	Pdrymth	Precipitation of Driest Month
BIO15	Pseason	Precipitation Seasonality (Coefficient of Variation)
BIO16	PwetQ	Precipitation of Wettest Quarter
BIO17	PdryQ	Precipitation of Driest Quarter
BIO18	PwarmQ	Precipitation of Warmest Quarter
BIO19	PcoldQ	Precipitation of Coldest Quarter
	Alt	Altitude
	tri	Terrain Roughness Index

Table 3.2: Climate and digital elevation variables

Bioclimatic variables in the BIOCLIM dataset are highly correlated as many of derived from some function of precipitation and temperature e.g. values of these variables in a coldest/wettest/hottest season. Models derived from datasets with well correlated variables may lead to erroneous results and variable selection is done to reduce the effects of multicollinearity (Garg and Tai, 2013). Correlated predictors are less of a problem where accurate prediction is the main goal, as it is in our application. However, as we do attempt some interpretation of variable

effects through partial dependence plots, we carried out a simple initial screening procedure to exclude variables that were highly correlated (above 0.7) with either mean annual temperature or mean annual precipitation. This resulted in the removal of five variables (see Table 3.3).

Excluded variable	Highest Correlation	Correlation
TmaxWmth	MAT	0.74
TminCmth	TcoldQ	0.87
TwetQ	PColdQ	-0.84
TwarmQ	MAT	0.82
Pwetmth	MAP	0..83
Pdrymth	PdryQ	0.94
PwetQ	MAP	0.86
PwarmQ	MAP	0.70

Table 3.3: Summary table of excluded variables and their highest correlated included variables.

In addition, where pairs of variables showed absolute correlations greater than 0.8, we removed one of these variables, based on a discussion with biologists as to biological relevance. This resulted in the removal of three further variables (see Table 3.3). The removal of correlated bioclimatic variables means that interpretation of variable effects must be done with caution. For example, minimum temperature in the coldest month is highly correlated with average temperature in the coldest quarter (0.87, see Table 3.3), and we removed the former. When interpreting the association between, for example, habitat suitability and the average temperature in the coldest quarter, we should bear in mind that a very similar relationship is likely to exist between habitat suitability and the minimum temperature in the coldest month.

The subset chosen consisted of eight bioclimatic variables, annual mean temperature; temperature seasonality; mean temperature of the driest quarter;

mean temperature of coldest quarter; annual precipitation; precipitation seasonality; precipitation of driest quarter and precipitation of coldest quarter, to be specific.

Future bioclimatic variables for model fitting

Future climate projections used in this study for the years 2050 (average for 2041-2060), and 2070 (average for 2061-2080) are generated by the MPI-ESM-LR model accessed from the WorldClim database. This enabled the temporal exploration of species' distribution patterns to determine potentially suitable areas for the species, via mapping. The MPI-ESM is a comprehensive Earth-System Model consisting of component models for the ocean, atmosphere and land surface, coupled through the exchange of energy, momentum, water and important trace gases such as carbon dioxide (Max Planck Institute for Meteorology, 2016). This model specifically, was used because it has been established to be a fairly good predictor of precipitation and temperature compared to other models (Giorgetta *et al.*, 2013)

Representative concentration pathways (RCPs) are greenhouse gas concentration scenarios used in climate modelling and research, based on emissions in the years to come. Of the four available RCP trajectories named after a range of possible radiative forcing values in the year 2100 relative to pre-industrial values, based on a discussion with biologists RCP8.5 was chosen because it assumes the worst case scenario given current human activity. The other three (RCP2.6, RCP4.5 and RCP6.0) assume a decrease in greenhouse gas concentrations at some stage within the next century (Moss *et al.*, 2010).

In addition to the bioclimatic variables, two forms of elevation information were leveraged into the dataset. The first being an average transect altitude value and the second a terrain roughness estimate. The terrain roughness index as defined by Riley *et al.* (1999), is used to classify location surfaces on a scale of level to extremely rugged. The index value for a point is computed using the variance between its

elevation and that of its eight neighbouring points.

Assessing the effect of data manipulations

Balancing response class frequencies

Our primary analyses and results are reported for the data as described above. However, previous research has shown that the performance of some of the classification algorithms we use to fit SDMs is compromised if data are not roughly equally distributed between response classes (He and Garcia, 2009)

Our data has substantially more absences than presences and is thus imbalanced, suggesting that better classification performance may be obtained by balancing the data. This can be done in a number of ways (see e.g. He and Garcia (2009)) but a simple approach is to subsample from the larger (i.e. absence) class. This is a commonly used technique (Drummond *et al.*, 2003), whereby all presence records were kept and sampling without replacement was carried out on the absence class, to create a dataset with equal response class frequencies. Despite subsampling potentially ignoring useful data, the scheme is attractive because it changes the input data rather than the algorithms (Drummond *et al.*, 2003).

Investigating the effect of balancing the response class frequencies was done by fitting SDMs on a balanced dataset containing the same number of absences and presences. We repeated this procedure 100 times to generate average differences between the balanced datasets and the original imbalanced dataset.

Including information about the known range

We use our SDMs to estimate habitat suitability across a larger area than that covered by the roadside survey – roughly, areas in Namibia and South Africa that lie to the west of the $26°E$ line of longitude. *Aloidendron dichotomum* experts are confident that the area covered by the roadside survey traverses the known range of *Aloidendron dichotomum*, and that a random sample of points outside of the known

distribution would be extremely likely to return an absence observation. Under these circumstances, it seems reasonable to ask whether our estimates of habitat suitability (inside and outside the known range) are affected by the addition of absence points outside the study area. The addition of absence points outside of the known species range has been found to generate more defined potential distributions and return higher accuracy scores (Chefaoui and Lobo, 2008). Adding these generated absences however increases the chance of SDMs picking up larger scale differences only, as opposed to local variations (Senay *et al.*, 2013).

We test this manipulation by randomly sampling a variable number of points (100, 1000, 5000) outside of the study area (east of $26°E$), and appending these points (classified as absences) to the dataset. To ensure consistency in data collection, these points were randomly sampled along the road network. We repeat this procedure 100 times to generate 100 "augmented" datasets. We then fit SDMs on each augmented dataset and extract average differences between the augmented datasets and the original dataset.

Predictive modelling methods

This section describes the approaches used in the construction of SDMs. This primarily covers three areas

- Specifying the data to be used: As described in the previous section, we construct a number of SDMs that differ in the response variable and whether the data is treated as is or manipulated.
  Response variables: binary presence/absence (all trees, or within each stage class juveniles, adult, dead); compositions
  Data: original; balanced; with additional absences (100, 1000, 5000)
  Timeframe: current habitat suitability, predicted habitat suitability in 2050, predicted habitat suitability in 2070

- Choice of a predictive modelling approach: our main results employ an ensemble of popular SDM methods (neural networks, boosting, random

forests, support vector machines, MaxEnt). In a later section we assess model variability across these approaches.

- Incorporating uncertainty into the modelling approach: many SDM approaches based on machine learning do not include explicit assessments of uncertainty in the predictions. We incorporate uncertainty in two ways: by bootstrap resampling from the original dataset and obtaining estimated prediction intervals, and using MESS maps.

Ideally supplementary inferences can be made from the output of each SDM type, potentially enriching one's knowledge of the species. All the machine learning methods for each SDM type were then applied to the species' distribution, under present and future climate scenarios. SDMs were constructed in R (R Core Team, 2017) using the packages neuralnet (Fritsch and Guenther, 2016); gbm (Ridgeway, 2017); randomForest (Liaw and Wiener, 2002); e1071 (Meyer *et al.*, 2017); dismo (Hijmans *et al.*, 2017); mvtboost (Miller, 2016); randomForestSRC (Ishwaran and Kogalur, 2017)

The data as is was randomly split into two samples for training and testing of the SDMs. The training dataset was made up of 80% of surveyed locations, and the remaining 20% was reserved for reporting on performance statistics used to asses model performance. Unless stated otherwise, the data manipulations as described in the previous section were performed on the training input data alone, and the test set left as is.

Distribution models

This section gives a brief explanation on the algorithms used. Hastie *et al.* (2001) covers the algorithms used here and more in greater detail, and Franklin (2010) with their specific application in mapping species distributions. SDMs were ranked by their ability to identify species' presence or lack thereof. Simpler generalised linear models favour interpretability, whereas a random forest and support vector machine offer greater predictive power with less interpretability. These machine/statistical

learning algorithms work as a black box and have neither a specific functional form nor interpretable parameters. Although some understanding of variable effects can be gained by examining partial dependence plots, for example.

MaxEnt

MaxEnt is a modelling approach that estimates a species' geographic distribution by finding the probability distribution with maximum entropy from presence data (Phillips *et al.*, 2006).

Given the presence records, the covariates are used to draw inferences about the true unknown distribution, π. In this instance for example, the expected terrain roughness index value under the estimated distribution $\tilde{\pi}$ must equal the average terrain roughness index from the observed species presences, and likewise for all other covariates. Squaring covariates (quadratic features) along with linear covariates, would ensure equal variance between the estimated and empirical distribution, and products of covariates would ensure the observed covariance structure is maintained in $\tilde{\pi}$.

These inferences can be seen as constraints on the estimated distribution. $\hat{\pi}$ is one of many possible estimates that satisfy all constraints derived from prior knowledge about π. However, what makes it the best estimate is that it will have the highest entropy. The principle of maximum entropy assigns the probability distribution closest to uniform as the best estimate because it expresses maximum uncertainty while agreeing with prior information, making it the least biased estimate one can select. This reduces the amount of prior information inherent to the distribution. Use of any other estimate would amount to arbitrarily assuming information which, by hypothesis, is not available (Shore and Johnson, 1980).

The joint distribution of the covariates is estimated using Gibbs sampling given $\mathbf{x}_i$, which is a vector representing the environment at each presence location, where

$$p(y = 1|\mathbf{x}_i) \quad \approx \quad q_\lambda(\mathbf{x}_i) = \frac{1}{\mathrm{Z}_\lambda(\mathbf{x}_i)} \exp\left(\sum_{j=1}^{m} \lambda_j f_j(\mathbf{x}_i)\right) \tag{3.1}$$

and

$$\mathrm{Z}_\lambda(\mathbf{x}_i) = \sum_{i=1}^{n} \exp\left(\sum_{j=1}^{m} \lambda_j f_j(\mathbf{x}_i)\right) \tag{3.2}$$

Here, λ is the weight of the marginal distribution, $f()$, for each of the j variables. Equation 3.2 is a normalising constant that ensures the probabilities sum to 1. Even with just presence-only data, MaxEnt effectively captures the realised species' distribution (Phillips *et al.*, 2006; Elith *et al.*, 2011). MaxEnt was designed as a presence-background method and its use with presence-absence data here is non-standard but sometimes done (Thibaud *et al.*, 2014; Phillips and Dudík, 2008). Background points were randomly sampled from absence data and default MaxEnt settings used. In comparison to some other machine learning algorithms, MaxEnt has the added advantage of mathematical interpretability, however the model has an increased risk of overfitting (Merow *et al.*, 2013).

Boosted Model

Tree based methods partition the feature space into subspaces and then fit a model to each subspace to estimate the response variable. The optimal partition is one that minimises the error associated with estimated response variables, and this is determined using recursive binary splitting (Hastie *et al.*, 2001). Boosting is one tree based method that aggregates several decision trees for improved predictive performance. The error functions for regression and classification trees are often the residual sum of squares (RSS) and the Gini index respectively (Qu *et al.*, 2002).

Small trees with a set number of splits are grown sequentially where error values from tree b are taken as the modified input to grow tree $b + 1$ until there are B trees. Each one of the trees accounts for the variation in the target variable not explained by preceding trees (Elith *et al.*, 2008). As a result, the first trees

represent the strongest amount of variation in the data and have a greater impact on error term reduction. Ensuing trees pick up less and less information in the data.

If $r^{(0)} = y$, where $r^{(b)}$ and $\hat{r}^{(b)}$ are the residuals and predictions of the b^{th} fitted tree respectively, then for $b = 1, \ldots, B$ fits a tree with d splits on the training set $(X, r^{(b-1)})$ to get:

$$r^{(b)} = r^{(b-1)} - \lambda \hat{r}^{(b)} \tag{3.3}$$

λ is the learning rate of the algorithm and d implies the number of interactions in the model. The boosted tree predictions are given by the equation:

$$\hat{y} = y - r^{B} = \sum_{b=1}^{B} \lambda \hat{r}^{(b)} \tag{3.4}$$

If the number of trees B is too large there is an increased risk of overfitting (Elith *et al.*, 2008). Therefore it is necessary to perform cross-validation for the total number of trees of the boosted model.

Random Forest

A random forest is another form of a tree-based method that decorrelates B full trees grown on B bootstrapped training samples (Hastie *et al.*, 2001). The same splitting criterion as boosted models is used but at each split a random subset of $m < p$ covariates are chosen to grow the trees. By not being able to consider all covariates as split candidates, this decorrelates trees and reduces the bias of having a very significant covariate that would make all trees look similar.

For each independent observation, the resulting random forest prediction is the average value or the majority class across all B trees, for regression and classification trees respectively (Breiman, 2001). Although, using a random forest for regression is not ideal because the model cannot predict beyond the range of the target variable in training data. This can be problematic, especially when

modelling species prevalence for example.

Unlike boosting, a large B does not increase the risk of overfitting and because trees are grown independently of one another, they can be grown in parallel, making this a computationally efficient algorithm (Breiman, 2001).

Neural Network

A neural network is a regression or classification model that consists of a collection of processing units called neurons (Kaastra and Boyd, 1996). The neurons are organised into three main parts: the input layer, hidden layers and output layer, which are all connected by the weights. These weights can be interpreted as the strength of the relationship between layers and their components. The size of any layer is a count of the number of neurons in that layer. Each of the neurons process a signal from the previous layer using an activation function whose output is the input for the following layer. What differentiates classification and regression neural networks is the activation function in the output layer neurons. The activation function used here was the sigmoid activation function.

Input layer size is determined by the number of features in the input data plus a bias term and the output layer by the number of target output variables. Network architecture is a deterministic factor of model accuracy. Selecting too few nodes in hidden layers results in a poor approximation by the final model and too many nodes results in increased computation and model complexity, not to mention an increased risk of overfitting (Hastie *et al.*, 2001).

Weights are randomly initialised, with those closer to the optimal solution reducing the computation time. The weight update process works in two phases at each iteration, until convergence (Hecht-Nielsen, 1992). An observation is picked at random to get a prediction, which is a linear combination of all network weights and layer outputs. This is the feed forward operation.

With the prediction, an error term is computed to be used in the back propagation when adjusting neuron weights. Minimising the error term is done by computing the change in the error term with respect to all weights within the network. For each layer this can be broken down into a delta term and previous layer output. If δ_i^l is defined as the change in the error term with respect to the signal of neuron i in layer l,

$$\delta_1^L = 2(\hat{y}_r - y_r) \cdot \frac{e^{-s_1^L}}{\left(1 + e^{-s_1^L}\right)^2} \tag{3.5}$$

where δ_1^L corresponds to the output layer neuron. Thus the equation for δ_i^l is

$$\delta_i^{(l-1)} = x_i^{(l-1)}\left(1 - x_i^{(l-1)}\right) \sum_{j=1}^{d^l} \mathrm{w}_{ij}^l \delta_j^l \tag{3.6}$$

The delta of a neuron is a linear combination of the deltas of connected neurons in the subsequent layer and their respective weights, multiplied by a term dependent on the output of that neuron. Equation 3.7 (Riedmiller and Braun, 1993) below is used to update the weights

$$\mathrm{w}_{ij}^l = \mathrm{w}_{ij}^l - \eta x_i^{(l-1)} \delta_j^l \tag{3.7}$$

where η is the learning rate. Including a complexity cost function in the error minimisation problem, which increases with the number of hidden layer nodes, can be done to prevent the model from fitting noise or by cross-validating on η.

Support Vector Machine

SVMs are a popular classification and regression method that make use of an optimal hyperplane in a $p-1$ feature space to identify patterns in the data, with maximum separability between classes (Suykens and Vandewalle, 1999). The hyperplane is the best dividing line between the two closest points from either class. The objective of the SVM is to maximise the margin, which is the distance between these two points. Once a hyperplane has been identified, these two points are then termed the support vectors.

A large margin reduces the likelihood of a data point being on the wrong side of the hyperplane, resulting in a misclassification (Hastie *et al.*, 2001). Maximising the margin also leads to a less complex model, increasing the algorithm's ability to generalise and reduces the risk of overfitting.

A hard margin problem would require perfect classification of observations. Optimising a soft-margin SVM allows for margin violations and transforms the problem into the Lagrangian form:

$$\max \quad \mathcal{L}(\alpha) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} y_n y_m \alpha_n \alpha_m \mathbf{x}_n^T \mathbf{x}_m \tag{3.8}$$

$$\text{subject to:} \qquad 0 \leq \alpha_n \leq C \tag{3.9}$$

$$\mathbf{y}^T \boldsymbol{\alpha} = 0, \tag{3.10}$$

where C is a regularisation term, α_n are Lagrange multipliers, y_n a presence/absence indicator of a given observation and $\mathbf{x}_n$ a vector of covariate information. Often data are not linearly separable, so inputs are mapped into a higher dimensional space for optimal separability using a kernel (Scholkopf *et al.*, 1997). Here, a radial basis kernel is used, which transforms the feature space into infinite dimensions. The radial basis function is defined by:

$$K(\mathbf{x}\ \mathbf{x}') = exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2) \tag{3.11}$$

where γ is a tuning parameter. The process of choosing values C and γ is hyper-parameter optimisation and is done using a grid search. In addition to performing non-linear classification, SVMs can efficiently perform regression. The functional form is almost identical to Equations 3.8 through 3.10 above. A study by Nassiri *et al.* (2014) covers this in more detail.

Ensemble Model

Ensemble modelling is a solution to methodological uncertainty that comes with the use of SDMs. The rationale behind it is that each of the predictions of the previously mentioned distribution models represent a possible species occurrence state. Given

that the individual outcomes contain some independent information, aggregation is a well established variance reduction technique (Thuiller *et al.*, 2009; Araújo and New, 2007; Bates and Granger, 1969), which would give a more robust projection of the species' distribution, with a lower mean error than any of the individual possible present or future outcomes. Ensemble forecasts can therefore be derived using either the median or weighted or unweighted averages of the different predictions of a single location, given the environmental gradient.

Assessing model performance

Reporting is done using unweighted mean ensemble models because these average over all other models and thus will be less variable in their predictions than any one model (Elith *et al.*, 2010; Araújo and New, 2007)

Where the model input target variable is a binary presence/absence indicator, the model output is a species' suitability index, given the environment. Models whose output was habitat suitability were assessed using three performance measures: classification accuracy, the receiver operating characteristic (ROC) curve and the area under curve (AUC) statistic.

Suitability estimates from the test data sample were transformed into binary data indicating species presence or absence for the purpose of defining confusion matrices, in order to determine classification accuracies. Transforming the suitability measure into a binary variable would require a defined threshold suitability level to determine species presence or absence. Picking a threshold level at which a species is believed to be present depends on model application. For example, modelling rare or endangered species would require a more conservative (higher) threshold than for range impact analysis (Hughes *et al.*, 2008). A fixed threshold level of 0.5 was used in this study, which is the threshold most commonly used (Buckley *et al.*, 2010; Manel *et al.*, 1999).

The ROC curve, depicts binary classifier accuracy as its class separation

threshold is varied. The probability of false identification (false positive rare) is plotted on the X axis, against the probability of identification (true positive rate) on the Y axis, for each threshold value between zero and one.

Within the ROC space, at point (0,0) all predictions are assigned to the negative class (absences) and there is no false positive error. Conversely, at point (1,1) all observations are predicted to be positive classifications (presences). A perfect classifier has an ROC curve passing through (0,1) and a poor classifier is close to the diagonal line $x = y$, which is what is expected from chance alone.

The AUC statistic is a single value used to compare classifiers in the ROC space. AUC is the area under the ROC curve and is a portion of the area in the ROC space which is equal to a unit square. The value of a classifier's AUC statistic will always be between zero and one. A poor classifier producing a diagonal line, will have an AUC equal to 0.5. No realistic classifier even with random guessing should have an AUC less than 0.5 (Fawcett, 2004).

Variable importance based on the random forest model was used to determine the relative importance of bioclimatic and biogeographic variables when predicting species presence or absence. The measure was calculated as the total decrease in node impurities from splitting on a variable, averaged over all trees. In classification, the node impurity is quantified by the Gini index (Liaw and Wiener, 2002). The measure has been found to show a bias towards correlated predictor variables, underestimating the true importance, as correlation between variables increases (Archer and Kimes, 2008). This should not be as of great concern in this dissertation, because of the removal of highly correlated bioclimatic variables, mentioned in the previous section.

Compositional data analysis

This is the modelling of proportions of each stage class at potential species locations. The nature of the problem demanded a multivariate analysis approach for results to be meaningful, with reference to species stage classes. In

compositional data analysis, a composition vector $\mathbf{x} = [x_1, x_2 \dots x_D]$ is made up of D parts where x_d represents the d^{th} component of $\mathbf{x}$. These D parts must have a constant sum, usually 1 or 100 if expressed as percentages. In this study $D = 3$ and $\mathbf{x}_i$ the vector of proportions at transect i is equal to $[juv_i, adult_i, dead_i]$, where $juv_i, adult_i$ and $dead_i$ are the proportions of juvenile, adult and dead individuals in transect i, respectively. The compositional models were fit to the normalised counts from actual presence locations in the training data sample.

The negative bias problem defined by Jackson (1997), points out the inappropriateness of standard multivariate linear regression when modelling proportions, because of the constraint on the D part vector to sum to a constant. There are a number of approaches to modelling multivariate proportions but we followed Aitchison's (2003) additive log-ratio transformation. This involved moving the D-part composition into a $(D-1)$ dimensional space defined by:

$$\mathbf{r} = \left[\log\left(\frac{x_1}{x_D}\right), \log\left(\frac{x_2}{x_D}\right) \dots \log\left(\frac{x_{D-1}}{x_D}\right)\right], \tag{3.12}$$

where the first $(D-1)$ proportions are divided by the last component. The inverse transformation is defined by:

$$\mathbf{x} = C\left[\exp(r_1), \exp(r_2) \dots \exp(r_{D-1}), 1\right], \tag{3.13}$$

where C is a closure operation that makes the D-part composition $\mathbf{x}$ sum up to a constant. The procedure in Equation 3.12 is perturbation invariant, which means the order of components and choice of denominator has no effect. Transforming the data by changing the order of components would result in the same outcome (Aitchison, 2003).

There is a significant amount of research on compositional data analysis (Pawlowsky-Glahn *et al.*, 2015; Pawlowsky-Glahn and Buccianti, 2011; Aitchison and Egozcue, 2005) but a standing drawback of current methods is how they deal with components which are legitimately zero, and not because of measurement rounding. These are often termed essential zeros. The method can not handle stage

class absences by definition, due to the log-ratio transformation. A stage class is considered absent if its component has a value of zero in the compositional vector. Component-wise zeros were imputed using a simplified version of Martín-Fernández *et al.* (2003)'s multiplicative replacement strategy defined by:

$$a_d = \begin{cases} \delta & \text{if } x_d = 0 \\ \left(1 - \Sigma_{k|x_k=0}\delta\right) x_d & \text{if } x_d > 0, \end{cases} \tag{3.14}$$

where δ is the small constant value to replace component-wise zeros and a_d is the new component value. Non-zero components are reduced by a factor dependant on the number of imputations required. This is one of many non-parametric replacement strategies(e.g see Martín-Fernández and Thió-Henestrosa (2006); Martín-Fernández *et al.* (2003)), and despite the fact they all assume the zeros are rounded, they are the most workable techniques available. Other approaches for treating essential zeros have been proposed by Aitchison *et al.* (2003).

Additionally, compositional data analysis can not handle species absences, which would be represented by $\mathbf{x}_i = [0, 0, 0]$. Thus a hurdle model with one process generating the zeros and another generating the positive values was applied. The reasoning behind this was that species presence and absence is determined by a binomial model, separate to one that influences the magnitude of species proportions. Compositional SDM estimates were therefore dependant on the binary SDMs defined previously, whose outcome was the conditional distribution of proportions, given species presence. Compositional distributions were estimated for locations in the study area with predicted habitat suitability above the 0.5 threshold. These are locations where the species would be considered present.

Once the processes defined in Equations 3.14 and 3.12 have been applied to the compositional vectors and in that order, standard unconstrained multivariate analysis can be carried out (Aitchison and Egozcue, 2005). Only a subset of the models could be implemented to identify and estimate the three part compositions of *Aloidendron dichotomum* populations using multivariate regression. These

models were neural networks, random forests and boosted trees.

The Aitchison distance between an observation and its predicted proportion vector was used as an evaluation measure in the compositional data analysis. This is a D dimensional distance metric between the two vectors, defined by Aitchison (1983) as

$$\Delta_S(\mathbf{x},\mathbf{y}) = \left[\sum_{i=1}^{D}\left\{\ln\frac{x_i}{g(\mathbf{x})} - \ln\frac{y_i}{g(\mathbf{y})}\right\}^2\right]^{1/2} \tag{3.15}$$

where $g(\cdot)$ denotes the geometric mean of the components in each vector. Each pair of observed proportions and predictions thus has an individual Aitchison distance.

Uncertainty modelling

Two kinds of uncertainty assessment were used: standard deviations of estimated habitat suitability obtained by bootstrapped predictions, and multivariate environmental similarity surface maps that show where predictions were made in areas that are different to the range of observed bioclimatic variables in the survey zone.

Bootstrap resampling

Resampling of the training dataset was carried out to account for random variation when approximating true distributions or statistics of interest. Samples of the same size were randomly drawn with replacement from the original training dataset repeatedly. For each sample the estimated habitat suitability and statistics of interest were computed. Results reported are the respective means of the approximated distributions and statistics, and the standard deviation used as a measure of variability, specifically for habitat suitability.

The procedure was repeated 100 times to generate SDM misclassification/accuracy distributions and an average variable importance. Final

habitat suitability distributional maps are the mean of distributional estimates derived from 20 of the 100 bootstrapped datasets. Uncertainty maps are the standard deviation of suitability estimates from the same 20 bootstrapped datasets. Fewer datasets were used for suitability and uncertainty maps because of the computational cost of predicting suitability for the over two million data points within the study area.

MESS maps

The multivariate environmental similarity surface (MESS) as defined by Elith *et al.* (2010) is a visual representation of how similar a point is to a group of surveyed points, given a set of predictors. It takes a hyper-dimensional rectangle viewpoint by analysing environmental coverage one bioclimatic variable at a time and reporting as novel those conditions outside the defined space (Zurell *et al.*, 2012). Points considered novel are where SDMs would be extrapolating suitability estimates.

The similarity of a point P is calculated as follows (Elith *et al.*, 2010):

1. Let $\min_j$ and $\max_j$ be the minimum and maximum values respectively, of variable V_j over the surveyed points
2. Let p_j be the value of variable V_j at point P
3. Let f_j be the percentage of surveyed points whose value of $V_j < \mathrm{p}_j$
4. The similarity of P with respect to V_j is then:
   - $(\mathrm{p}_j - \min_j)/100(\max_j - \min_j)$ if $\mathrm{f}_j = 0$
   - $2\mathrm{f}_j$ if $0 < \mathrm{f}_j \leq 50$
   - $2(100 - \mathrm{f}_j)$ if $50 \leq \mathrm{f}_j < 100$
   - $(\max_j - \mathrm{p}_j)/100(\max_j - \min_j)$ if $\mathrm{f}_j = 100$
5. The multivariate environmental similarity (MES) of point P is the minimum of its similarity with respect to each variable

Sites within the environmental range of the surveyed points return a positive value, increasing with similarity to the surveyed points. A negative value points out a site where at least one of the variables for said site had a value outside of the range of the environment of the surveyed set of points.

As useful as the approach is in the identification of unsampled environmental spaces, it is limited by its use of the minimum similarity with respect to each bioclimatic variable as an indicator for overall similarity (Owens *et al.*, 2013). Two different points may have the same similarity index value based on different variables.

RESULTS

From the distribution modelling there were 21 predicted distributions in total given the 2 target variable types, for live individuals and the 3 individual stage classes, given the 3 climate scenarios. Together with the predicted distribution maps, were maps depicting prediction uncertainty, the impact of data manipulations and distributional change across time. For the sake of brevity, only the distributional maps relevant to the analysis will be shown in this section. Additional noteworthy materials that supplement contents of this section can be found in Appendix A. Unless stated otherwise, reporting is on live *Aloidendron dichotomum* individuals (i.e. excluding dead). Presence is therefore the occurrence of either an adult or juvenile individual or both.

This section will firstly go on to present distribution modelling results based on the data as it was collected in the survey. Secondly, findings on the potential impact of climate change on the species' distribution are reviewed. The section will then go on to discuss the importance of an even amount of presence and absence points in species distribution modelling datasets. Next, the addition of generated species absences from a different geographical range into training data will be investigated. Lastly, the outcome of compositional distribution models will be examined and sensitivity to model type selection explored.

Spatial distribution patterns of *Aloidendron dichotomum*

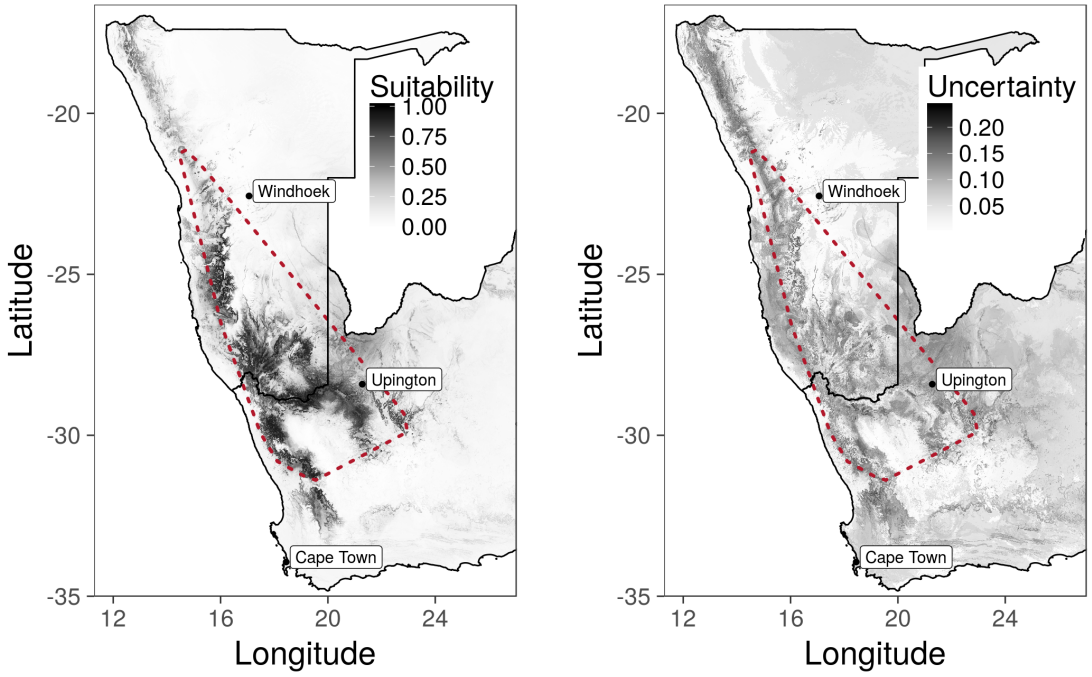
Estimated habitat suitability maps of the study area are the point-wise means from 20 individual ensemble SDMs. Each set of underlying candidate models used to construct the ensemble were fit to a dataset bootstrapped from the training dataset "as is". 10-fold cross-validation was used for hyper-parameter selection. For each candidate model, selected combinations of tunable parameters were based on accuracy and model complexity. Cross-validation results for the individual candidate models can be found in Appendix A.

This section first considers presences, showing maps of habitat suitability under current climatic conditions, and discussing the uncertainty of those estimates. We then assess which bioclimatic variables were most influential on suitability estimates, and assess the nature of these effects with partial dependence plots. Finally, habitat suitability under future climatic conditions is considered.

Estimated current habitat suitability of *Aloidendron dichotomum* (based on current "as is" data)

Using species presences

Modelling under current climate conditions indicated a large overlap between the surveyed area and the modelled distribution (see Figure 4.1a). *Aloidendron dichotomum* populations extended all the way along the West Coast but never reached the coast. The highest concentration of favourable conditions for this species was between the orange river along the South Africa-Namibia border and approximately $28°S$, with two arms extending south into the Little Karoo between approximately $17°E$ and $22°E$. Habitat suitability gradually decreased towards Windhoek, and was generally lower outside of the sampled species' distribution.



(a) Suitability

(b) Uncertainty

Figure 4.1: Current (a) estimated habitat suitability and (b) prediction uncertainty for live *Aloidendron dichotomum*. Darker areas indicate greater suitability for species occurrence / increased uncertainty. The dotted line is a convex hull of the sampled species presences.

The extent of highly suitable environmental conditions for the species was however not limited to the survey area. Results pointed towards some suitable areas of species incidence outside of the sampled distribution where *Aloidendron dichotomum* are not known to occur. Noticeable areas of species incidence outside of the sampled distribution were to the North and South of the sample zone, along the West Coast.

Figure 4.2 is a MESS map under current climate conditions, given the survey locations and eight bioclimatic variables chosen. Environmental conditions outside the sampled distribution were in general very different to the surveyed points. Conversely, sites within the environmental range of the surveyed points were located inside the survey area. Thus results from suitability estimates need to be interpreted with caution when extrapolating to outside the surveyed area. High suitability areas were generally associated with high environmental gradient

similarity to the survey points. The opposite held for areas characterised by low suitability between 0.1 and 0.3. These areas were mostly associated with very little environmental gradient similarity.

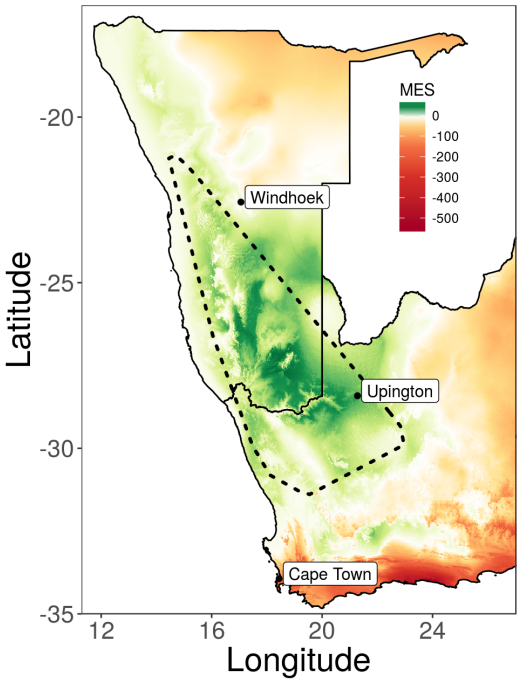


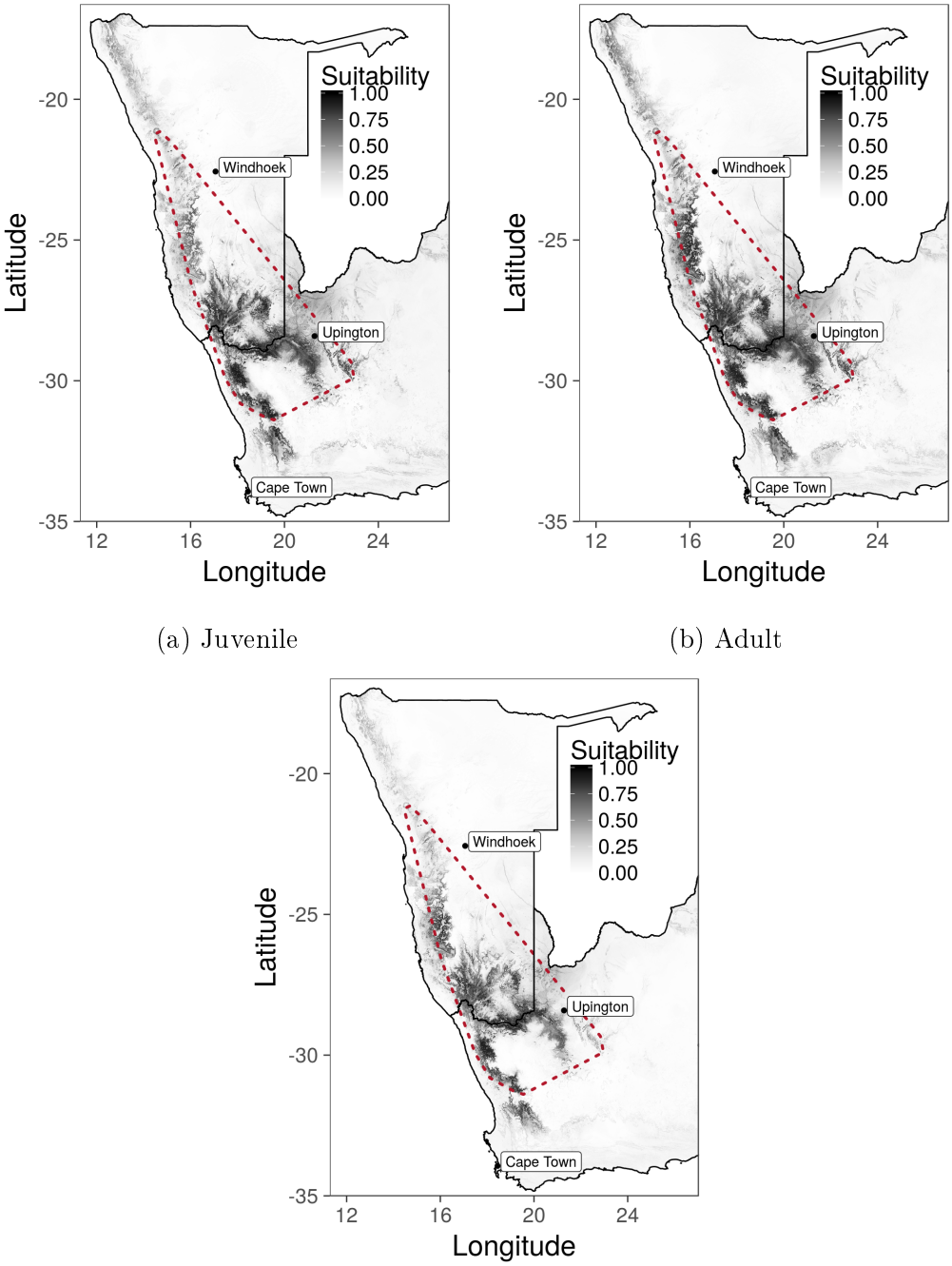
Figure 4.2: MESS map under current climate conditions. Darker green indicates greater similarity, darker red greater dissimilarity and white is no similarity.

Ensemble SDM uncertainty estimation revealed that much of the variation in suitability predictions occurred within the sampled distribution (see Figure 4.1b). Uncertainty was seen to be generally lower in areas where the species would be considered to be absent because of the very low predicted suitability, in comparison to within the sampled distribution.

Areas of high prediction uncertainty (from bootstrap estimates) are the areas where the MESS maps indicated greatest similarity, which communicated quite different information. Despite the high environmental similarity, the variability in suitability estimates may be a result of greater predicted suitability values, or non-bioclimatic factors, such as the omission of biotic factors in the modelling process. Taking non-bioclimatic factors into consideration could result in greater confidence in suitability estimates, as results show that some of the variability in

the estimates is left unexplained by SDMs.

Within each stage class

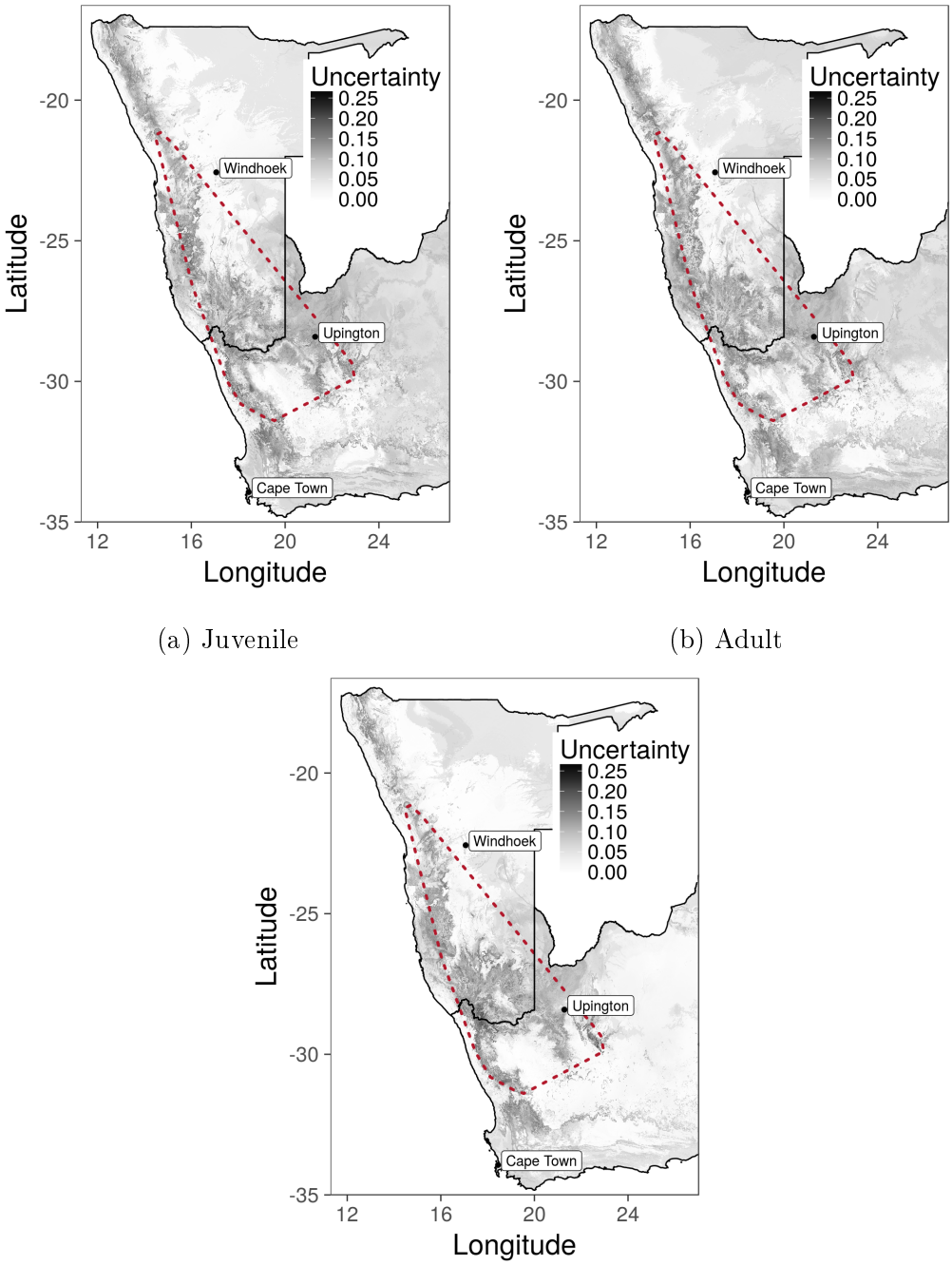


(a) Juvenile

(b) Adult

(c) Dead

Figure 4.3: Suitability estimates of (a) juvenile, (b) adult and (c) dead *Aloidendron dichotomum*. Darker areas indicate a higher environmental suitability for the species.



(a) Juvenile

(b) Adult

(c) Dead

Figure 4.4: Suitability uncertainty of (a) juvenile, (b) adult and (c) dead *Aloidendron dichotomum*. Darker areas indicate greater uncertainty in suitability estimates for the species.

All stage class distribution estimates had a resemblance to that of the full species' distribution (see Figure 4.3). The suitability estimates for juvenile and dead individuals between $25°S$ and $30°S$ inside the species' distribution, were typically lower when compared to the adult stage class distribution. Looking at habitat suitability at a stage class level did not add much to analysis.

Aloidendron dichotomum stage class uncertainty estimation showed the distribution of dead individuals to be less variable outside the sampled distribution, in comparison to the other two stage classes (see Figure 4.4).

Influence of predictors

Amongst the ten predictors, on average the terrain roughness index and altitude respectively had the greatest variable importance when determining *Aloidendron dichotomum* distribution patterns (see Figure 4.5).

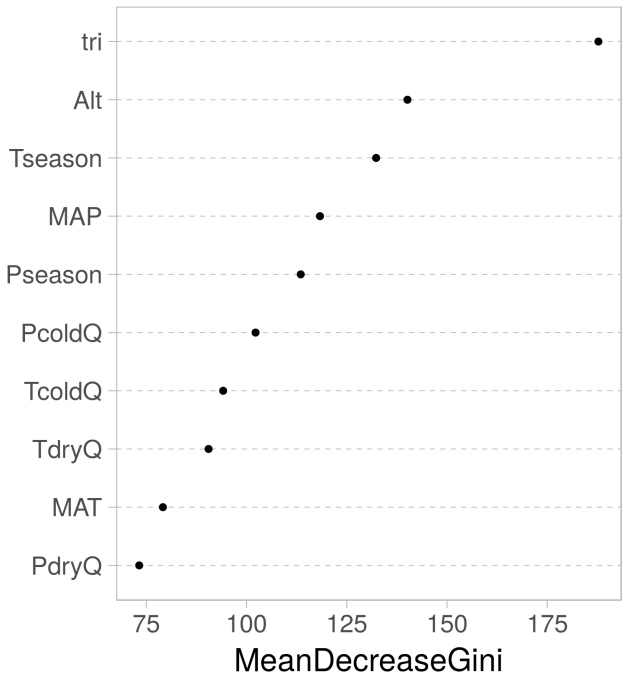
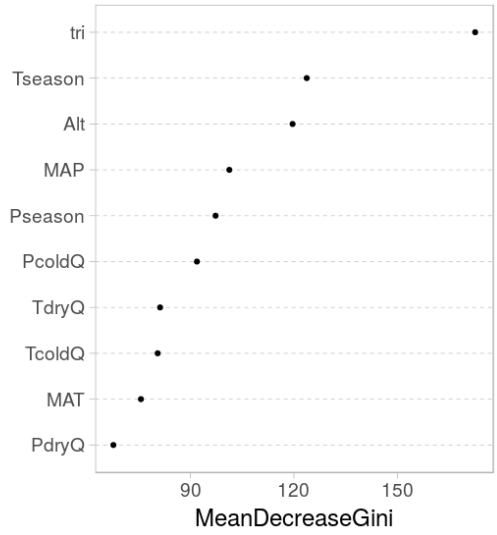
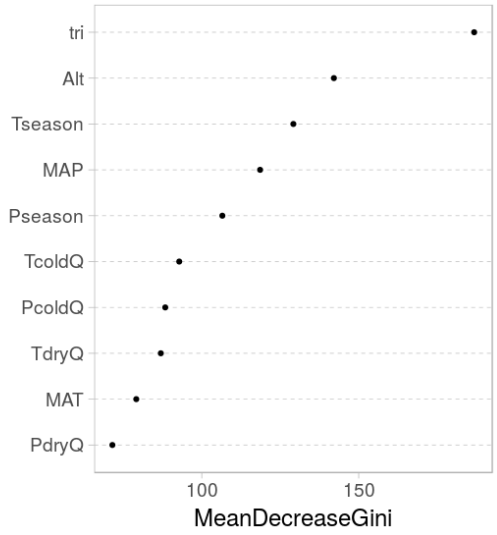


Figure 4.5: Average covariate variable importance ranked by the mean decrease in node impurity. The greater the mean decrease in node impurity the more important the variable.

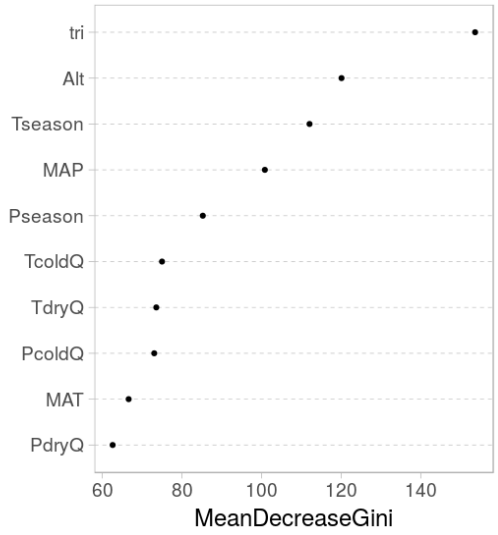
Of the bioclimatic predictors, precipitation based variables were generally more deterministic of *Aloidendron dichotomum* presence in comparison to temperature based variables. For the most part there was no change in the relative importance of predictors when species' distributions were estimated at a stage class level (see Figure 4.6).



(a) Juvenile



(b) Adult



(c) Dead

Figure 4.6: Average covariate variable importance ranked by the mean decrease in node impurity for each stage class. The greater the mean decrease in node impurity the more important the variable.

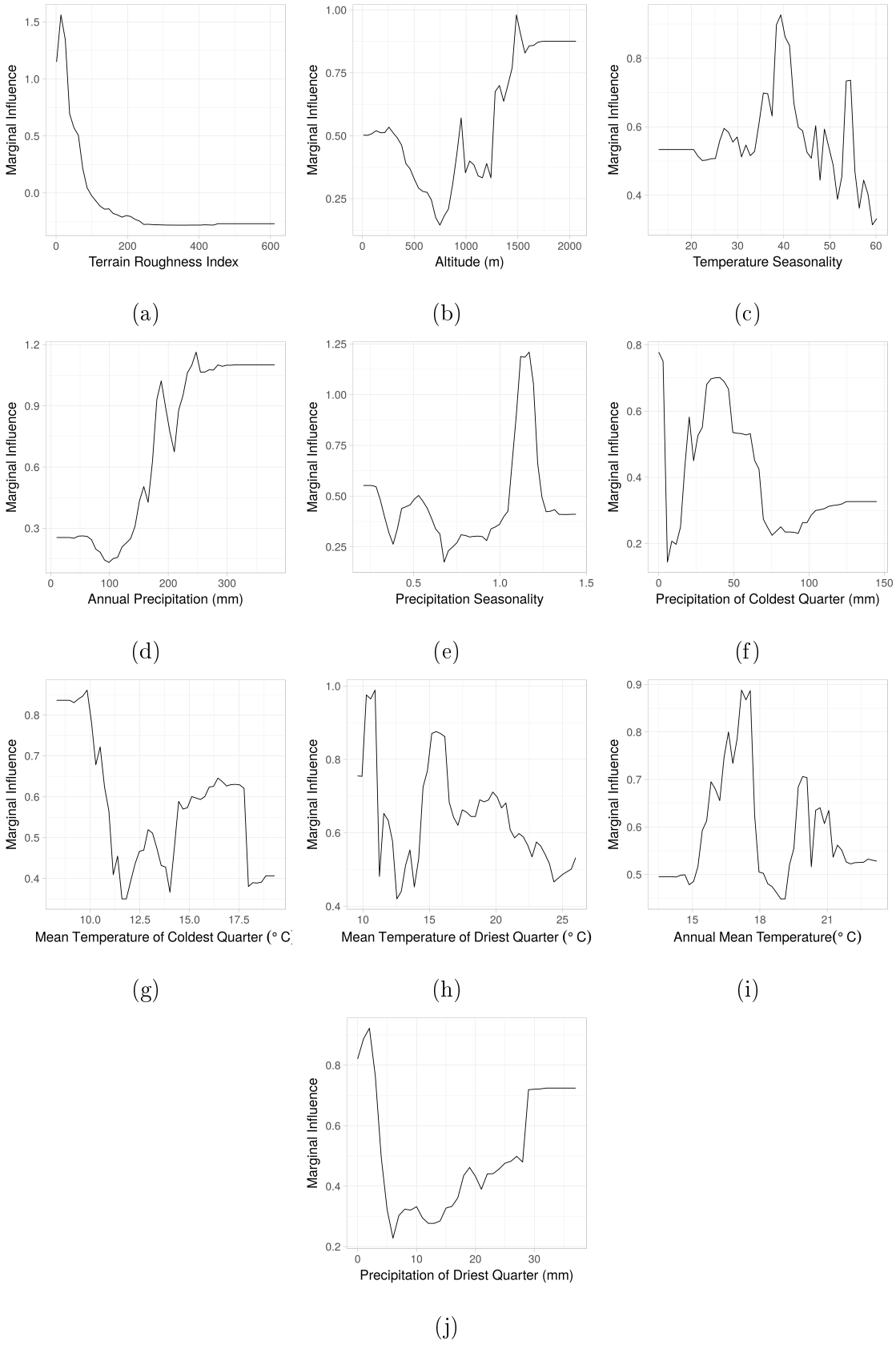


Figure 4.7: *Aloidendron dichotomum* partial dependence plots. The y-axis is the contribution of the variable to a positive classification, on a logit scale. Note the different vertical scales for each plot

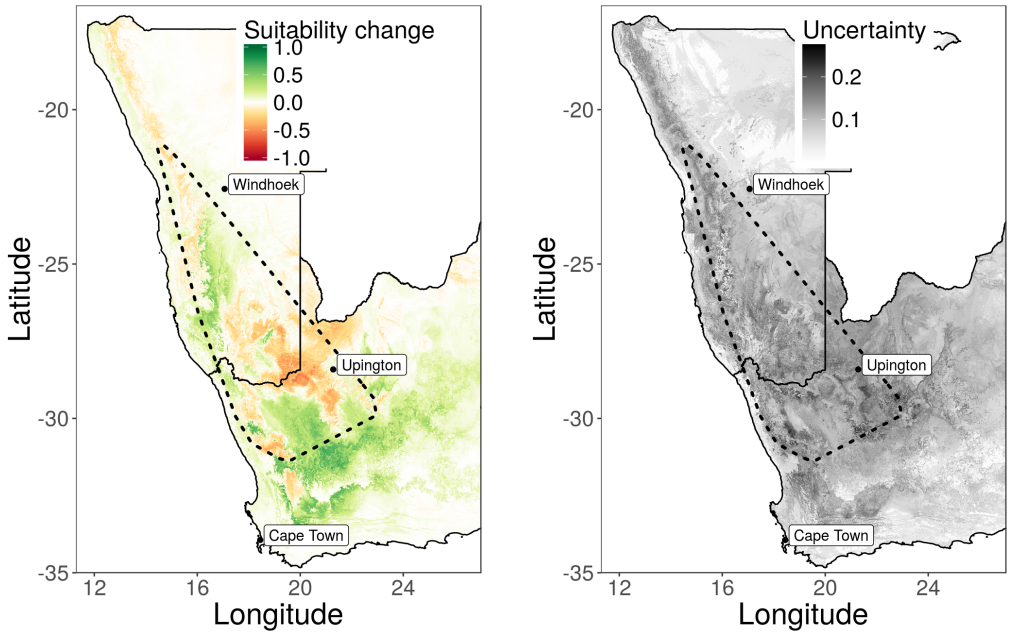
Partial dependence plots visualise the marginal influence of a variable on model predictions, given the average influence of all other variables (Friedman, 2001). Results show that values of terrain roughness greater than 20 begin to have a negative influence on suitability predictions (see Figure 4.7a). A lower terrain roughness index value is more characteristic of high species habitat suitability. All else being equal, there was a greater chance of species occurrence between precipitation seasonality of approximately 1.05 and 1.25 (see Figure 4.7e). Altitude and annual precipitation exhibited some interaction, whereby values above 750m and 100mm respectively, began to have a positive influence on suitability predictions (see Figures 4.7b & 4.7d). All else being equal, driest quarter temperatures between approximately 10.5℃ and 14.3℃ were less characteristic of species presence, than in temperatures both above and below this range (see Figure 4.7h).

Although, given the general complexity in estimating a functional relationship between the target variable and covariates, partial dependence plots are unlikely to uncover the complete detailed functional form (Friedman, 2001). Some of the output such as the positive influence of annual precipitation on presence, and the range of precipitation in the driest quarter characteristic of species presence defined above, may well be an accurate representation of the underlying relationships. However, others such as the right-hand-side of Figure 4.7c, do not look to be anything with a biological/physiological explanation and may be a result of fitting noise occurring in the data.

Estimated future habitat suitability of *Aloidendron dichotomum*

Using species presences

Projections of *Aloidendron dichotomum*'s distribution implied a noticeable south-easterly shift in *Aloidendron dichotomum* habitat suitability by 2050 (see Figure 4.8a). It is important to note that a decrease or increase in the suitability of an area did not necessarily indicate changing presence or absence, but rather a shift in environmental suitability for the species.



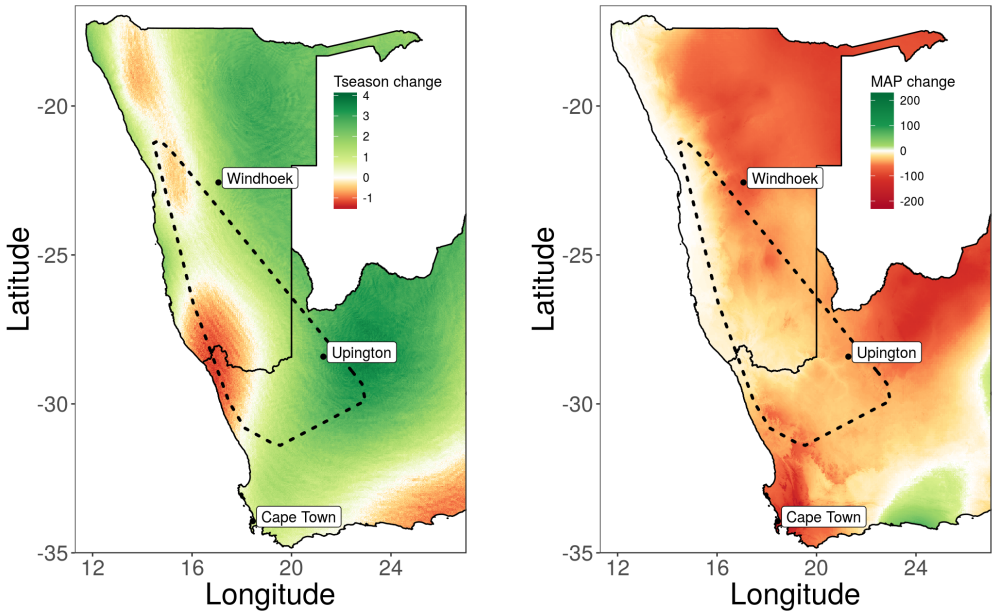
(a) Suitability (b) Uncertainty

Figure 4.8: The (a) Current to 2050 change in habitat suitability and (b) associated uncertainty for live *Aloidendron dichotomum*

A large negative shift in environmental suitability was observed inside the defined species' distribution, within the Riemvasmaak Conservancy and to the east of the area, towards Upington. Patterns of negative change also occurred along the coast, extending as far north as Mowe Bay in Namibia. Some potential population recovery was however observed in the southern region of the distribution, with a south-easterly expansion into the Little Karoo.

The 2050 climate scenario saw temperature becoming more variable and a drop in annual precipitation for a large part of the survey area (see Figure 4.9). Temperature seasonality and annual precipitation were two of the most important bioclimatic variables according to Figure 4.5. The positive change in estimated habitat suitability by the western boundary of the survey zone could be attributed to a combination of the declining temperature seasonality, and little to no decrease in annual precipitation in the area.

Climate scenarios anticipated drier conditions across the study area, seen by a general decline in annual precipitation levels, with the exception some substantial positive change by the Garden Route national park which is around 24°E, south-east of the survey zone range (see Figure 4.9b). Some increase in habitat suitability in that area was not surprising to see, given the species' strong relation to precipitation (Midgley *et al.*, 2009).



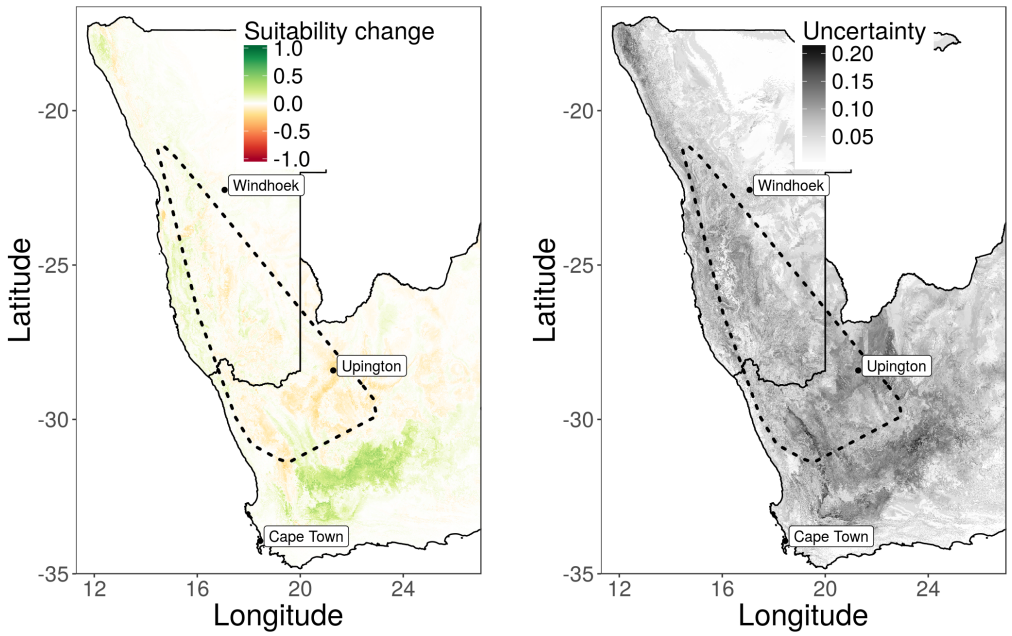
(a) Temperature seasonality

(b) Annual precipitation

Figure 4.9: The Current to 2050 change in (a) temperature seasonality and (b) annual precipitation, two of the most important bioclimatic variables.

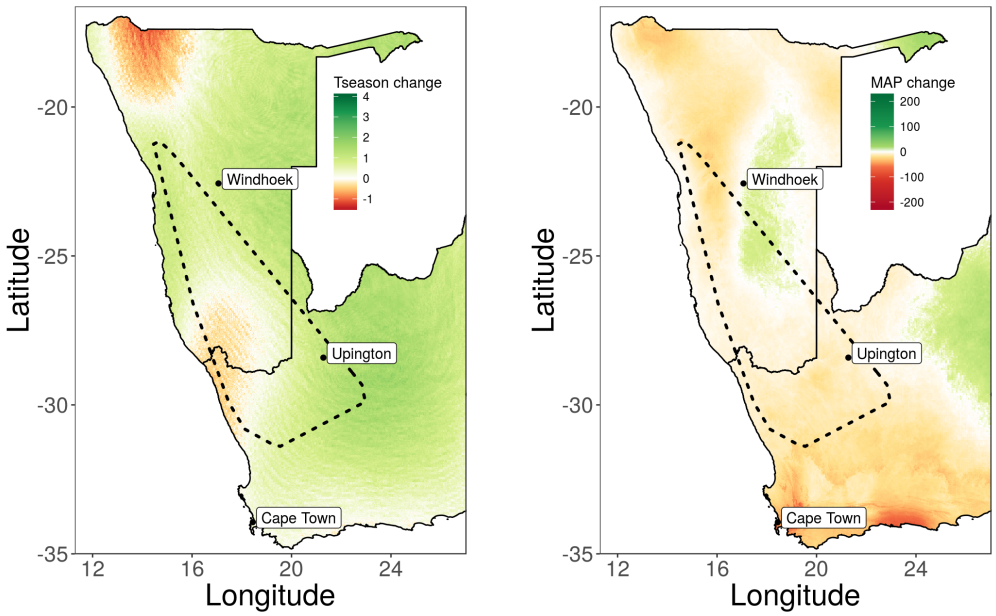
The 2070 climatic scenario implied a less dramatic shift in suitability, still in a south-easterly direction (see Figure 4.10a). Under this scenario a noticeable amount of the predicted favourable conditions under the 2050 climate were maintained. Some of the area close to Upington had continual negative change in suitability. Even so, the area in the direction of the Little Karoo was becoming progressively favourable for species occurrence.

Again, increasingly variable temperature and drier conditions were observed from the 2050 scenario to 2070, but to a lesser extent than from current conditions (see Figure 4.11). Much of the positive change in suitability estimate occurred where there was a slight increase in temperature seasonality and a drop in annual rainfall, both of which would make recruitment less probable(Jordan and Nobel, 1979). Despite their relative importance, results show that these two bioclimatic variables alone may not be sufficient to explain the trend between the potential climate change and the distribution of conditions suitable for *Aloidendron dichotomum* occurrence.



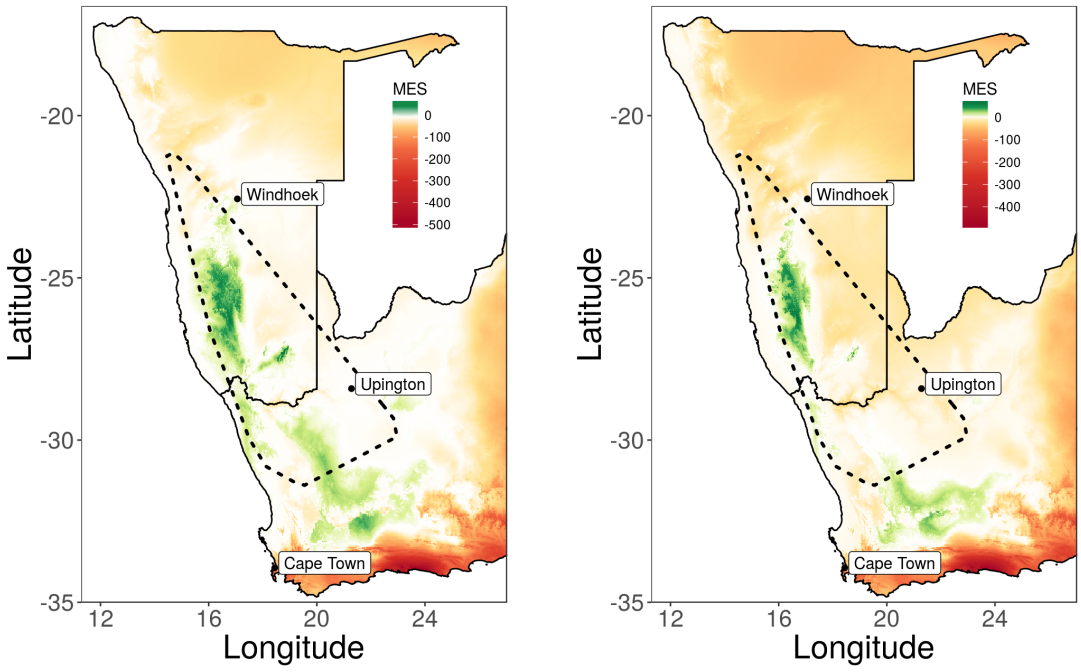
(a) Suitability (b) Uncertainty

Figure 4.10: The (a) 2050 to 2070 change in habitat suitability and (b) associated uncertainty for live *Aloidendron dichotomum*



(a) Temperature seasonality (b) Annual precipitation

Figure 4.11: The 2050 to 2070 change in (a) temperature seasonality and (b) annual precipitation, two of the most important bioclimatic variables.



(a) 2050 (b) 2070

Figure 4.12: MESS map under future climate conditions. Darker green indicates greater similarity, darker red greater dissimilarity and white is no similarity.

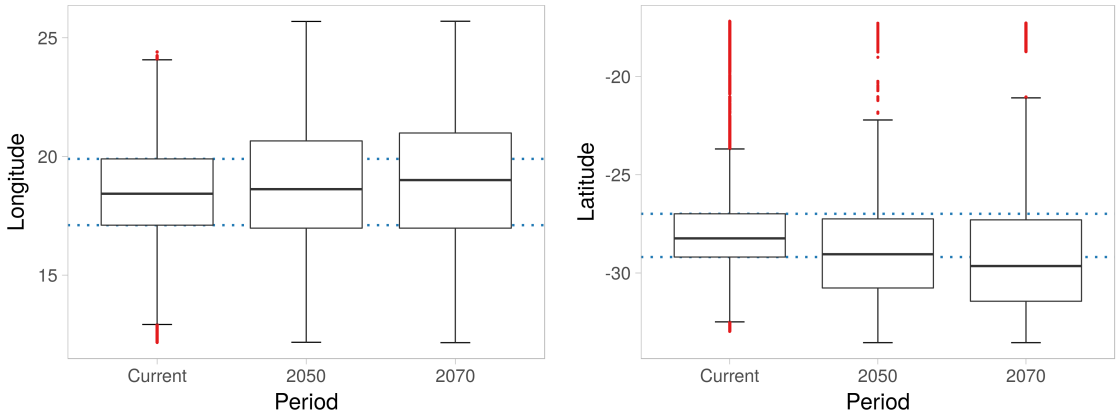
Comparison of future climates to the current survey points revealed SDMs were largely predicting into considerably different environmental gradients, both inside and outside the survey zone (see Figure 4.12). MESS maps indicated a substantial decrease in the spread of sites characterised by environmental gradients similar to that of the survey data points. The 2050 scenario largely had little to no similarity to the survey data locations with respect environmental gradients. By 2070, environmental gradients for regions within the sampled distribution were becoming increasingly novel and there was even fewer areas of highly similar environmental gradients.

The shift in environmental similarity from current conditions to the 2050 scenario was greater in comparison to the change between 2050 and 2070. A common feature of both climate scenarios was the concentration of very similar environmental gradients within the species' distribution south of Windhoek, up to the South Africa-Namibia border by the Richtersveld World Heritage Site. This area was predicted to experience an initial increase in habitat suitability (see Figure 4.8a). Additionally, similar environmental gradients outside of the sampled distribution towards the Little Karoo were expected to experience an increase in habitat suitability over time.

Future uncertainty estimates were to a great degree similar, exhibiting generally greater uncertainty in areas of estimated positive habitat suitability change (see Figures 4.8b & 4.10b). Note though, that the average changes in 2070 estimated suitability maps were smaller, which meant uncertainty as a percentage of change was greater in 2070.

A temporal analysis of *Aloidendron dichotomum*'s latitudinal and longitudinal range revealed a south-easterly shift in suitable species conditions over time. An eastward expansion in habitat suitability was observed, whilst maintaining the western boundary, alongside the poleward shift in suitability (see Figure 4.13). The positioning of the distribution of *Aloidendron dichotomum* shifted with each

subsequent climate scenario in a south easterly direction.



(a) Longitudinal range

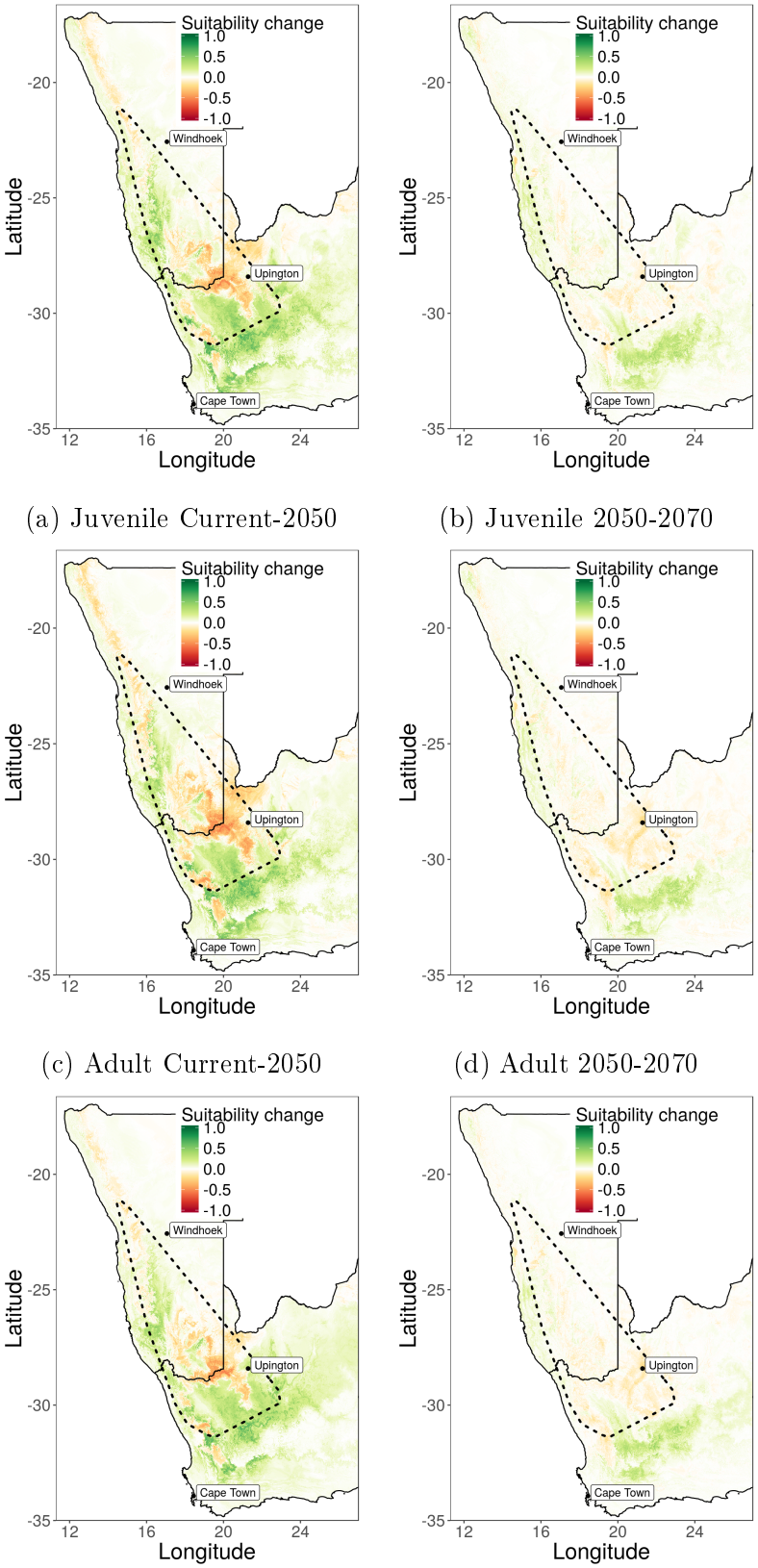
(b) Latitudinal range

Figure 4.13: Box-plots of *Aloidendron dichotomum* (a) longitudinal, and (b) latitudinal ranges given predicted species presence. Dotted lines represent the respective current upper and lower quartiles of presence and can be used to investigate any temporal shifts given a climate change scenario. Red points represent outliers.

Within each stage class

Estimated suitability change under the 2050 climate scenario for all stage classes within the survey zone was comparable to a great extent, moving in a south-easterly direction (see Figure 4.14). The dead stage class however was estimated to experience less of a negative change in habitat suitability around the Riemvasmaak Conservancy (see Figure 4.14e), in comparison to the other two stage classes. Of the three stage classes, the juvenile population was seen to have the greatest positive change in estimated habitat suitability by 2050 (see Figure 4.14a). The positive shift in habitat suitability extended into the Little Karoo and occurred as far east as Kimberly.

As with previous results, the estimated shift in environmental suitability between 2050 and 2070 was less striking, with a continued decrease in suitability within the sampled distribution and some positive change south of the sampled distribution. Species' stage class latitudinal and longitudinal ranges presented a sizeable expansion to the south-east with each subsequent climate scenario (see Figure 4.15)



(a) Juvenile Current-2050

(b) Juvenile 2050-2070

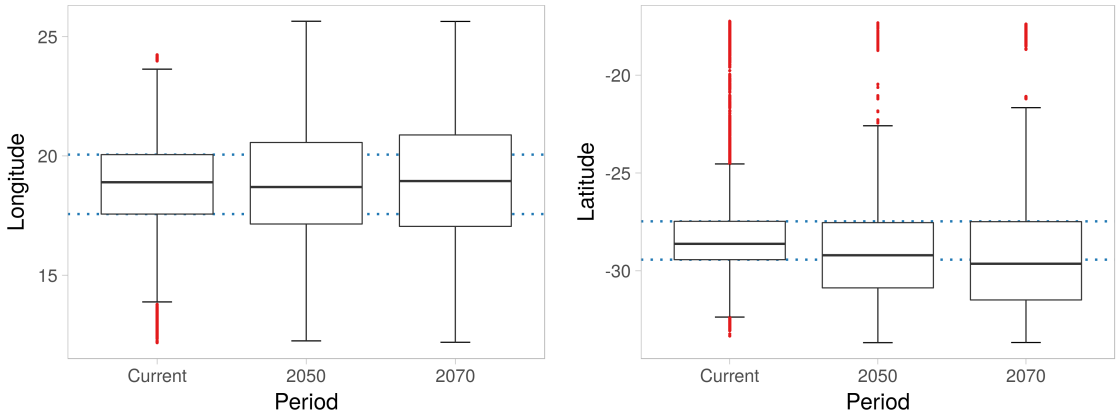
(c) Adult Current-2050

(d) Adult 2050-2070

(e) Dead Current-2050

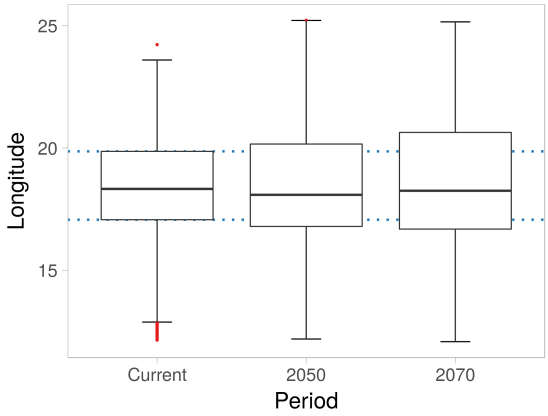
(f) Dead 2050-2070

Figure 4.14: Period-on-period change in habitat suitability for each *Aloidendron dichotomum* stage class.

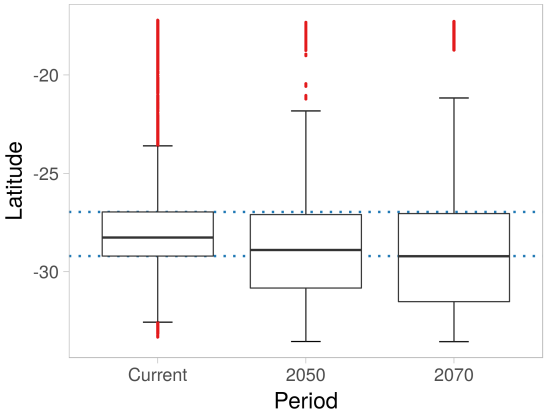


(a) Juvenile longitudinal

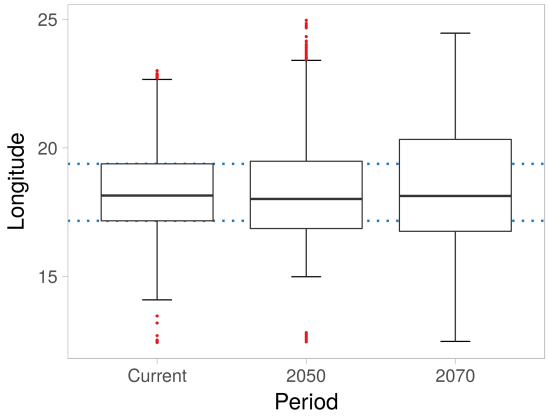
(b) Juvenile latitudinal



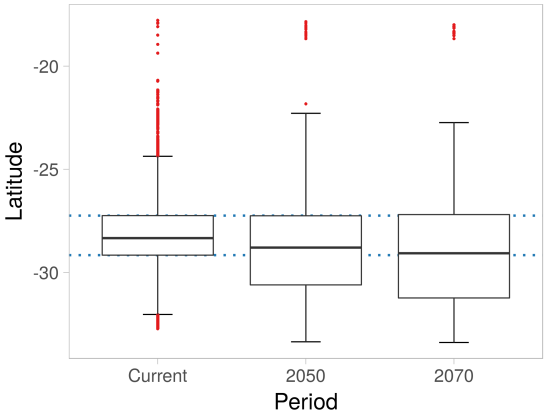
(c) Adult longitudinal



(d) Adult latitudinal



(e) Dead longitudinal



(f) Dead latitudinal

Figure 4.15: Box-plots of *Aloidendron dichotomum* longitudinal and latitudinal ranges given stage class species presence. Dotted lines represent the respective current upper and lower quartiles of presence and can be used to investigate any temporal shifts given a climate change scenario. Red points represent outliers.

Assessing the effect of data manipulations

Downsampling to create equal class sizes

Figure 4.16 shows that ROC curves for SDMs fit to imbalanced and balanced datasets were nearly identical, suggesting that in this case creating equal-sized response classes by downsampling from the larger class did not materially affect classification accuracy. AUC values for ensemble models fit to the two dataset types were similar, further proving downsampling to be ineffective at error reduction. Models were compared on the same test dataset with no manipulations, one which reflects the true underlying proportion of presences.

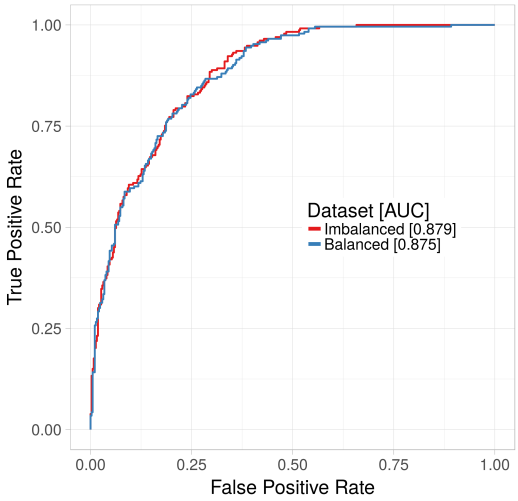


Figure 4.16: Balanced and imbalanced class dataset ROC curve, depicting binary classifier accuracy as the class separation threshold is varied.

Imbalanced and balanced data SDMs had an average accuracy of 80% and 79% respectively (see Figure 4.17). Presence and observed absence accuracies were however different for the two datasets as expected, given the constant threshold of 0.5 as the number of absences in data were reduced. Average classifier performance on absences (false positive rate) was 8% higher compared to that of balanced class data SDMs, and presences were more accurately predicted (12% higher true positive rate) with balanced class data.

Whether this is a desirable trade-off depends on the application and situation, one can not make a judgement. However, overall, it is clear that balancing response class frequencies made little difference to accuracy and just redistributed where that accuracy came from.

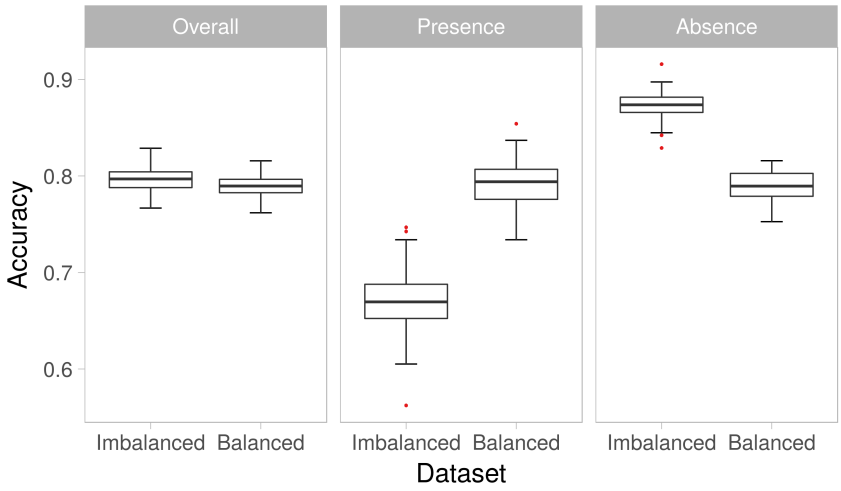


Figure 4.17: Bootstrapped overall, presence (true positive) and absence (true negative) model accuracies for balanced and imbalanced class data. Red dots represent outliers

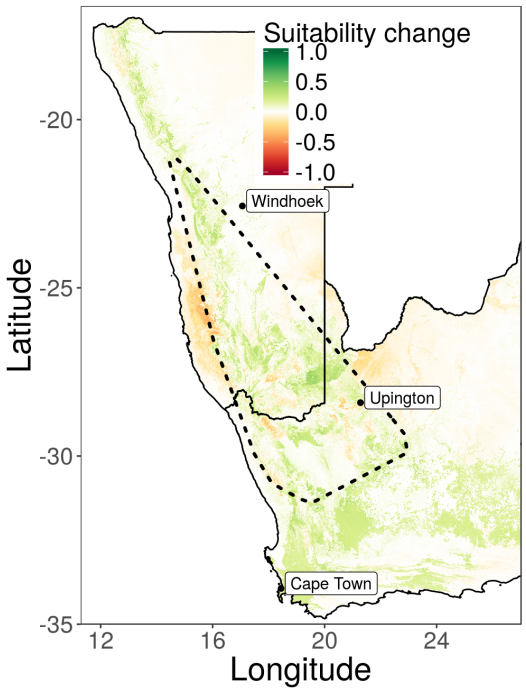


Figure 4.18: Change in suitability estimates given balanced *Aloidendron dichotomum* response class frequencies.

One however has to consider the effect of equal class data on the interpretation of results, because removing absences increases the prior probability that an observation is a presence. This results in an arbitrary increase in habitat suitability across the study area as seen in Figure 4.18. The suitability in the study area generally increased, as expected, and this is acceptable if one only interprets the results strictly relatively, but some suitability outside the survey zone actually decreased, accentuating the difference between inside and outside the survey zone. This was an unintended side effect of balancing.

Given the evidence above and the implications of down-sampling when interpreting results, providing the SDM equal class data was unnecessary. In keeping with imbalanced response class frequencies, Barbet-Massin *et al.* (2012) and Merow *et al.* (2013) put forward that it is necessary to have very large absence datasets in order to produce accurate species distributions, but these apply to presence-only data.

Adding absences outside of the known distribution

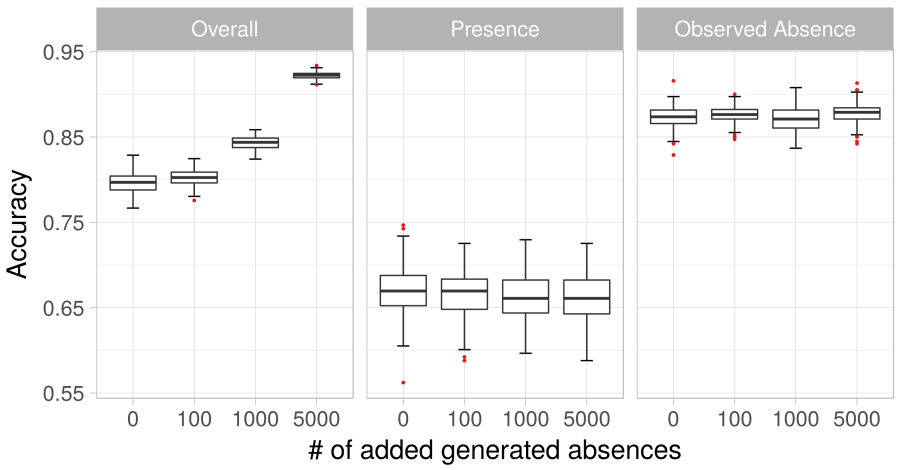


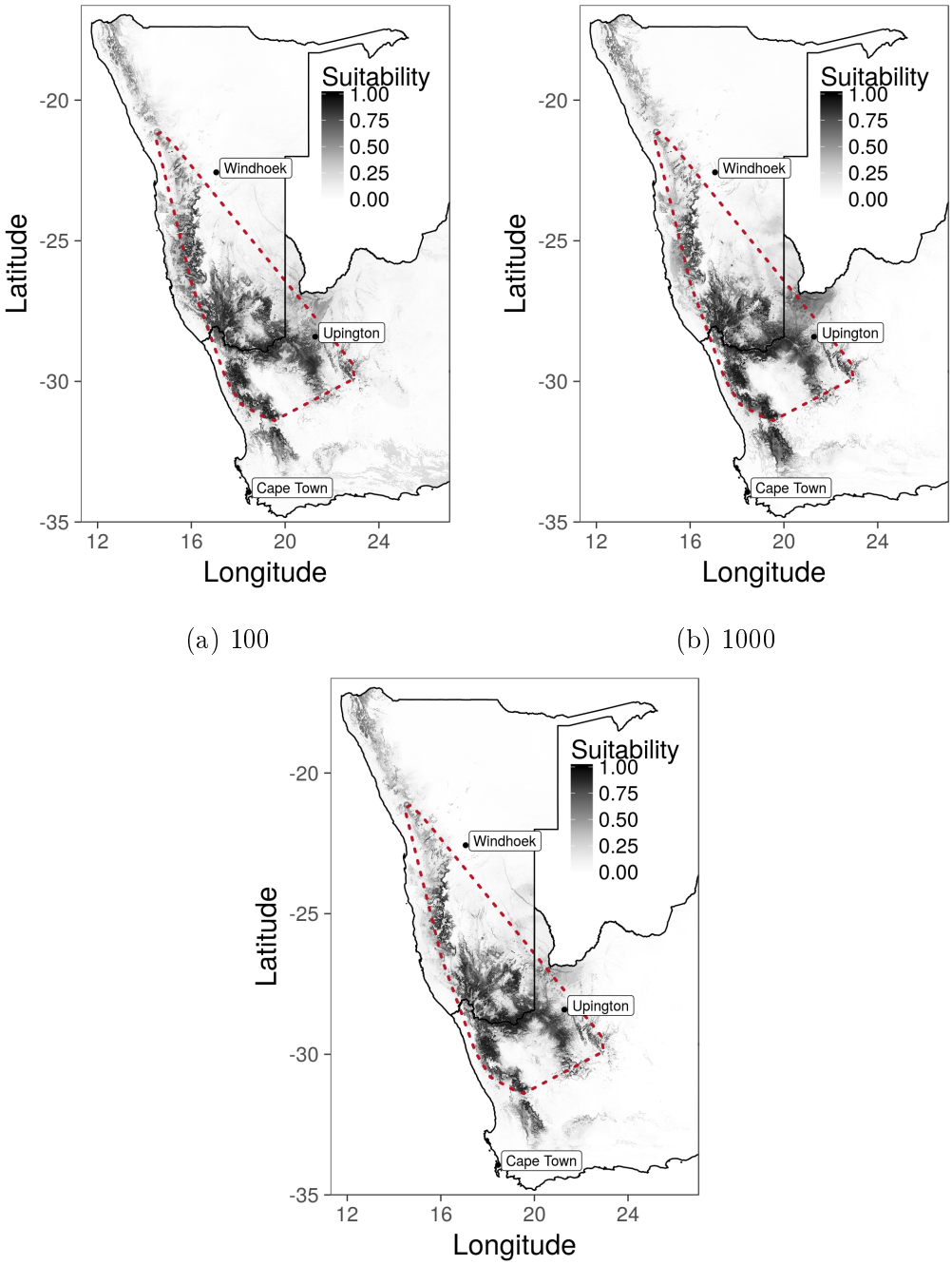
Figure 4.19: Bootstrapped overall, presence (true positive) and absence (true negative) model accuracies for varying amounts of added generated species absences. Red dots represent outliers

Three different sample sizes of generated absence data were considered: 100, 1000 or 5 000 absences. On average, overall model accuracy increased as more of the generated absences were added to the input data, because the generated absences were predicted well (see Figure 4.19). The accuracy of known species presences and absences remained the same.

The approximately 13% increase in overall accuracy given an additional 5000 generated absences was attributed to the consistently perfect detection of the generated species absences. No improvement was seen in the average accuracy of species presence and observed absence, which slightly deviated around approximately 66% and 87% respectively. On this basis, added absences did not appear to at all to impact the classification accuracy of observed presences or observed absences.

An arbitrary amount of generated absences could therefore be appended to input data. As with balancing classes, what matters is the impact on interpretability of results, given that this would alter the prior probability of species presence.

Once generated absences were introduced into data this typically caused lower estimated habitat suitability both inside and outside of the survey zone, when compared to estimates from the data as is (see Figures 4.21 & 4.1a). The overall visible decrease in suitability intensified within the study area as more absences were added to training data. Some positive change in habitat suitability was observed within the survey zone in areas estimated to have high suitability in the absence of any data manipulations. Suitability estimates inside the survey zone were impacted by the additional qualitative information about the known species distribution, making the data manipulation undesirable. Ideally, the additional absences would adjust estimates only in the spaces where SDMs are extrapolating.

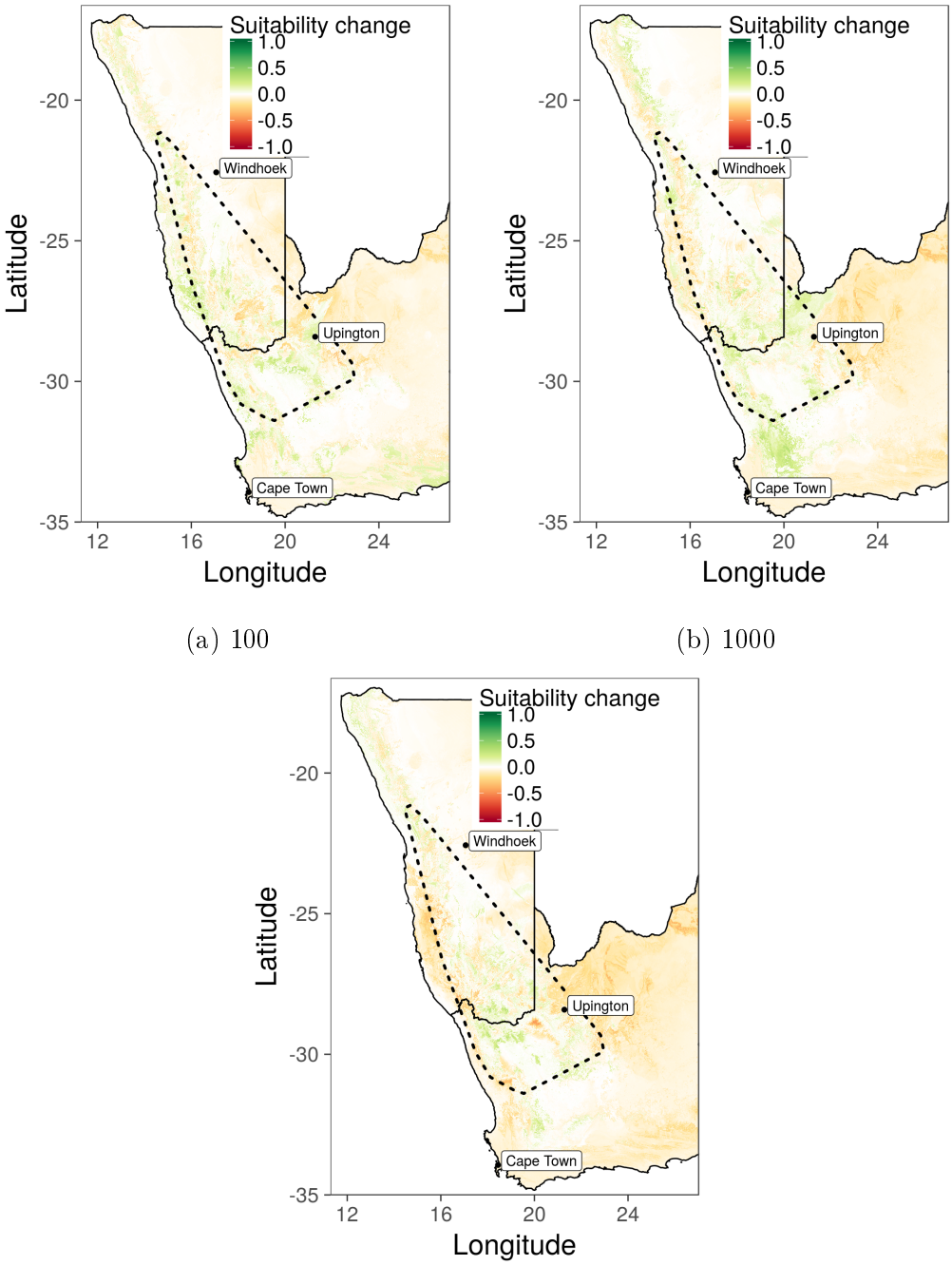


(a) 100

(b) 1000

(c) 5000

Figure 4.20: Suitability estimates given (a) 100 (b) 1000 and (c) 5000 added *Aloidendron dichotomum* absences. Darker areas indicate a higher environmental suitability for the species.



(a) 100

(b) 1000

(c) 5000

Figure 4.21: Change in suitability estimates given (a) 100 (b) 1000 and (c) 5000 added *Aloidendron dichotomum* absences.

Assessing the effect of different predictive models

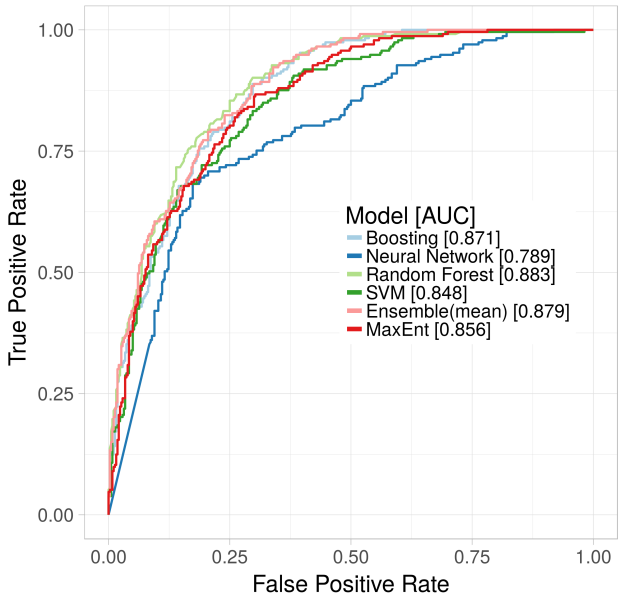


Figure 4.22: Predictive model ROC curves, depicting binary classifier accuracy as the class separation threshold is varied.

Upon comparison of true positive and false positive rates, classifier performance was largely similar with no outright worst model. Based on ROC curves, none of the candidate models were considered perfect classifiers (see Figure 4.22). The AUC statistic communicated there is little difference in performance among the models, with the exception of the neural network which returned a noticeably lower statistic.

Noteworthy models were the random forest and ensemble which on average had the lowest misclassification error rate at 19% (see Figure 4.23). The neural network SDM had the most varied performance between 20% and 27% misclassification when determining species presence or absence. On average the neural network SDM was 5% worse than the best. The best single recorded performance was produced by the ensemble model at a 17% misclassification rate.

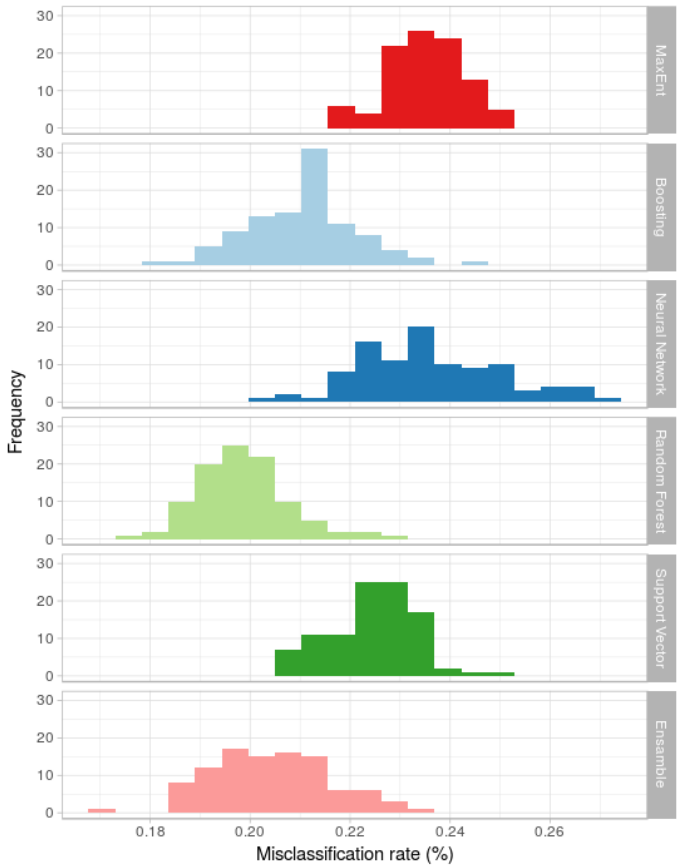
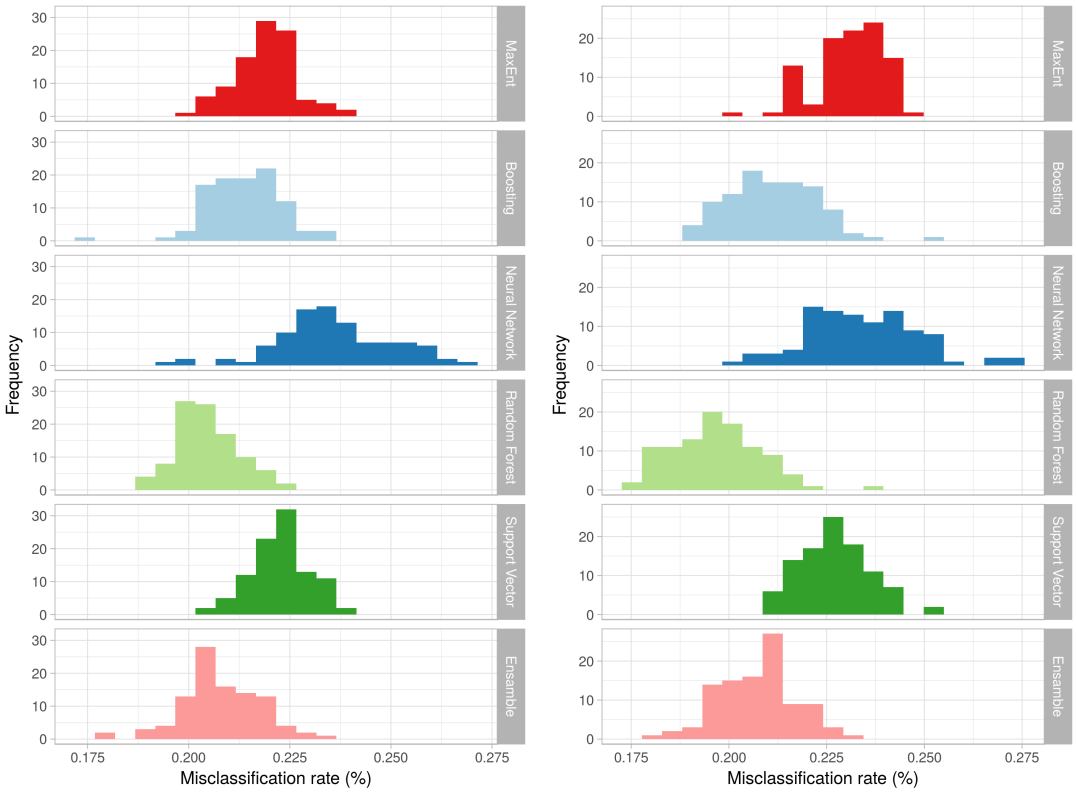


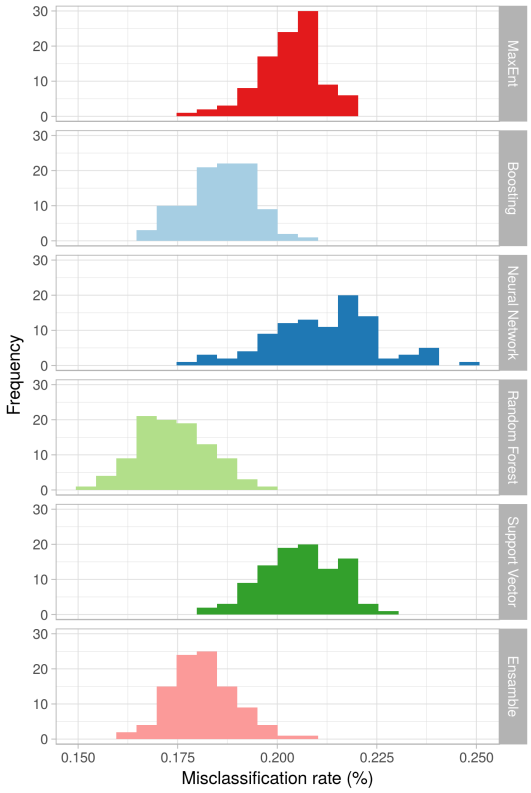
Figure 4.23: Histogram of bootstrapped misclassification rate for the individual classifiers. Species presence is the occurrence of live *Aloidendron dichotomum* individuals.

Classifier performance at a stage class level maintained the same pattern as for presences, where there was relatively little difference, between models (see Figure 4.24). Species' detection performance across models was more similar for juveniles, least so for adults, and intermediate for dead individuals. Juvenile and dead stage class returned generally lower misclassification errors in comparison to the adult stage class as expected, because at a stage class level response class frequencies had more absences of the two stage classes.



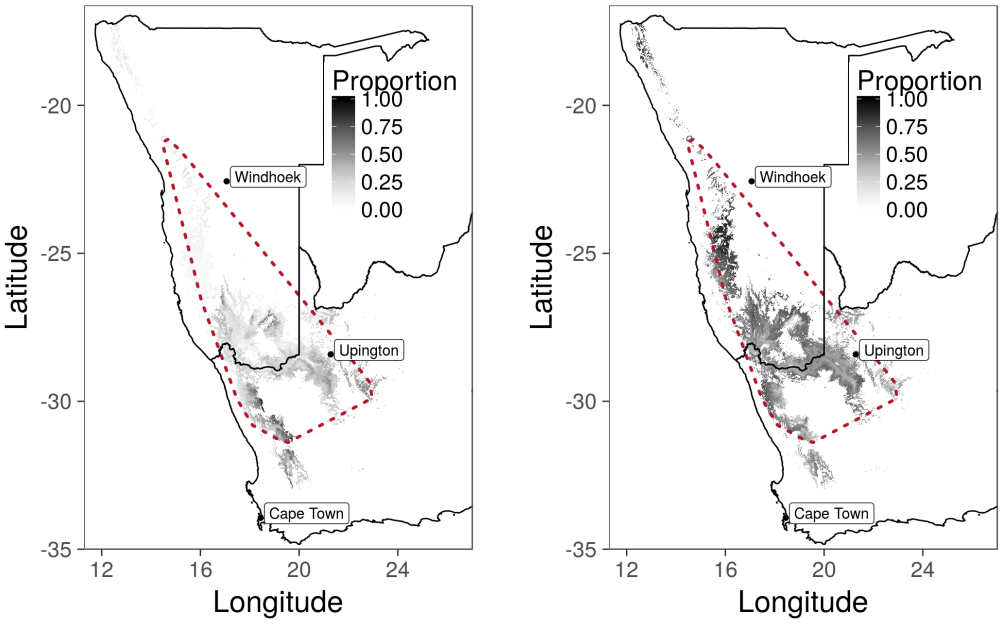
(a) Juvenile

(b) Adult



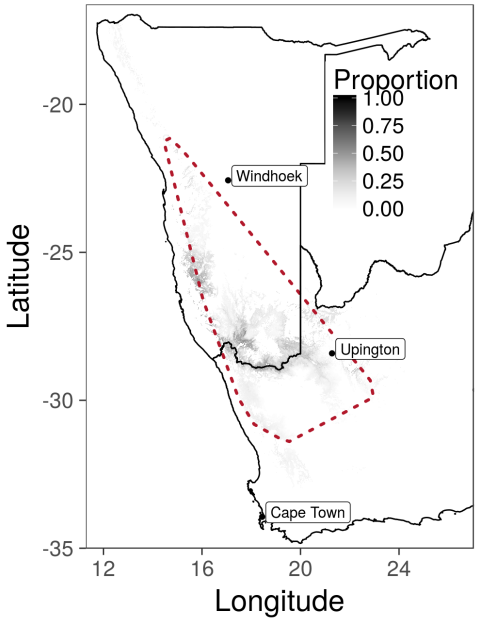
(c) Dead

Figure 4.24: Histogram of misclassification rate for the individual classifiers within each stage class.



(a) Juvenile

(b) Adult

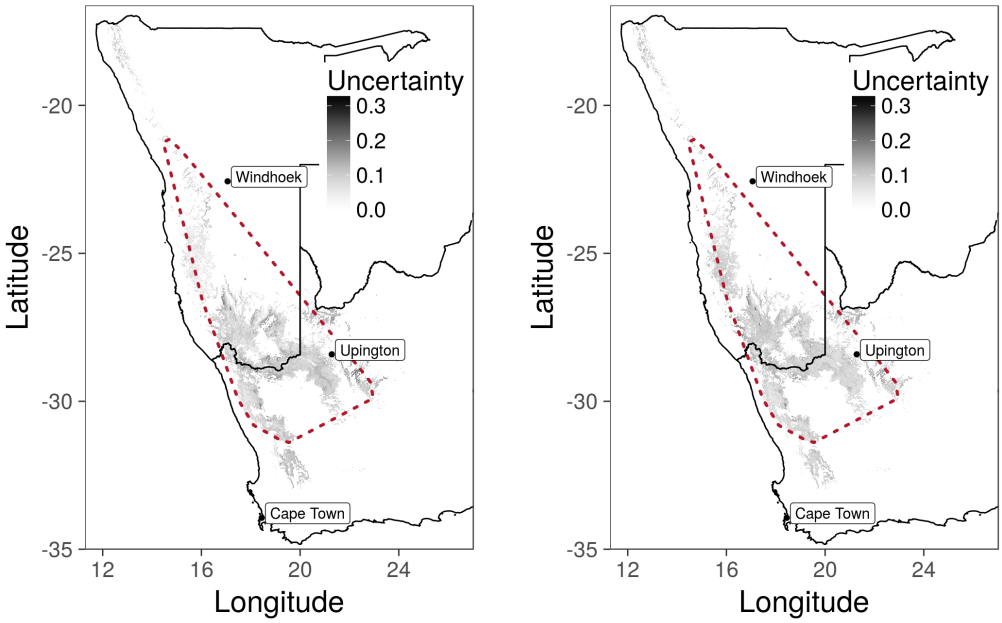


(c) Dead

Figure 4.25: Current proportional density estimates of (a) juvenile (b) adult and (c) dead *Aloidendron dichotomum*, given estimated presence. Darker areas indicate a greater proportion for the species stage class.

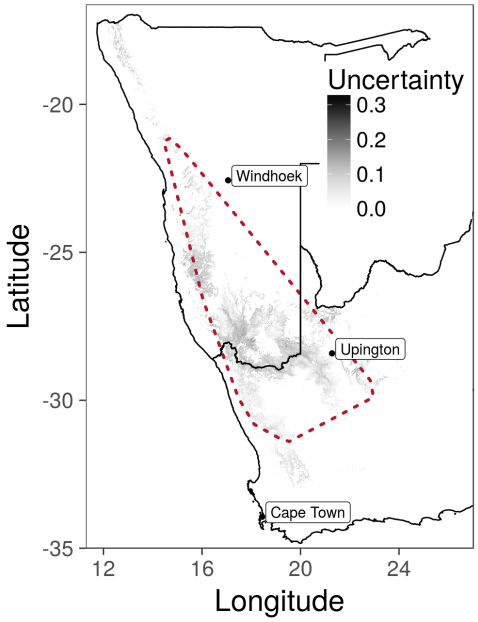
Proportional density distribution estimates predicted adult stage class dominant populations throughout the species' distribution. This result was as expected, because significantly more adult individuals were recorded in training data. Populations predicted to have a considerable proportion of dead individuals were between approximately $17°E$ and $21°E$, along the South Africa-Namibia border line (see Figure 4.25c). These *Aloidendron dichotomum* populations on the northern edge of the Riemvasmaak Conservancy, had up to a 63% composition of dead individuals. The juvenile stage class was estimated to have less of a geographical presence within the survey area, in comparison to the dominant adult population especially in the north of the study area, in Namibia.

The uncertainty associated with adult stage class compositional estimates appeared to be constant across space, with no areas within the study area with a visible concentration of high compositional uncertainty (see Figure 4.26b). There was a noticeable decrease in uncertainty associated with dead state class compositions south of the Gariep River valley, an area where *Aloidendron dichotomum* populations were characterised by very low dead stage class compositions (see Figure 4.26c). Juvenile stage class compositional uncertainty decreased in the opposite direction to the dead stage class, again into areas with low estimated juvenile proportions (see Figure 4.26a).



(a) Juvenile

(b) Adult



(c) Dead

Figure 4.26: Current proportional density uncertainty of (a) juvenile, (b) adult and (c) dead *Aloidendron dichotomum*, given estimated presence. Darker areas indicate greater uncertainty in compositional estimates for the species stage class.

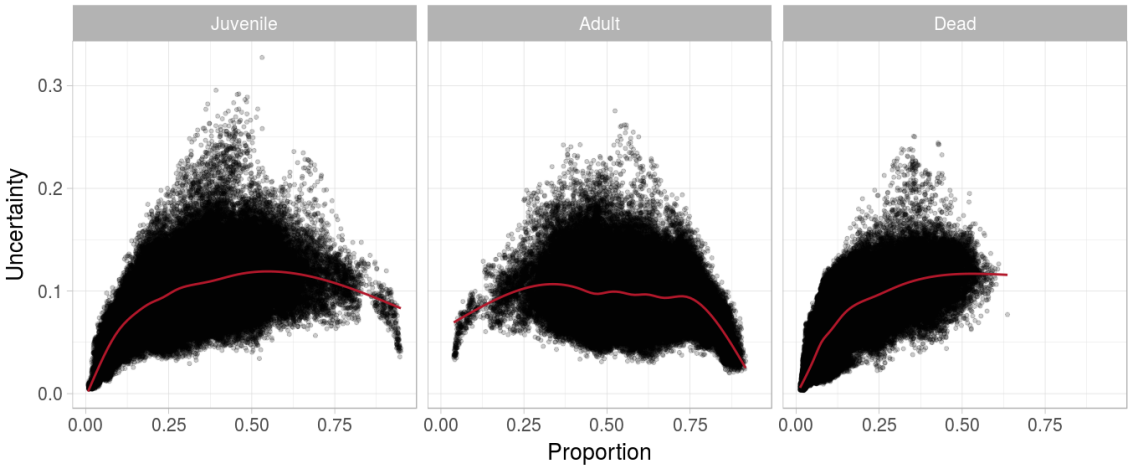


Figure 4.27: Stage class proportion estimate vs uncertainty. The red line is a generalised additive model (GAM) of uncertainty as a function of the estimated proportion.

The relationship between proportion estimates and uncertainty resembled an upside down U-shape in both juvenile and adult stage classes (see Figure 4.27). This also held for the dead stage class, but dead stage class proportion estimates were never large enough to to get the full shape of the relationship with uncertainty. The greatest uncertainty for live individuals was observed mostly at around a 50% composition. Larger estimated adult proportions were in general associated with lesser uncertainty, and the opposite held for the other two stage classes. Therefore, adult dominant populations were estimated with greater certainty in comparison to other composition combinations. This was as expected because populations in training data were skewed towards the adult stage class.

Given optimal species occurrence conditions under the current climate scenario, the average adult stage class proportion was estimated to be twice as large in comparison to the average juvenile composition, and 3.53 times that for dead individuals (see Table 4.1). In survey data, the average adult population proportion was 3.3 times bigger than the average dead proportion and approximately twice as large for juveniles. Ratios deviated slightly under the current climate scenario, but the bias towards the adult stage class was less prominent given future climate scenarios. The largest ratio shift was seen with the dead stage class, where by 2070 average adult population proportion was 2.42 times larger than the average dead

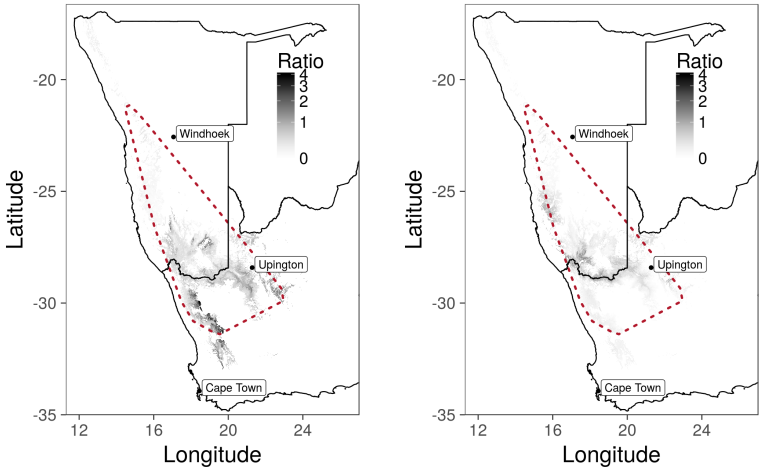
proportion. This means that the proportion of dead individuals was predicted to grow greatly by 2050 and 2070.

Period	Adult:Juvenile	Adult:Dead
Current	2.11	3.53
2050	2.09	2.71
2070	2.00	2.42

Table 4.1: Ratio of predicted average *Aloidendron dichotomum* adult stage class compositions to estimated average juvenile and dead compositions.

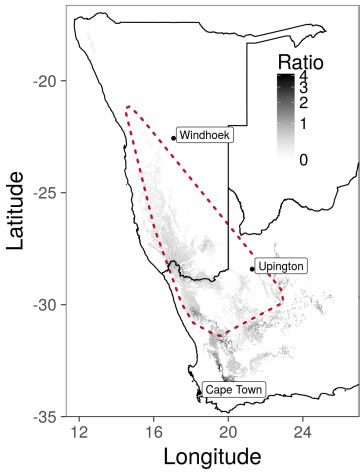
The relative abundance of juvenile and dead individuals however was not constant over space. Proportional density estimates for the juvenile stage class were poleward increasing (see Figure 4.25a), and the opposite held for the dead stage class, whose presence increased in the direction of Windhoek (see Figure 4.25b). Figure 4.28a shows juvenile populations become increasingly abundant in a poleward direction, relative to the dominant adult stage class. Under future climate scenarios, juvenile stage class compositions in the Karoo region were estimated to be up to as much as 4 times bigger than adult compositions (see Figure 4.28c & 4.28e). Dead stage class compositions were for the most part, a constant fraction of the adult stage class (see Figure 4.28). Their relative abundance intensified between $28°S$ and $29°S$ in the Gariep River valley, under current and future climate scenarios.

In keeping with MESS maps in Figure 4.12, under future climate scenarios there was comparatively greater certainty in estimates of the relative abundance of the dead stage class ratios to that of juvenile compositions. Substantial dead compositions were found mostly within the survey zone, where there was some similarity to surveyed environmental conditions. Whereas juvenile abundant populations were estimated to occur where there was no environmental similarity to surveyed locations, therefore requiring extrapolation.

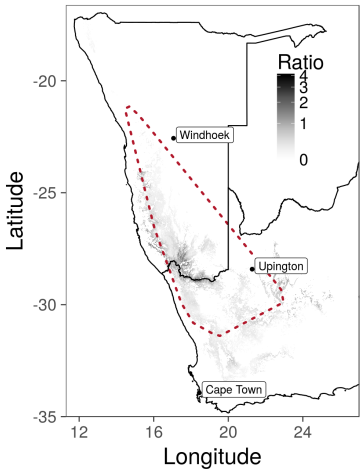


(a) Juvenile current

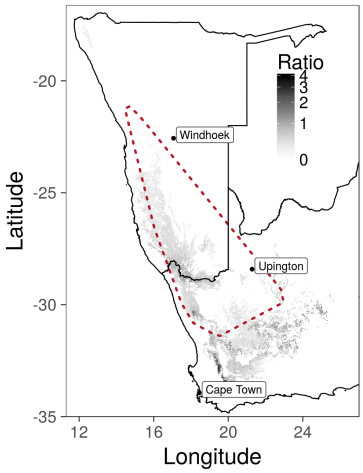
(b) Dead current



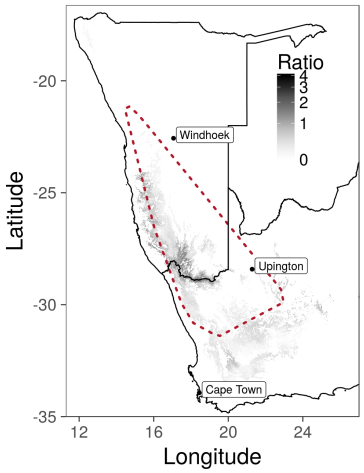
(c) Juvenile 2050



(d) Dead 2050



(e) Juvenile 2070



(f) Dead 2070

Figure 4.28: *Aloidendron dichotomum* juvenile and dead stage class compositions as a ratio of the adult population. Colour scale values have been adjusted using a Box-Cox transformation ($\lambda_1 = 0$ and $\lambda_2 = 1$), to make smaller values visible

While there was no discernible error pattern on test dataset predictions, which is desirable in any statistical analysis, ternary diagrams based on the test dataset made it evident that this did not hold for the observed values (see Figure 4.29). Data points with component-wise zeros or heavily skewed towards an individual stage class returned higher error rates.

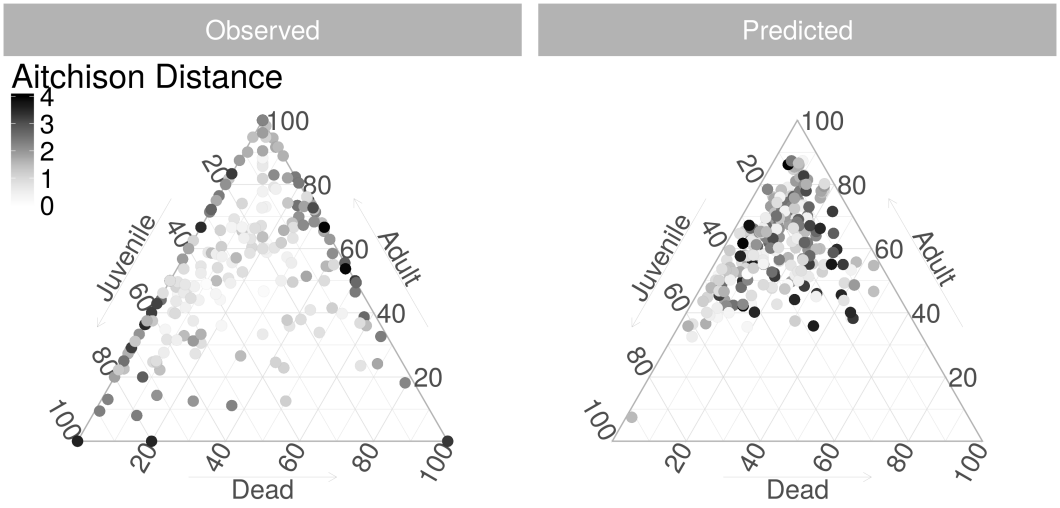


Figure 4.29: Ternary plots of test set observed *Aloidendron dichotomum* proportions (left) and their composition predictions (right). Axis arrows indicate the direction of increasing proportions for the respective stage classes. Symbol colour describes the Aitchison distance between pairs of observed and predicted proportions. A white symbol is an accurate prediction and darker symbols are poorer predictions.

Much of the test observations had more than a 80% proportional density of adults (see Figure 4.29). These observations were also slightly skewed towards smaller proportional densities of the other stage classes. Nevertheless, as a result of having to impute essential zeros, juvenile and dead stage class densities were over-predicted. This was seen in ternary plots, where despite capturing the distribution of data points, estimates were less spread out. Conversely, there was an overestimation of adult compositions for data points with low proportional values.

DISCUSSION

Distribution patterns under current climate conditions

SDM estimates presented a large overlap of habitat suitability with the surveyed area. Alongside this was a larger than observed *Aloidendron dichotomum* distribution, having identified suitable conditions outside of the surveyed area, where the species is not known to exist (see Figure 4.1a). Here, populations may be non-existent due to dispersal failure, amongst other things.

Foden (2002) put forward the suggestion that suitability predictions are accurate and that the *Aloidendron dichotomum* distribution is larger than what is actually seen, because herbivory patterns in the past were more intense. The change in herbivory patterns, which were previously limiting, will see populations extending into other territories within the Karoo.

In light of this, the reliability of distribution estimates when extrapolating to conditions outside of the sampled distribution or the future must be questioned (Guisan and Thuiller, 2005). MESS maps showed much of the study area outside of the survey zone to contain novel environmental gradients even under the current climate scenario, thus requiring models to extrapolate (see Figure 4.2). It is important to know where climates are novel because this guides where models are most uninformed and estimates require interrogation (Elith *et al.*, 2010). MESS maps aim to quantify the uncertainty associated with extrapolation. *Aloidendron dichotomum* habitat suitability was therefore captured relatively better within the survey zone than outside.

Precipitation-derived variables were generally favoured by the distribution models when determining species habitat suitability, the most important being the annual precipitation (see Figures 4.5 & 4.6). This is as anticipated, given the species' strong relation to rainfall (Midgley *et al.*, 2009). For the most part, partial dependence plots of the precipitation-derived covariates exhibited clear ranges of positive influence on species presence (see Figure 4.7).

Habitat suitability within stage classes was largely similar in geographic distribution with slight differences (see Figure 4.3). The adult stage class was estimated to generally have more areas of high suitability in comparison to the other stage classes. Furthermore, juvenile habitat suitability was poleward increasing. Variations in stage class distributions could be attributed to drought tolerance differences within stage classes (Jackson *et al.*, 2009). In addition to association with hotter temperatures, the equatorward region of the species' distribution had a greater representation of conditions highly suitable for adult populations, considering their likely resilience due to greater water storage capacity (Jack *et al.*, 2016)

High suitability estimates for the dead stage can be used to identify previously habitable environmental conditions, given the estimated absence of the other stage classes. Furthermore, conditions with a substantial composition of the dead stage class can be identified as potentially becoming unsuitable for the occurrence of live individuals. This would necessitate an investigation into the causes of mortality.

During data collection, there was no discrimination of the dead stage class (i.e adult or juvenile), which is not ideal given the differing climate tolerances of the stage classes (Jack *et al.*, 2016). The likely result would be a poor estimate of the relationship between the dead stage class and climate. Jack *et al.* (2016) found a poor relationship between climate and dead *Aloidendron dichotomum* individuals in comparison to the other two stage classes. Incorporating other important forms

of mortality on *Aloidendron dichotomum* populations in SDMs may not entirely compensate for the poor relationship with bioclimatic variables but would likely lead to better suitability and compositional estimates. Windthrow is one form of mortality that has been highlighted as such (Jack *et al.*, 2014).

Currently, juvenile stage class proportion estimates are poleward increasing and the adult stage class is proportionally dominant within the species' distribution (see Figure 4.25 & 4.28). Findings of equatorward mortality agreed with Jack *et al.* (2014), who however attributed a significant amount of *Aloidendron dichotomum* deaths to uprooting, due to a combination of shallow soils and high wind speeds. Dead individuals were estimated to only occur within the species' distribution, largely in an area which also had the greatest concentration of high suitability for the dead stage class (see Figure 4.3c & 4.25c). Estimates of the relative abundance of the dead stage class agreed with Jack (2012).

Predicted test data proportional distributions exhibited an underestimation of the adult stage class, and an over representation of the other two stage classes (see Figure 4.29). Models captured the general shape of the ternary distribution of points, but not the spread of values, a similar conclusion to Jack *et al.* (2016).

Potential impacts of climate change

Climate scenarios may see a south-easterly period-on-period shift in environmental conditions suitable for *Aloidendron dichotomum* occurrence (see Figure 4.8a & 4.10a). When investigating the effects of climate change, it is arguable to expect an immediate increase in mortality of the species in previously habitable conditions. This is attributable to the species' resilience even in unfavourable environmental gradients because of evolutionary characteristics, especially within the adult stage class (Jack, 2012).

By 2050 regions south-east of the species' current distribution were expected to experience an increase in habitat suitability, with an estimated longitudinal range

extending to Kimberly (see Figure 4.8a). This trend continued into the 2070 scenario, where in addition, greater levels of drought impact were estimated within the current species' distribution, characterised by decreasing habitat suitability (see Figure 4.10a).

Dispersal into habitable conditions outside of the current defined distribution may become possible with time, naturally or as a result of human action. Hughes *et al.* (1996) however suggest that long-lived species will struggle to advance into new habitats because they cannot migrate fast enough. Furthermore, predicted temperatures and drought prevalence make seedling survival increasingly unlikely (Jordan and Nobel, 1979).

Unlike the results presented here, Foden (2002) found no overlap between the geographic distribution of observed habitable conditions for *Aloidendron dichotomum* and the estimated suitability distributions under the future climate scenarios. Foden *et al.* (2007) reported a similar shift in the climatic envelope of poleward increasing habitat suitability.

MESS maps however were a cause for reduced confidence in future *Aloidendron dichotomum* habitat suitability estimates (see Figure 4.12). Comparison of future climates to the current survey points signalled a riskier prediction space both inside and outside the current species' distribution because of dissimilar environmental gradients. Throughout time, the species maintained the western boundary of its longitudinal range (see Figure 4.13), which is where a continued concentration of very similar environmental gradients within the species' distribution was also observed. Additional uncertainty in estimation is brought in by the initial assumption of pseudo-equilibrium, which becomes of greater concern when projecting distribution into the future (Elith *et al.*, 2010; Elith and Leathwick, 2009). The equilibrium of now may not be the same 30-50 years from now.

At a stage class level, most of the negative change in suitability given future

climate estimates occurred within the species' distribution, most notably just west of Upington (see Figure 4.14). A negative change in suitability in uninhabited areas outside of the species' distribution meant dispersal into said areas became increasingly unlikely. Estimates to the south of the leading edge of the current species' distribution suggested successful *Aloidendron dichotomum* recruitment even up until 2070.

Jack *et al.* (2016) found warmer temperatures were predominantly associated with an increase in *Aloidendron dichotomum* mortality and a decline in the proportion of adults. Likewise, here the average relative abundance of the dead stage class is expected to increase under future climate scenarios (see Table 4.1).

The relative increase in mortality however may have been associated with the failure of the compositional models to accurately predict small proportional values, coupled with estimation uncertainty from predicting into novel environmental gradients. Such a result immediately highlighted an area of potential compositional SDM methodological improvement. Without it, investigating the potential effect of climate change on population demographics would produce slightly misleading results. An emphasis on the accuracy of compositional SDMs is necessary, especially with component-wise zeros, otherwise it is an overestimation of the minority stage classes/species and underestimate of the dominant class/species.

Species distribution model inputs

Again, SDMs are often built on climate data alone, and this by itself may not be adequate enough to precisely explain potential changes in the species' distribution (Meier *et al.*, 2010). Presence/absence models are rather extensively researched in distribution modelling literature (Franklin, 2010) and have sound methodology. Focusing on identifying functionally relevant predictors would see the simplest of models competing with the more complex algorithms. This applies to distribution models in general, regardless of the target variable being modelled.

The above output, and for distribution models in general is valid only if a species is able to occupy all suitable habitats. The possibility of individuals occurring at the fringes of favourable species conditions is not given due consideration, because the modelling process rarely takes into account adaptation from evolution, or land transformation. Incorporation of some measure of species tolerance into SDMs would be beneficial and result in a larger but likely more defined distribution.

The lack of additional geological information (e.g. soil type) or other non climatic influences, may be the explanation as to why favourable *Aloidendron dichotomum* conditions are predicted east of the current defined distribution, and not seen. Attributing *Aloidendron dichotomum*'s presence/absence to more than just bioclimatic factors would be potentially gainful towards distribution modelling. In the absence of data at a finer scale required to achieve this, one could look to identifying an optimal suitability threshold with the probability distribution models. Quality of presence and absence information is key to a useful and accurate SDM and it is important to acknowledge the disadvantages that come with the data collection process (Guillera-Arroita, 2016)

CONCLUSION

This study investigated the ability of machine/statistical learning methods to approximate the suitable climate of *Aloidendron dichotomum* in its various stage classes using distributional maps. Secondary to this was considering the potential effect of climate change on the distribution of the species, and assessing the effect of different data manipulations in the hope of estimating distributions with greater precision.

A conclusion to be drawn from evidence presented, is that the macroclimate is adequate to determine *Aloidendron dichotomum* distribution patterns. Ensemble models returned an average classification accuracy rate of 80% on test data, alongside an AUC statistic of 0.88. Much of the modelled habitat suitability had an overlap with the surveyed area. Distribution estimates were however wider than expected, suggesting suitable conditions where the species is not known to exist. Habitat suitability modelled at a stage class level did not add much insight to analysis, because only minor differences in the general suitability pattern of all stage classes was observed.

An investigation into the potential impact of climate change suggested a gradual poleward shift of the species under worst case climate scenarios. Alongside this was an increasingly shrinking distribution of conditions highly suitable for *Aloidendron dichotomum* occurrence.

Whether future habitat suitability could be accurately modelled was contentious, given some of the results and limitations. The limiting correlative

nature of the constructed SDMs was of particular concern when predicting under future climate scenarios, given the likely changing collinearity structures (Dormann *et al.*, 2013). Furthermore, a substantial amount of the study area was considered to contain novel environmental gradients with reference to the sampled data points, in the future climate scenarios. SDMs were therefore extrapolating to areas with high levels of quantifiable uncertainty.

Some machine learning algorithms perform better if data is equally roughly equally distributed between response classes (He and Garcia, 2009). To investigate whether this applied to the problem at hand, class frequencies were balanced by downsampling from the larger absences class. This was done repeatedly, to average out the sampling variability from downsampling. Balancing response class frequencies in training data did not materially affect the accuracy of SDMs, which dropped by a at percentage point to 79%. The manipulation made little difference to accuracy and just redistributed the driver of overall accuracy from absences to presences.

The addition of absence points outside of the known species range has been found to generate more defined potential distributions and return higher accuracy scores (Chefaoui and Lobo, 2008). This manipulation was investigated by randomly sampling a variable number of points outside of the study area, to append to training data as absences. Adding generated *Aloidendron dichotomum* absences outside of the study area from where it is known not to exist did not affect SDM accuracy on observed presences and absences, which slightly deviated around approximately 66% and 87% respectively. Appending the data points however did impact the distribution of estimated suitability within the survey zone, making the manipulation undesirable.

In comparison to stage class habitat suitability, compositional data analysis gave some insight into the relative distribution of stage classes across space. The adult population was determined to be the generally proportionally dominant stage

class, however this relative dominance was not constant across space. Compositional estimates saw poleward increasing juvenile proportions and the opposite held for the adult stage class. Dead individuals were predominantly concentrated within the species range. Considering the tolerance of the adult stage class to varying climate conditions, a shift in the suitable climate of *Aloidendron dichotomum* is unlikely to be followed by an immediate shift in observed species presences. However, much of the future predicted presences outside of the survey area had estimated juvenile compositions larger than that of the adult stage class.

Despite the slight deviation from ecological theory, SDMs proved to be an effective tool in monitoring species distribution patterns. Reflection on past and current SDM literature revealed a significant amount of potential for SDM improvements. Going forward, with respect to *Aloidendron dichotomum* presence, access to additional functionally relevant variables or data at a finer scale, has the potential to produce greater precision in distribution models. Additionally, a revised sampling technique, not restricted to the road network, would likely increase the availability of presence records.

With SDMs in general, addressing the shortcomings from a methodological standpoint is recommended. Emphasis on this would be greatly beneficial to the scientific community both in the biodiversity and statistical learning fields, potentially sparking new areas of interest.

Bibliography

Aitchison, J. (1983). Principal component analysis of compositional data, *Biometrika* pp. 57–65.

Aitchison, J. (2003). A concise guide to compositional data analysis, cda work, *Girona* **24**: 73–81.

Aitchison, J. and Egozcue, J. J. (2005). Compositional data analysis: where are we and where should we be heading?, *Mathematical Geology* **37**(7): 829–850.

Aitchison, J., Kay, J. W. *et al. (2003). Possible solution of some essential zero problems in compositional data analysis.*

Araújo, C. B., Marcondes-Machado, L. O. and Costa, G. C. (2014). The importance of biotic interactions in species distribution models: a test of the eltonian noise hypothesis using parrots, Journal of Biogeography **41***(3): 513–523.*

Araújo, M. B. and New, M. (2007). Ensemble forecasting of species distributions, Trends in ecology & evolution **22***(1): 42–47.*

Archer, K. J. and Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures, Computational Statistics & Data Analysis **52***(4): 2249–2260.*

Barbet-Massin, M., Jiguet, F., Albert, C. H. and Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: how, where and how many?, Methods in Ecology and Evolution **3***(2): 327–338.*

Bates, J. M. and Granger, C. W. (1969). The combination of forecasts, Journal of the Operational Research Society **20***(4): 451–468.*

Breiman, L. (2001). Random forests, Machine learning **45***(1): 5–32.*

Buckley, L. B., Urban, M. C., Angilletta, M. J., Crozier, L. G., Rissler, L. J. and Sears, M. W. (2010). Can mechanism inform species' distribution models?, Ecology letters **13***(8): 1041–1054.*

Chefaoui, R. M. and Lobo, J. M. (2008). Assessing the effects of pseudo-absences on predictive distribution model performance, Ecological modelling **210***(4): 478–486.*

De'Ath, G. (2007). Boosted trees for ecological modeling and prediction, Ecology **88***(1): 243–251.*

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J. et al. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance, Ecography **36***(1): 27–46.*

Drummond, C., Holte, R. C. et al. (2003). C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling, Workshop on learning from imbalanced datasets II*, Vol. 11, Citeseer.*

Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC. M. Overton, J., Townsend Peterson, A., J. Phillips, S., Richardson, K., Scachetti-Pereira, R., E. Schapire, R., Soberón, J., Williams, S., Wisz, M. S. and Zimmermann, N. E. (2006). Novel methods improve prediction of species' distributions from occurrence data, Ecography **29***(2): 129–151.*

Elith, J., Kearney, M. and Phillips, S. (2010). The art of modelling range-shifting species, Methods in ecology and evolution **1***(4): 330–342.*

Elith, J. and Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time, Annual Review of Ecology, Evolution, and Systematics **40***(1): 677.*

Elith, J., Leathwick, J. R. and Hastie, T. (2008). A working guide to boosted regression trees, Journal of Animal Ecology **77***(4): 802–813.*

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E. and Yates, C. J. (2011). A statistical explanation of maxent for ecologists, Diversity and distributions **17***(1): 43–57.*

Fawcett, T. (2004). Roc graphs: Notes and practical considerations for researchers, Machine learning **31***(1): 1–38.*

Foden, W. (2002). A Demographic Study ofAloe dichotoma in the Succulent Karoo: Are the Effects of Climate Change Already Apparent?*, unpublished PhD thesis, University of Cape Town, South Africa.*

Foden, W., Midgley, G. F., Hughes, G., Bond, W. J., Thuiller, W., Hoffman, M. T., Kaleme, P., Underhill, L. G., Rebelo, A. and Hannah, L. (2007). A changing climate is eroding the geographical range of the namib desert tree aloe through population declines and dispersal lags, Diversity and Distributions **13***(5): 645–653.*

Franklin, J. (2010). Mapping species distributions: spatial inference and prediction*, Cambridge University Press.*

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine, Annals of statistics *pp. 1189–1232.*

Fritsch, S. and Guenther, F. (2016). neuralnet: Training of Neural Networks. *R package version 1.33.*
***URL:** https://CRAN.R-project.org/package=neuralnet*

Garg, A. and Tai, K. (2013). Comparison of statistical and machine learning methods in modelling of data with multicollinearity, International Journal of Modelling, Identification and Control **18***(4): 295–312.*

Giorgetta, M. A., Jungclaus, J., Reick, C. H., Legutke, S., Bader, J., Böttinger, M., Brovkin, V., Crueger, T., Esch, M., Fieg, K. et al. (2013). Climate and carbon

cycle changes from 1850 to 2100 in mpi-esm simulations for the coupled model intercomparison project phase 5, Journal of Advances in Modeling Earth Systems **5***(3): 572–597.*

Guillera-Arroita, G. (2016). Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities, Ecography .

Guisan, A. and Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models, Ecology letters **8***(9): 993–1009.*

Hastie, T., Friedman, J. and Tibshirani, R. (2001). The elements of statistical learning*, Vol. 1, Springer series in statistics Springer, Berlin.*

He, H. and Garcia, E. A. (2009). Learning from imbalanced data, IEEE Transactions on knowledge and data engineering **21***(9): 1263–1284.*

Hecht-Nielsen, R. (1992). Theory of the backpropagation neural network, Neural networks for perception*, Elsevier, pp. 65–93.*

Hijmans, R. J., Phillips, S., Leathwick, J. and Elith, J. (2017). dismo: Species Distribution Modeling. *R package version 1.1-4.*
***URL:** https://CRAN.R-project.org/package=dismo*

Hoffman, M. T., Carrick, P., Gillson, L. and West, A. (2009). Drought, climate change and vegetation response in the succulent karoo, south africa, South African Journal of Science **105***(1-2): 54–60.*

Hughes, G. O., Thuiller, W., Midgley, G. F. and Collins, K. (2008). Environmental change hastens the demise of the critically endangered riverine rabbit (bunolagus monticularis), Biological Conservation **141***(1): 23–34.*

Hughes, L., Cawsey, E. and Westoby, M. (1996). Climatic range sizes of eucalyptus species in relation to future climate change, Global Ecology and Biogeography Letters *pp. 23–29.*

Ishwaran, H. and Kogalur, U. (2017). Random Forests for Survival, Regression, and Classification (RF-SRC). *R package version 2.5.0.*
URL: *https://cran.r-project.org/package=randomForestSRC*

Jack, S. (2012). Revisiting Aloe dichotoma's suitability as an indicator of climate change in southern Africa, *unpublished PhD thesis, Masters dissertation. University of Cape Town.*

Jack, S., Hoffman, M., Rohde, R. and Durbach, I. (2016). Climate change sentinel or false prophet? the case of aloe dichotoma, Diversity and Distributions **22***(7): 745–757.*

Jack, S. L., Hoffman, M. T., Rohde, R. F., Durbach, I. and Archibald, M. (2014). Blow me down: A new perspective on aloe dichotoma mortality from windthrow, BMC ecology **14***(1): 1.*

Jackson, D. A. (1997). Compositional data in community ecology: the paradigm or peril of proportions?, Ecology **78***(3): 929–940.*

Jackson, S. T., Betancourt, J. L., Booth, R. K. and Gray, S. T. (2009). Ecology and the ratchet of events: climate variability, niche dimensions, and species distributions, Proceedings of the National Academy of Sciences **106***(Supplement 2): 19685–19692.*

Jordan, P. W. and Nobel, P. S. (1979). Infrequent establishment of seedlings of agave deserti (agavaceae) in the northwestern sonoran desert, American Journal of Botany *pp. 1079–1084.*

Kaastra, I. and Boyd, M. (1996). Designing a neural network for forecasting financial and economic time series, Neurocomputing **10***(3): 215–236.*

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest, R News **2***(3): 18–22.*
URL: *http://CRAN.R-project.org/doc/Rnews/*

Manel, S., Dias, J.-M. and Ormerod, S. J. (1999). Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a himalayan river bird, Ecological modelling **120***(2): 337–347.*

Margules, C. and Austin, M. (1990). Nature conservation: cost effective biological surveys and data analysis*, CSIRO PUBLISHING.*

Martín-Fernández, J. A., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation, Mathematical Geology **35***(3): 253–278.*

Martín-Fernández, J. and Thió-Henestrosa, S. (2006). Rounded zeros: some practical aspects for compositional data, Geological Society, London, Special Publications **264***(1): 191–201.*

Max Planck Institute for Meteorology (2016). Mpi-esm – verc.
***URL:** https://verc.enes.org/models/earthsystem-models/mpi-m/mpi-esm*

Meier, E. S., Edwards Jr, T. C., Kienast, F., Dobbertin, M. and Zimmermann, N. E. (2011). Co-occurrence patterns of trees along macro-climatic gradients and their potential influence on the present and future distribution of fagus sylvatica l., Journal of Biogeography **38***(2): 371–382.*

Meier, E. S., Kienast, F., Pearman, P. B., Svenning, J.-C., Thuiller, W., Araújo, M. B., Guisan, A. and Zimmermann, N. E. (2010). Biotic and abiotic variables show little redundancy in explaining tree species distributions, Ecography **33***(6): 1038–1048.*

Merow, C., Allen, J. M., Aiello-Lammens, M. and Silander, J. A. (2016). Improving niche and range estimates with maxent and point process models by integrating spatially explicit information, Global Ecology and Biogeography **25***(8): 1022–1036.*

Merow, C., Smith, M. J. and Silander, J. A. (2013). A practical guide to maxent for modeling species' distributions: what it does, and why inputs and settings matter, Ecography **36***(10): 1058–1069.*

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. and Leisch, F. (2017). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. *R package version 1.6-8.*
URL: *https://CRAN.R-project.org/package=e1071*

Midgley, G., Altwegg, R., Guo, D. and Merow, C. (2009). Are quiver trees a sentinel for climate change in arid southern africa, Cape Town: The South African National Biodiversity Institute .

Midgley, G. F., Chown, S. L. and Kgope, B. S. (2007). Monitoring effects of anthropogenic climate change on ecosystems: A role for systematic ecological observation?, South African Journal of Science **103***(7-8): 282–286.*

Midgley, J., Cowling, R., Hendricks, H., Desmet, P., Esler, K. and Rundel, P. (1997). Population ecology of tree succulents (aloe and pachypodium) in the arid western cape: decline of keystone species, Biodiversity & Conservation **6***(6): 869–876.*

Miller, P. (2016). mvtboost: Tree Boosting for Multivariate Outcomes. *R package version 0.5.0.*
URL: *https://CRAN.R-project.org/package=mvtboost*

Moss, R. H., Edmonds, J. A., Hibbard, K. A., Manning, M. R., Rose, S. K., Van Vuuren, D. P., Carter, T. R., Emori, S., Kainuma, M., Kram, T. et al. (2010). The next generation of scenarios for climate change research and assessment, Nature **463***(7282): 747–756.*

Nassiri, H., Najaf, P. and Amiri, A. M. (2014). Prediction of roadway accident frequencies:: Count regressions versus machine learning models, Scientia Iranica. Transaction A, Civil Engineering **21***(2): 263.*

Owens, H. L., Campbell, L. P., Dornak, L. L., Saupe, E. E., Barve, N., Soberón, J., Ingenloff, K., Lira-Noriega, A., Hensz, C. M., Myers, C. E. et al. (2013). Constraints on interpretation of ecological niche models by limited environmental ranges on calibration areas, Ecological Modelling **263***: 10–18.*

Pawlowsky-Glahn, V. and Buccianti, A. (2011). Compositional data analysis: Theory and applications, *John Wiley & Sons.*

Pawlowsky-Glahn, V., Egozcue, J. J. and Tolosana-Delgado, R. (2015). Modeling and analysis of compositional data, *John Wiley & Sons.*

Phillips, S. J., Anderson, R. P. and Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions, Ecological modelling **190***(3): 231–259.*

Phillips, S. J. and Dudík, M. (2008). Modeling of species distributions with maxent: new extensions and a comprehensive evaluation, Ecography **31***(2): 161–175.*

Qu, Y., Adam, B.-L., Yasui, Y., Ward, M. D., Cazares, L. H., Schellhammer, P. F., Feng, Z., Semmes, O. J. and Wright, G. L. (2002). Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients, Clinical chemistry **48***(10): 1835–1843.*

R Core Team (2017). R: A Language and Environment for Statistical Computing, *R Foundation for Statistical Computing, Vienna, Austria.*
***URL:** https://www.R-project.org/*

Ridgeway, G. (2017). gbm: Generalized Boosted Regression Models. *R package version 2.1.3.*
***URL:** https://CRAN.R-project.org/package=gbm*

Riedmiller, M. and Braun, H. (1993). A direct adaptive method for faster backpropagation learning: The rprop algorithm, Neural Networks, 1993., IEEE International Conference on, *IEEE, pp. 586–591.*

Riley, S. J., DeGloria, S. D. and Elliot, R. (1999). A terrain ruggedness index that quantifies topographic heterogeneity, intermountain Journal of sciences **5***(1-4): 23–27.*

Scholkopf, B., Sung, K.-K., Burges, C. J., Girosi, F., Niyogi, P., Poggio, T. and Vapnik, V. (1997). Comparing support vector machines with gaussian

kernels to radial basis function classifiers, IEEE transactions on Signal Processing **45***(11): 2758–2765.*

Senay, S. D., Worner, S. P. and Ikeda, T. (2013). Novel three-step pseudo-absence selection technique for improved species distribution modelling, PloS one **8***(8): e71218.*

Shore, J. and Johnson, R. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy, IEEE Transactions on information theory **26***(1): 26–37.*

Suykens, J. A. and Vandewalle, J. (1999). Least squares support vector machine classifiers, Neural processing letters **9***(3): 293–300.*

Tadross, M., Jack, C. and Hewitson, B. (2005). On rcm-based projections of change in southern african summer climate, Geophysical Research Letters **32***(23).*

Thibaud, E., Petitpierre, B., Broennimann, O., Davison, A. C. and Guisan, A. (2014). Measuring the relative effect of factors affecting species distribution model predictions, Methods in Ecology and Evolution **5***(9): 947–955.*

Thuiller, W. (2004). Patterns and uncertainties of species' range shifts under climate change, Global Change Biology **10***(12): 2020–2027.*

Thuiller, W., Lafourcade, B., Engler, R. and Araújo, M. B. (2009). Biomod–a platform for ensemble forecasting of species distributions, Ecography **32***(3): 369–373.*

Verner, J., Morrison, M. L., Ralph, C. J. et al. (1986). Wildlife 2000. Modeling habitat relationships of terrestrial vertebrates.*, University of Wisconsin Press.*

Vogel, J. (1974). The life span of the kokerboom, Aloe **12***(2): 66–68.*

Wiens, J. A., Stralberg, D., Jongsomjit, D., Howell, C. A. and Snyder, M. A. (2009). Niches, models, and climate change: assessing the assumptions and uncertainties, Proceedings of the National Academy of Sciences **106***(Supplement 2): 19729–19736.*

WorldClim (2016). WorldClim - Global Climate Data bioclim.
URL: *http://www.worldclim.org/bioclim*

Zimmermann, N. E., Edwards, T. C., Graham, C. H., Pearman, P. B. and Svenning, J.-C. (2010). New trends in species distribution modelling, Ecography **33***(6): 985–989.*

Zurell, D., Elith, J. and Schröder, B. (2012). Predicting to new environments: tools for visualizing model behaviour and impacts on mapped distributions, Diversity and Distributions **18***(6): 628–634.*

APPENDIX

Hyper-parameter selection

Results reported below are the cross validated species detection accuracy of the various statistical learning models, from hyper-parameter selection. The ensemble model is not included here, because it did not require parameter selection.

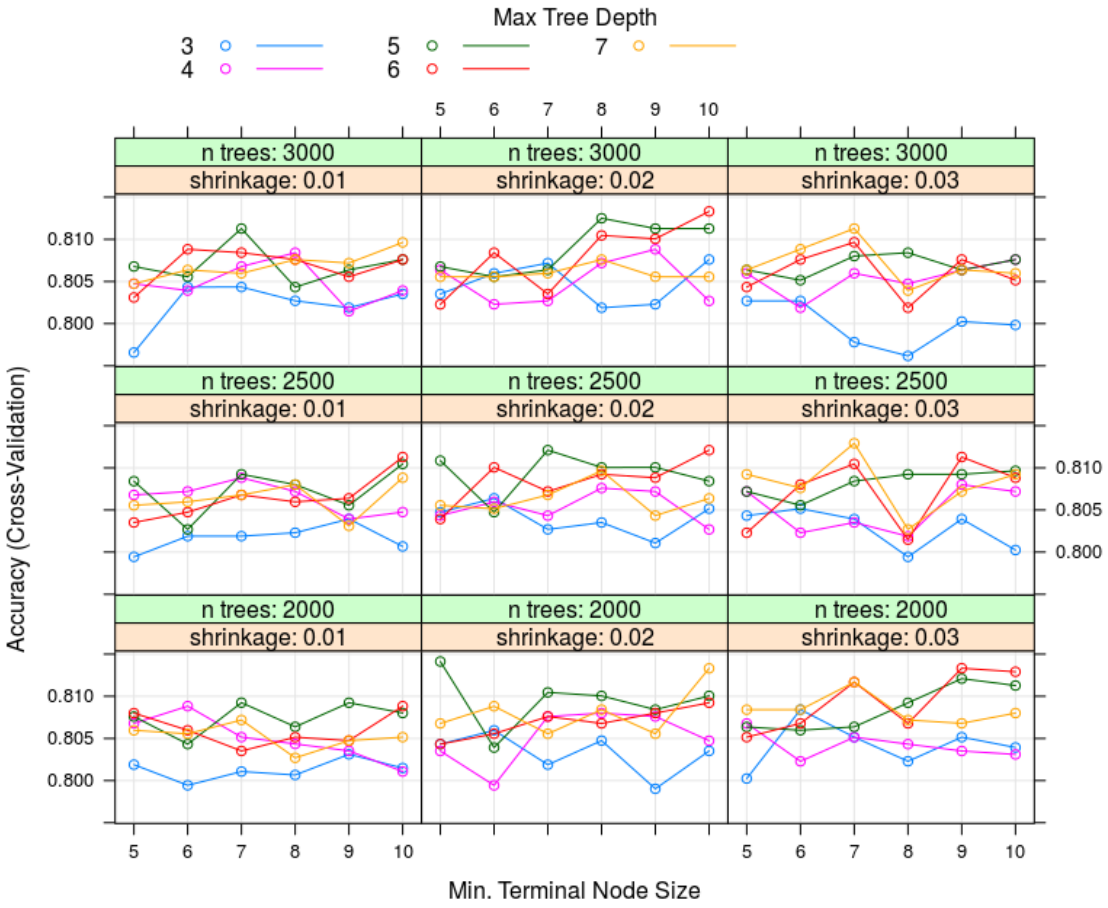


Figure A.1: Boosted model parameter tuning results

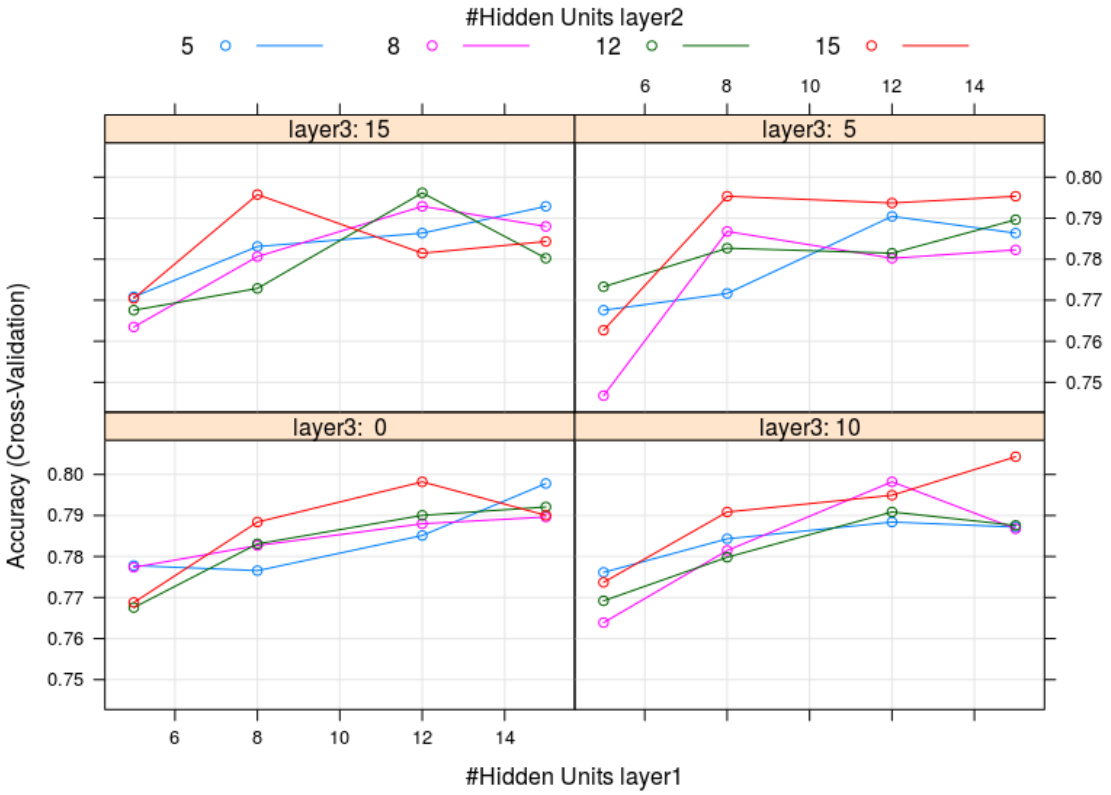


Figure A.2: Neural network parameter tuning results

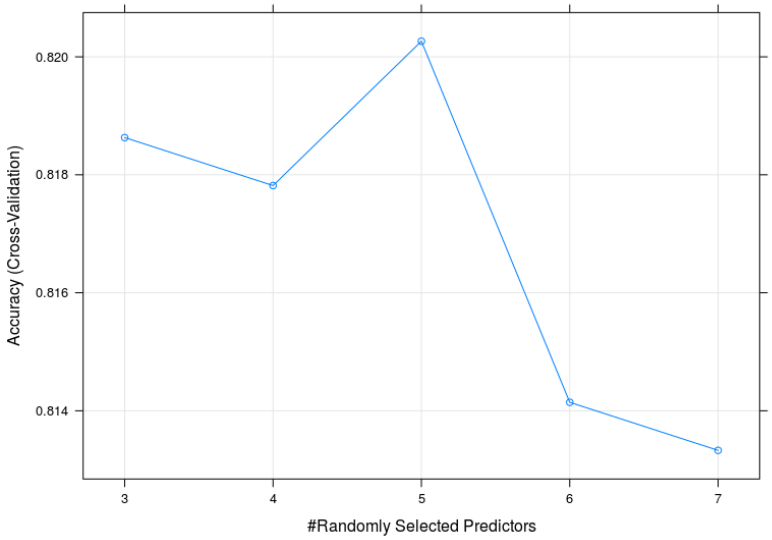


Figure A.3: Random forest parameter tuning results

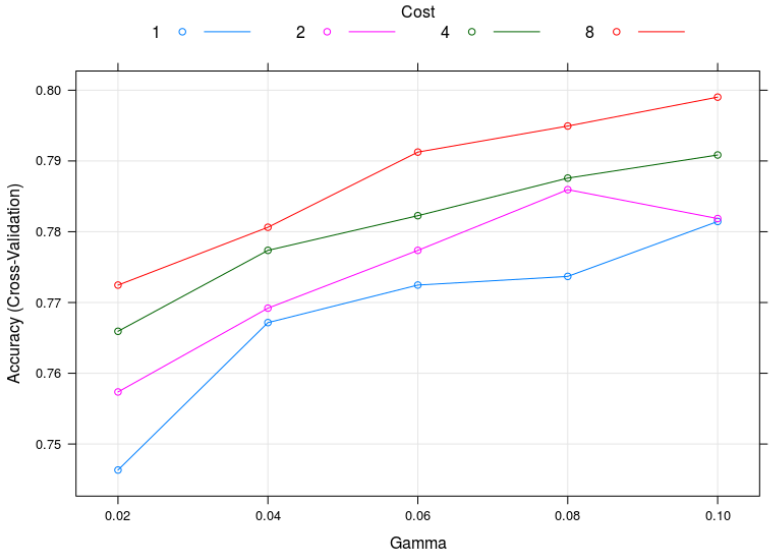


Figure A.4: Support vector machine parameter tuning results