Predicting reptile species distributions and biogeographic patterns within Kruger National Park

Jody Michael Barends

University of the Western Cape, Department of Biodiversity and Conservation Biology, Private, bag X 17, Bellville, 7535, South Africa.

A thesis submitted in fulfilment of the requirements for the degree of *Magister Scientiae* in the Department of Biodiversity and Conservation Biology, Faculty of Science, University of the Western Cape

Supervisor: Dr. Bryan Maritz

October 2018

Predicting reptile species distributions and biogeographic patterns within Kruger National Park

Jody Michael Barends

Keywords

Biodiversity monitoring; Biogeography; Data bias; Geographic information systems (GIS); Kruger National Park; MaxEnt; Reptile fauna, SANParks; Species distribution modelling; Thresholds of potential concern.

Abstract

Predicting reptile species distributions and biogeographic patterns within Kruger National Park

J.M Barends

MSc. Thesis, Department of Biodiversity and Conservation Biology, University of the Western Cape.

Knowledge of global reptile ecology is limited and there remains much to understand in terms of detailed reptile species information, including that of their distributions. In South Africa, despite being one of SANParks best-studied reserves, surprisingly little is known about the distributions and spatial ecology of reptiles within Kruger National Park (KNP). Management within KNP follows a strategic adaptive management strategy which monitors the statuses of animals using species or group specific indicators. Indicators are given predetermined upper and lower ranges of acceptable fluctuation before actions are taken. These ranges are referred to as thresholds of potential concern (TPCs), and for reptiles these are based on changes to their distributions across the landscape of KNP.

An apparent lack of high-quality reptile distribution data inhibits the effective monitoring of the statuses of these animals within KNP, which in turn limits management and conservation options. In this study, I use several methods to quantify available reptile occurrence data which formed the foundations for predicting the distributions of these species across KNP by means of species distribution modelling, with a view to gaining novel insight into reptile assemblage structure across the landscape of KNP.

I collated 7118 locality records representing 127 reptile species occurring within KNP. In quantifying these, I effectively performed a gap analysis of KNP and found that large areas of the park were poorly-sampled, with nearly 68 % of all records occurring within 2 km of infrastructure. Despite challenges relating to spatial scale and data bias, using an ecological niche modelling approach I predicted the geographical distributions of 119 reptile species across KNP at a resolution of 1 km x 1 km. I used these distributions to infer species' presence or absence at any given 1 km x 1 km grid cell across KNP and to subsequently quantify assemblage membership via hierarchical cluster analyses. I predicted that at least nine taxonomically distinct, spatially-segregated reptile assemblages are present within KNP, with each appearing to be correlated with changes in landscape features across the park.

My work has identified important gaps in our understanding of the distributions of reptiles in KNP that will drive future sampling efforts. Moreover, my modelled predictions offer multiple testable hypotheses in terms of species' presences and absences that will direct future research efforts in KNP, and potentially aid in the monitoring of reptiles more generally across protected areas within South Africa.

Declaration

I declare that "**Predicting reptile species distributions and biogeographic patterns within Kruger National Park***"* is my own work, that it has not been submitted for any degree or examination at any university, and that all sources I have used or quoted have been indicated and acknowledged by complete references.

Full name: Jody Michael Barends **Date:** October 2018

Signature: .....

Acknowledgments

Special thanks go to Bryan Maritz, Robin Maritz, Donovan Tye, Graham Alexander, and all others involved in the Reptile Diversity in African Savannas (RDAS) field course. This includes UWC students Ielhaam Bassier, Riaaz Mohamed, and Silindokuhle Tokota, as well as several students from University of Mpumulanga and University of Venda who assisted me with field sampling. I would like to thank Herpetological Conservation International (HCI), Nsasani Trust, Organisation for Tropical Studies (OTS), African Snakebite Institute, and Comair Limited for their funding and donations. The efforts of all involved allowed me to complete a critical component of my project and this was greatly appreciated.

I would like to thank my colleagues within the Maritz lab who assisted me throughout the duration of this project. Their support, discussions, and friendships have greatly aided in the completion of this thesis. To my family and friends, thank you for providing a constant stream of support and encouragement. A special thank you goes to Ielhaam Bassier for always being there for me, academically and otherwise, without whom the enjoyment of my time working on this project would have been significantly less.

I would also like to acknowledge Darren Petersen, Guin Zambatis, the Ditsong Museum, OTS, and the Animal Demography Unit for providing me with reptile occurrence data.

To Bryan Maritz, your guidance and supervision has been consistently provided to me at the highest degree possible. All of your efforts and contributions towards my academic career and experiences have been appreciated beyond measure.

Lastly, thank you to those who provided funding towards the completion of this project. This includes the NRF, Ada and Bertie Levenstein foundation, and Merseta. Through the combined contributions of these organisations my financial expenses were covered which allowed me to focus all of my efforts on the successful completion of this thesis.

Contents

List of Tables

List of Figures

List of Appendices

Preface

The core of this thesis involved predicting the distributions of all reptile species occurring within Kruger National Park in order to assess the viability of using a predictive framework as an alternative to on-site sampling as a means of monitoring changes in the populations of these animals. This primarily involved the use of species distribution modelling to not only predict where each individual reptile species occurs within and across Kruger National Park, but also identify reptile assemblages within this same geographical space. Whilst I was able to successfully make these predictions, I found that the models I produced were not to as high a standard as expected. Each species distribution model was evaluated individually and although I deemed overall model performance across all species as acceptable, several of these were unlikely to be accurate representations of those species true distributions as they did not appear to make biological sense in accordance to what would be expected in reality. As such, the results I obtained here relating to reptile species distributions and the subsequent analyses thereof were interpreted with the knowledge that they represented probable rather than true distributional patterns of reptiles within Kruger National Park.

Chapter 1: General introduction

1.1 Biodiversity loss and its management in South Africa

Biodiversity loss is an ongoing global problem (Myers et al. 2000; Butchart et al. 2010; Cardinale 2012). Even within areas specifically designated to protect and conserve biodiversity, species loss still occurs, and the management thereof remains a complex and challenging issue (Jackson and Gaston 2008). To prevent loss of species within protected areas, governing authorities have developed a range of tools, policies, and strategies which aim to combat species loss and maintain biological diversity within their jurisdictions (Martin et al. 2009). However, since each protected area has unique conservation goals and objectives these approaches vary across the global protected area network (Chape et al. 2005).

In South Africa, the majority of protected areas are managed by South African National Parks (SANParks), an organization that makes use of a strategic adaptive management (SAM) strategy to achieve their conservation objectives (Parr et al. 2009; Roux and Foxcroft 2011). This strategy involves a 'learn by doing' approach that functions on specific informational needs to ensure that their key mandate of maintaining maximum biological diversity within their borders is met (Ferreira et al. 2011). This approach heavily relies on monitoring biodiversity to assess whether any predetermined thresholds of potential concern (TPCs) are crossed (Venter et al. 2008). TPCs are ranges of upper and lower limits of variables which are used to monitor the statuses of taxonomic groups and other variable environmental conditions within protected areas (Gillson and Duffin 2007; Parr et al. 2009). In the event of a TPC being crossed, SANParks undertakes necessary action to attempt to mitigate any undesired effects like loss of species.

1.2 Reptile monitoring and TPCs

TPCs are taxon-specific, with monitoring of some groups being easier than others. For reptiles as a group, TPCs are based on monitoring changes in reptile populations at given sites at three year intervals (Ferreira et al. 2011), but this difficult to achieve. As a group, reptiles are generally difficult to study and monitor because of the low and variable detection rates that these animals possess, especially at local scales (Durso et al. 2011). Detectability for many reptile species is so low that they could remain undetected within intensively sampled areas for years at a time (Mazerolle et al. 2007). Moreover, reptiles have low dispersal capabilities in general and are unlikely to leave suitable habitats (Sahlean et al. 2014). As such, these animals are therefore likely to remain present within sites for long periods of time but may not always be detected during sampling. This presents a potential flaw within current monitoring protocols as animals could be incorrectly deemed absent from sites, and as detecting these animals is of critical importance towards observing changes in TPCs these inaccuracies could inhibit assessments.

Ecologists have tried to statistically account for the low detection rates of various groups of animals, including reptiles (Thomas et al. 2010). For example, the field of occupancy modelling was developed to account for imperfect detectability of species during sampling due to low detection rates (Kery 2011). These types of models allow us to estimate true populations and determine the probability of detecting a species at a given site within a given detection event. This could then be used to infer the presence or absence of a species within said site even if it is not detected during sampling, and this approach has been successful in previous studies (for example: Mcgrath et al. 2015). However, the detectability of many reptile species is so low that an occupancy modelling approach would not be feasible for monitoring as it requires extensive presence/absence data from multiple sampling events that the current monitoring protocol simply does not allow for.

Reptile data is often limited or poor (Lee and Jetz 2010; Bates et al. 2014; Jetz and Freckleton 2015), and poor data hinders the TPC approach within the SAM strategy (Ferreira et al. 2011). Low detectability of reptile species is the number one reason why the currently defined idea of monitoring changes to reptile populations to assess TPCs will not be successful. In addition, issues relating to spatial scale also hinders these approaches. Reptile occurrence data are often collected at a relatively fine-scale, but most reptile species are broadly distributed across KNP. This mismatch can potentially cause confusion and inhibit analyses that rely on these data as spatial scale is an important factor to consider in the application of species occurrence data.

1.3 Problem statement

The management and monitoring of reptile populations in protected areas in South Africa is hindered by a lack of high quality data as a result of reptiles possessing extremely low detection probabilities. The absence of such data hinders our ability to assess the distributions of reptile species, monitoring protocols, and predict reptile communities. To fix this, here I attempt to quantify available reptile occurrence data, assess its biases, and develop a framework for predicting reptile species distributions and community assemblages.

1.4 Study area

My study focussed on Kruger National Park (KNP), the largest protected area in South Africa. KNP spans an area of approximately 20 000 km^2 across the Limpopo and Mpumalanga provinces (Figure 1.1) and is home to a vast array of biodiversity. KNP is home to nearly 2000 floral species and more than 850 faunal species (Parr et al. 2009). Of these, there are approximately 34 amphibian species (Du Toit et al. 2003), 505 bird species (Harrison et al. 1997), 49 fish species (Skelton et al. 2001), 147 mammal species (Du Toit et al. 2003), and 120 reptile species (Branch 1998; Bates et al. 2014).

The high diversity and richness of biodiversity within KNP is largely due to the heterogeneous savanna ecosystem encompassing the park (Venter et al. 2008). KNP is primarily dominated by large areas of subtropical woodlands with an assortment of vegetation types present that provides suitable habitats for numerous species (Scholtz et al. 2014). Furthermore, KNP has a range of differing geologies with several substrate types being present throughout. This includes basaltic and granite rocks which weather into clay-rich and sandy soils respectively (Kulmatiski et al. 2017). In terms of climate, KNP is classified as a summer-rainfall region and experiences an average of 300 – 500 mm, and 500 – 700 mm of rainfall annually within the northern and southern regions of the park respectively (Du Toit et al. 2003; Kulmatiski et al. 2017).

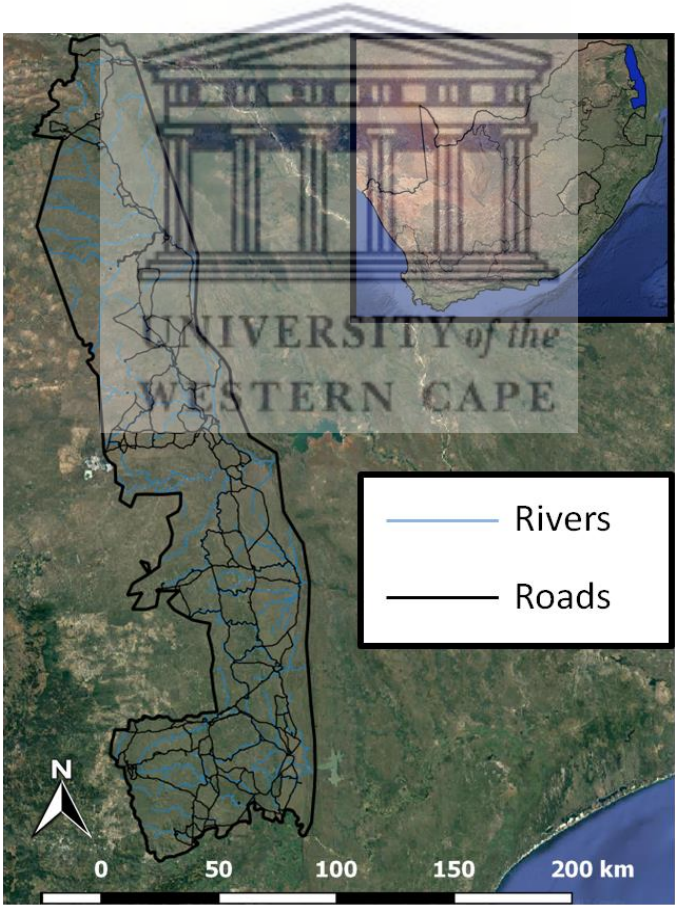


Figure 1.1: Geographical location and aerial view of Kruger National park, highlighting riverine and infrastructural areas.

1.5 Approach

1.5.1 Data:

Reptiles are often difficult to study as data are not always readily available for these animals (Powney et al. 2010; Bohm et al. 2013). This is largely because these taxa tend to have highly cryptic lifestyles which makes them difficult to detect (Maritz et al. 2007; Mcgrath et al. 2015). This creates particular challenges relating to our knowledge of their distributions, especially at localized scales (Bates et al. 2014). To properly understand and conserve a species it is important to know where that species occurs, and this largely requires species occurrence data. This type of data provides information relating to where individuals were observed, usually in the form of GPS co-ordinates with locality descriptions. Reptile species occurrence data formed the foundation of this study.

1.5.2 Spatial scale:

Choosing an appropriate spatial scale for data analyses is an important consideration for studies relating to the application of species occurrence data (Hurlbert and Jetz 2007). The scale at which data are assessed directly impacts on observable patterns and trends, and therefore selecting an appropriate spatial scale is critical to testing different hypotheses. Choice in spatial scale may be selected based on several criteria, including: the scale at which data were collected, the scale of additional variables (such as climate data), the extent of the area of the study site, the ecology of the species in question, or the specific hypotheses to be tested (Atkinson and Tate 2000; Rahbek 2005). As such, there is no universally accepted appropriate scale, with resolutions varying per study. For example, South African atlas projects such as FrogMAP and SABAP have operated at relatively broad resolutions of quarter degree grid cells (approximately 15 km x 15 km; Minter et al. 2004; Underhill and

Brooks 2016) whereas studies relating to species distribution modelling tend to operate at severely finer resolutions of 1 km x 1 km (For example, Pearson et al. 2007).

Here, I opted for a spatial scale of 1 km x 1 km for several reasons. Primarily, I chose this resolution as I aimed to predict reptile distributions as finely-scaled as possible and this scale represented the finest resolution at which reliable climate data, which is critical for species distribution modelling, was available (Hijmans et al. 2005). Secondly, KNP is significantly smaller in area than the study sites of the above mentioned atlas projects which operated at a national scale. For this study, operating at as broad a spatial scale as used in those studies would be detrimental to assessing trends in reptile occurrences within the context of localized sites within KNP and so it made sense to use a smaller spatial scale. Lastly, A fine-scale resolution offers more precise outputs than broader datasets and are ideal for quantifying biological patterns like distributions and regionalization (Kreft and Jetz 2010).

1.5.3 Chapter focus:

In this thesis, chapter 2 directly investigates trends in currently available reptile occurrence data. By collating and geospatially filtering all available records of reptile occurrences from numerous sources at a scale of 1 km x 1 km I was able to gauge patterns of known reptile species richness across KNP using GIS techniques. Additionally, I was able to assess the extent of data gaps within the park where reptile occurrences were few or not available and infer patterns of biases in reptile sampling. I then discuss the implications of the limited nature of these occurrence data and its use within species distribution modelling.

Chapter 3 focuses on predicting reptile species distributions across KNP using an ecological niche modelling approach via Maximum Entropy (MaxEnt) software. In this chapter I produce models predicting distributional ranges for nearly all known reptile species occurring within KNP, from which I then estimated reptile species richness across KNP at a spatial

scale of 1 km x 1 km. This approach is recommended as a potential alternative to on-site sampling for the purposes of monitoring changes in reptile populations (Ferreira et al. 2011), and I discuss the potential implementation of such an approach including the limitations and drawbacks. In addition, I provide insight into potential reptile spatial ecology across KNP.

In chapter 4 I describe the process and results of empirically evaluating the predictions of reptile species distributions I made in chapter 3 via ground-truthing on-site within KNP. This involved the use of standardized reptile sampling techniques to capture reptiles to produce a testable dataset to compare model predictions against.

In chapter 5 I use the predictions of reptile species distributions made in chapter 3 as a baseline to infer reptile assemblages across KNP. Using clustering techniques, I provide insight into biogeographical patterns of reptiles in KNP by delineating the park into spatially segregated biogeographic units based on compositional similarity between grid cells. In this chapter I address the benefits and challenges of dividing KNP into separate units for conservation purposes.

In chapter 6, I provide a synthesis of my findings and summarize the major conclusions of this thesis. I conclude with recommendations for future studies.

1.6 Major limitations

Major analyses within this thesis were dependent on reptile species occurrence data and these data were limited. Occurrence data were collated from several sources and data quality was not uniform amongst these. A large portion of recorded reptile occurrences were not provided with accurate GPS co-ordinates, and required estimates based on locality descriptions which were not always unambiguous. Given that the core of this project was modelling based, this was not a major constraint as the software I used (MaxEnt) is not overtly affected by subtle changes in GPS co-ordinates (Baldwin 2009), but other analyses may have been affected,

such as assessments of the spatial arrangements of reptiles across KNP, and geographical biases of data in relation to environmental factors. More importantly however, these data were biased (described in further chapters) and although I took steps to mitigate the effects of these biases it is unlikely that the results of all analyses were completely unaffected.

Chapter 2: A gap analysis of KNP reptile occurrence data

2.1 Introduction

Mapping the distributions of species is an important tool in biodiversity conservation and management. Not only are maps of the actual or potential distributions of species important for ecological research (Storch et al. 2003; Franklin 2010), they are also useful in furthering our understanding of the formations of biological assemblages. By knowing where each species occurs within a given area, we are able to identify which of these occur together spatially to form localized biological communities or assemblages (Feria and Peterson 2002; Kreft and Jetz 2010). Moreover, insight into the spatial arrangements of species within protected areas could inform conservation or management decisions (Ferrier and Guisan 2006). In the absence of intensive sampling, quantifying species arrangements are usually dependent on high quality species distribution maps, but these are often unavailable.

Species distribution maps are usually constructed based on recorded occurrences of individuals within a given area. This often involves plotting species locality data onto a given map and creating polygons that encompass all occurrence points (for example, see Branch 1998). Each polygon is referred to as the extent of occurrence for that particular species (Sardo-Palamera et al. 2012). In cases where occurrence data are limited, this may result in maps which do not adequately represent a species true distributional range. Furthermore, species occurrence data are often biased. The presence of geographical, spatial, and taxonomic bias within occurrence data is a persisting issue (Newbold 2010), particularly for species with low detection rates such as reptiles where occurrences are infrequently recorded (Durso et al. 2011). Available data are often limited or biased and this can result in distribution maps which do not reflect reality, thereby hindering their effectiveness (Botts et al. 2011).

An alternative means of estimating a species distribution is that of plotting its area of occupancy. The area of occupancy differs from the extent of occurrences in that it employs the use of grid cells rather than polygons to map out the presence of a species (Burgman and Fox 2002). The resolution of grid cells will affect the accuracy of the estimated distribution, with broader resolutions predicting larger areas of occupancy than fine scale resolutions (Barbosa et al. 2010; Sardo-Palamera et al. 2012). Generally, distributions based on areas of occupancy contain fewer commission errors than those based extents of occurrence (Gaston and He 2011) but contain higher omission rates as a result of incomplete sampling.

To create high-quality, fine-scale distribution maps of reptile species occurring in KNP, detailed species occurrence information would be required, but these data are sparse, largely due to the general difficulties associated with studying reptiles. For example, many reptiles have low detection rates and are thus not often observed within their natural habitats (Durso et al. 2011), but moreover, a general lack of reptile focused research within the recent past has directly contributed to the persistence of the limited availability of such data (Bohm et al. 2013; Tolley et al. 2016). Available distribution data of reptiles within KNP are therefore either potentially outdated (Pienaar 1978) or spatially too broad to be useful within a localized management context (Branch 1998; Bates et al. 2014).

Available occurrence data could be used to construct distribution maps for reptiles within KNP, but these data may be biased. In this chapter I therefore aim to a) collate and synthesize available KNP reptile occurrence data from museum and literature sources, b) critically assess geographical, spatial, and taxonomic biases within these data, and c) discuss the potential effects of these biases in the application of these data for use within a species distribution modelling framework. I further aim to assess the degree to which biases in sampling represent the full environmental niche space of the park. This will aid in identifying knowledge gaps within KNP.

2.2 Methods

2.2.1 Species occurrence data:

I compiled a database of records of reptile occurrences in KNP. To build this database, I collated locality data of reptile species from published literature (Pienaar 1978; Branch 1998; Bates et al. 2014), museum records (Ditsong National Museum of Natural History), field data (obtained from Organization for Tropical Studies), and a virtual museum platform (Reptile map project; http://vmus.adu.org.za). Additionally, I provided novel records from my own sampling (explained in section 2.2.2). All reptile occurrences consisted of presence-only data which I georeferenced as accurately as I was able based on provided locality descriptions.

Some records did not include sufficient locality information, thereby casting uncertainties on their accuracies. I omitted all records which did not contain GPS co-ordinates and had dubious locality descriptions. Additionally, I found several duplicates of records across the various data sources. For example, occurrences reported within Pienaar (1978) were also included within museum databases. In these instances, I filtered my database to only include the original occurrence record and eliminated repetition of records. Overall, this process resulted in my database consisting of 7118 geospatially unique occurrences representing 127 reptile species. To prevent ambiguity within species classifications, I taxonomically updated each record to align with those presented in Bates et al. (2014). I summarized my database to identify taxonomic biases in sampling. To test the hypothesis that representation was not even among families, I counted the numbers of records and numbers of species per family and performed a linear regression analysis to test if a relationship was present between those.

2.2.2 Field work:

For ten days in April 2017, I searched for reptiles within KNP. Whilst limited, and far from a complete survey of reptiles due to the short sampling period, this survey allowed me to

supplement available reptile occurrence data with a minor amount of novel records. Surveying consisted of incidentally searching for individual reptiles within publically accessible infrastructural areas such as campsites and along the major roads of KNP. I recorded the GPS co-ordinates of all reptiles that I observed, identified to the species level (using field guides and expert opinion). Additionally, I assisted with biodiversity surveys carried out by the Organization for Tropical Studies at three publically inaccessible sites within the park (Site 1: 22° 43" 35' S, 31° 22" 40' E; Site 2: 22° 42" 31' S, 31° 02" 57' E; Site 3: 22° 40" 0' S, 30° 59" 03' E), where I collected additional reptile occurrence data.

2.2.3 Spatial arrangement analyses:

To assess and identify sampling bias within reptile occurrence data, I divided KNP into equal sized grid cells of 1 km x 1 km (30 arc seconds) each. To determine the proportion of KNP for which reptile occurrence data exists and quantify the extent of unsampled areas at this resolution, I plotted all reptile occurrences from my database of occurrence records into those 1 km x 1 km grid cells within GIS software (QGIS version 3.2.3). I then counted the number of grid cells which contained at least one reptile occurrence record via the 'points per polygon' tool and used those numbers to create a heatmap of reptile occurrences across KNP.

To identify spatial patterns of sampling bias within grid cells where occurrence records were present, I tested the hypothesis that the numbers of reptile occurrences were unevenly distributed across those cells. To do this, I counted the frequencies of the numbers of reptile occurrence records per grid cell. Furthermore, to test the effects of spatial scale on these patterns, I resampled the data at the following resolutions: 2 km x 2 km, 4 km x 4 km, and 9 km x 9 km (pentad scale). These resolutions allowed me to better identify patterns of sampling bias within KNP and operating at the pentad scale also provided a means of

comparison to previous studies (For example, South African bird atlas project 2; SABAP2; Underhill and Brooks 2016).

To identify if geographical bias in sampling of reptile occurrences was present, I tested the hypothesis that reptile occurrences were biased towards infrastructure. To do this, I performed a linear regression analysis to test if a relationship exists between the number of recorded reptile occurrences within each grid cell, and the proximity of each grid cell to the infrastructure network of KNP. The infrastructure network consisted of any grid cell which contained a campsite, picnic spot, or road. For this analysis I assigned the log (distance to infrastructure per grid cell) as the independent variable, and the log (number of recorded occurrences per grid cell) as the dependent variable, and I used a significance level of 5 % to test if there was an increase or decrease in reptile occurrences as the distance to infrastructure increased. I carried out all analyses using SPSS statistics software version 23.

2.2.4 Climate and environmental data:

I obtained several climatic, environmental, and infrastructural data layers from a variety of sources to act as predictor variables regarding the environmental variability of KNP. In total, I produced 27 individual layers (Appendix 1), each at a resolution of 30 arc seconds (1 km x 1 km). Of these, I obtained the 20 climatic variables relating to temperature, precipitation, and altitude from the Worldclim database (Hijmans et al. 2005; http://www.worldclim.org). Using those I generated 'slope' and 'aspect' layers based on the 'altitude' layer via the 'spatial analyst toolbox' within ArcGIS software version 10.4. I obtained environmental layers in the form of a 'vegetation type' layer from the South African National Biodiversity Institute (bgis.sanbi.org) and a 'soil type' layer from the Soil and Terrain database (SOTER; www.isric.org). Finally, I acquired additional data layers from SANParks relating to the infrastructure and water networks of KNP. Using those layers, I then produced a 'distance to

water' (km) layer using 'Euclidean distance' tool within ArcGIS, with water including all rivers, dams, and lakes within KNP.

I chose these 27 layers as each of these could play a role in affecting reptile distributions across KNP based on the ecology of each species in question. For example, reptiles are ectotherms that require external temperature sources to regulate their body temperature (Branch 1998) and because these requirements differ per species (Brown et al. 2014), I therefore chose to include multiple temperature and precipitation variables of differing time periods and intensities. Similarly, because there is such a wide variety of reptiles within KNP (Pienaar 1976), including terrestrial, arboreal, fossorial, and aquatic species (Bates et al. 2014), I included variables to represent the habitats of these animals. These layers included the presence of and distance to water bodies, human infrastructure, vegetation type, soil type, altitude, slope, and aspect.

2.2.5 Summarizing climate and environmental data:

Climate and environmental layers, particularly those obtained from Worldclim, are usually highly correlated with one another (Demsar et al. 2013). To reduce the effects of correlation between layers, I performed a principal component analysis to summarize these layers into a number of new, uncorrelated representatives of overall environmental variability of KNP. I performed this within R software version 3.4 via the 'rasterPCA' function of the 'rstoolbox' package (Leutner and Horning 2016). This analysis produced 27 new, uncorrelated principal components that summarized the variability of the original 27 layers. Following this, I compared the proportions of environmental variance each principal component contributed towards overall environmental variability. Based on Jackson (1993), I used a cumulative proportion of variance of 85 % as a stopping rule. Once components cumulatively reached

this percentage, I deemed those as sufficient in representing overall environmental variability of KNP and omitted the remaining components from further analyses.

The results of my analysis (explained below) showed that reptile occurrences were biased towards infrastructure. I therefore tested the hypothesis that the environmental range of the infrastructure network of KNP did not represent overall environmental variability across the entirety of the park. To do this, I separated each principal component layer into two new layers, one representing infrastructure, and the other non-infrastructure. Infrastructure layers contained only those grid cells associated with the infrastructure network of KNP, and non-infrastructure layers contained all grid cells except those. I then determined kernel density estimates for these separated layers using the 'density' function within the 'raster' package (Hijmans and van Etten 2016) in R software and compared the distributions of environmental variability of these via two sample Kolmogorov-Smirnov tests for each of the six principal components.

2.3 Results

2.3.1 Summary of recorded reptile occurrences:

The 7118 reptile occurrences within my database varied taxonomically (Table 2.1). The majority of occurrences records belonged to 61 lizard species (with a combined 3434 recorded occurrences, 48 % of all records), and 59 snake species (with a combined 2944 records, 41 % of all records). The remaining records belonged to six species of chelonians and one species of crocodilian, with these making up a low proportion of total records (8 % and 3 % respectively).

Table 2.1: Summary of records of reptile occurrences within Kruger National Park.

Group	No. of species	No. of records	Percentage of total records (%)
Lizards	**61**	**3434**	**48**
Agamidae	3	214	3
Amphisbaenidae	8	195	3
Chamaeleonidae	1	154	2
Cordylidae	9	206	3
Gekkonidae	14	708	10
Gerrhosauridae	4	344	5
Lacertidae	6	272	4
Scincidae	14	1096	15
Varanidae	2	245	3
Snakes	**59**	**2944**	**41**
Colubridae	10	621	9
Elapidae	7	403	6
Lamprophiidae	29	1233	17
Leptotyphlopidae	5	221	3
Pythonidae	1	125	2
Typhlopidae	3	130	2
Viperidae	4	211	3
Chelonians	**6**	**530**	**8**
Pelomedusidae	3	230	3
Testudinidae	3	300	4
Crocodylians	**1**	**210**	**3**
Crocodylidae	1	210	3
Σ	**127**	**7118**	**100**

Records were unevenly distributed amongst reptile families. I found a significant relationship between the number of species and the number of records per reptile family (Linear regression analysis: $F_{1, 17} = 27.71$, $P < 0.01$, $r = 0.78$; Figure 2.1), with the trend being that families with more species had more records available. As such, representation amongst families was not equal with the Lamprophiidae family (29 species) having had the most recorded occurrences and Pythonidae (one species) the fewest. Whilst expected for these families, others had greater numbers of records than what would be expected based on their species diversity assuming all species occur throughout the park (Figure 2.2). This suggested the presence of taxonomic bias at the family level for reptile sampling within KNP.

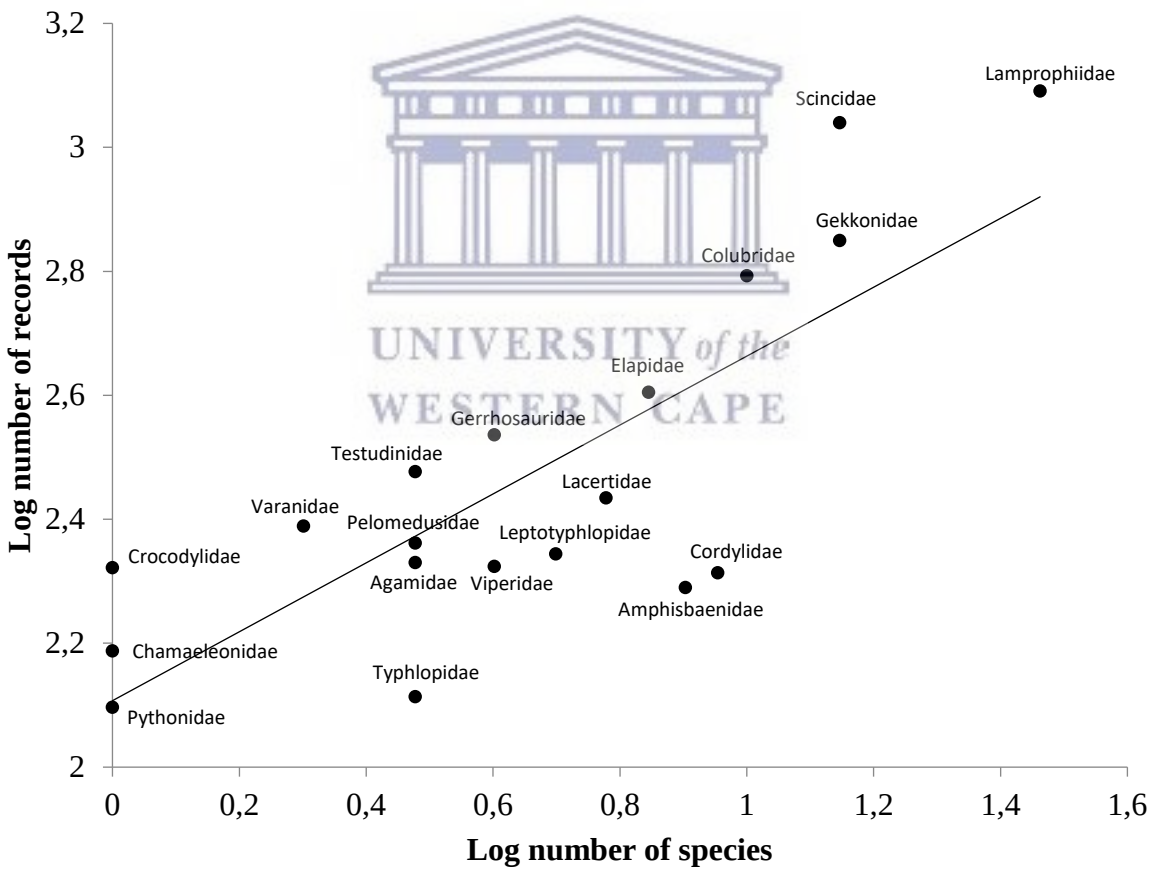


Figure 2.1: Log numbers of occurrences versus log numbers of species per reptile family.

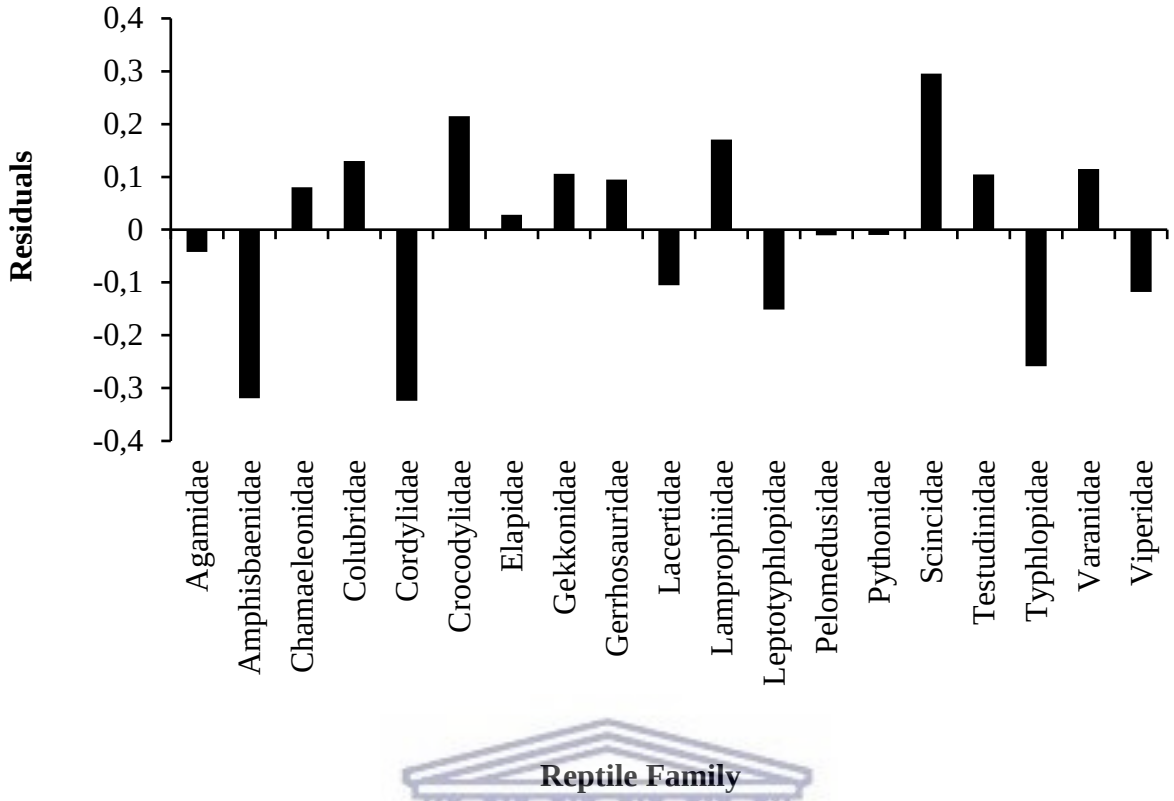


Figure 2.2: Residuals of numbers of occurrences versus numbers of species per reptile family.

2.3.2 Spatial arrangement of occurrence data:

When dividing KNP into separate 1 km x 1 km grid cells, I produced a total of 21761 grid cells (n = 21761). In plotting my dataset of reptile occurrences within these, I found that only 1751 of these grid cells contained at least one occurrence record. These grid cells made up approximately 8 % of KNP, with the remaining 92 % (20010 cells) representing a noticeably severe gap of data deficient areas (Figure 2.3). As the majority of the park contained no data, this suggested that there have been strong sampling biases within KNP in terms of sampling locations. In addition, of the 1751 grid cells containing data, the majority of these had relatively few recorded reptile occurrences (< 10 records). Additionally, the frequencies of reptile occurrence records were not evenly distributed across these cells, with the number of occurrence records per cell ranging from 1 – 118 (Figure 2.4).

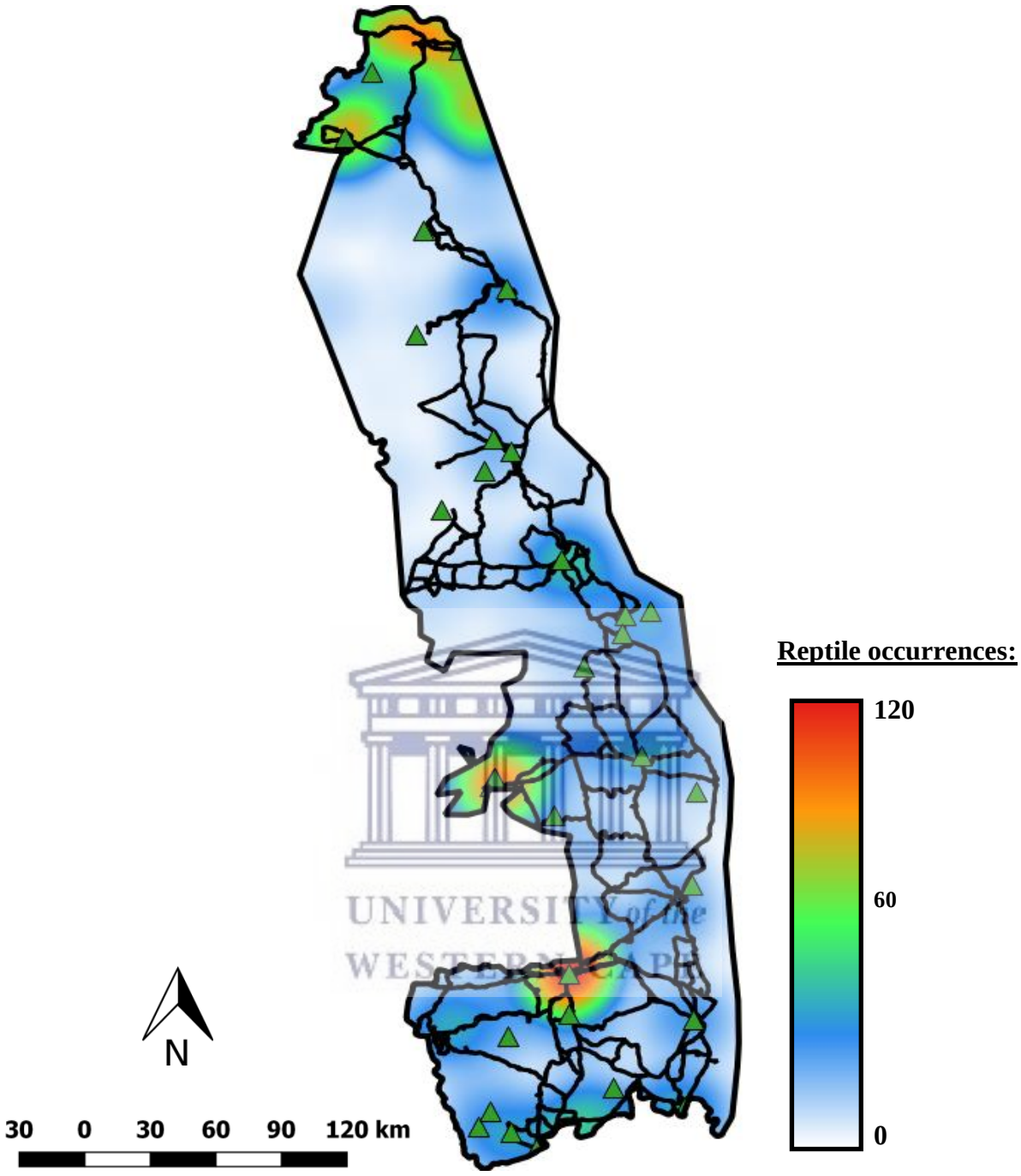


Figure 2.3: Heatmap of reptile occurrences within KNP. Green triangles represent major campsites within the park.

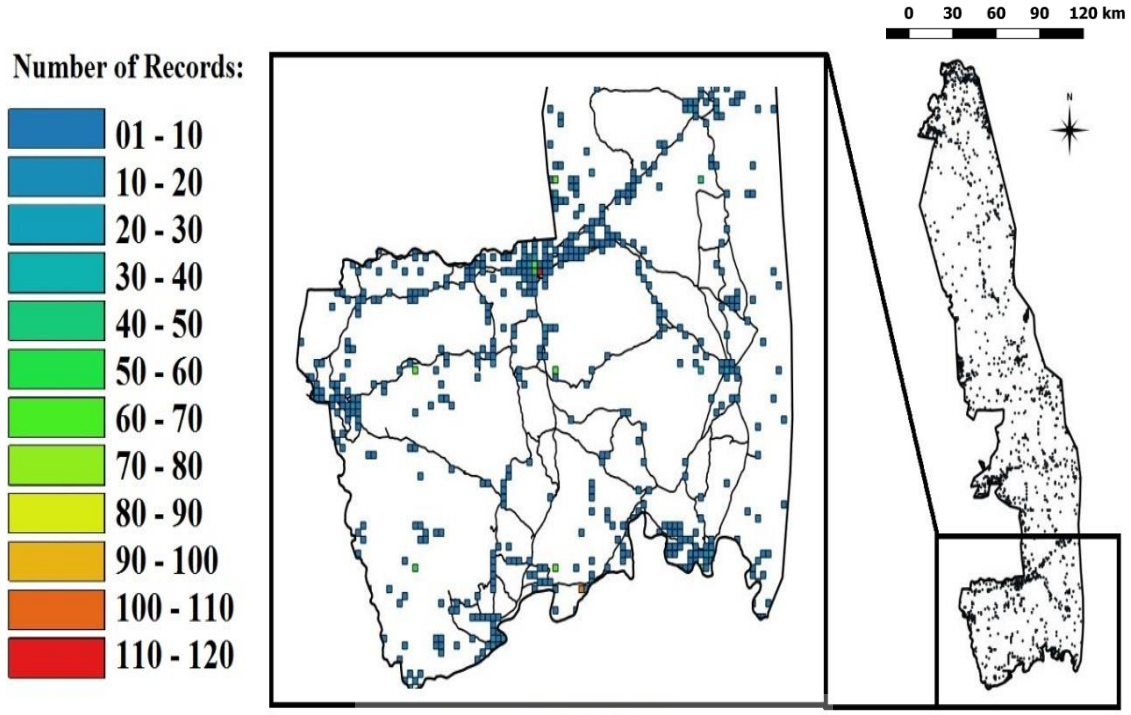


Figure 2.4: Numbers of recorded reptile occurrences per 1 km x 1 km grid cells within KNP (n = 21761), with emphasis on the southern region of the park. White areas represent grid cells containing no data.

At a scale of 1 km x 1 km most grid cells contained few reptile occurrence records. I observed a pattern which showed that as the number of records per cells increased, the number of those cells decreased (Linear regression analysis: $F_{1, 114} = 9.34$, $P < 0.01$, $r = 0.08$). I found that the majority of grid cells contained only a single record (911 cells, 52 % of cells; Figure 2.5a). This pattern held true at resolutions of 2 km x 2 km and 4 km x 4 km respectively (Figure 2.5b and 2.5c) but was not present at the pentad scale (9 km x 9 km; Figure 2.5d). Records at the pentad scale were relatively evenly spread across the different frequencies of grid cells. This suggests that spatial sampling bias is more apparent at fine scale resolutions than at broader levels. This was likely due to factors such as human presence being more impactful at finer spatial scales as these scales offer a more precise summation of reptile occurrences at specific sites.

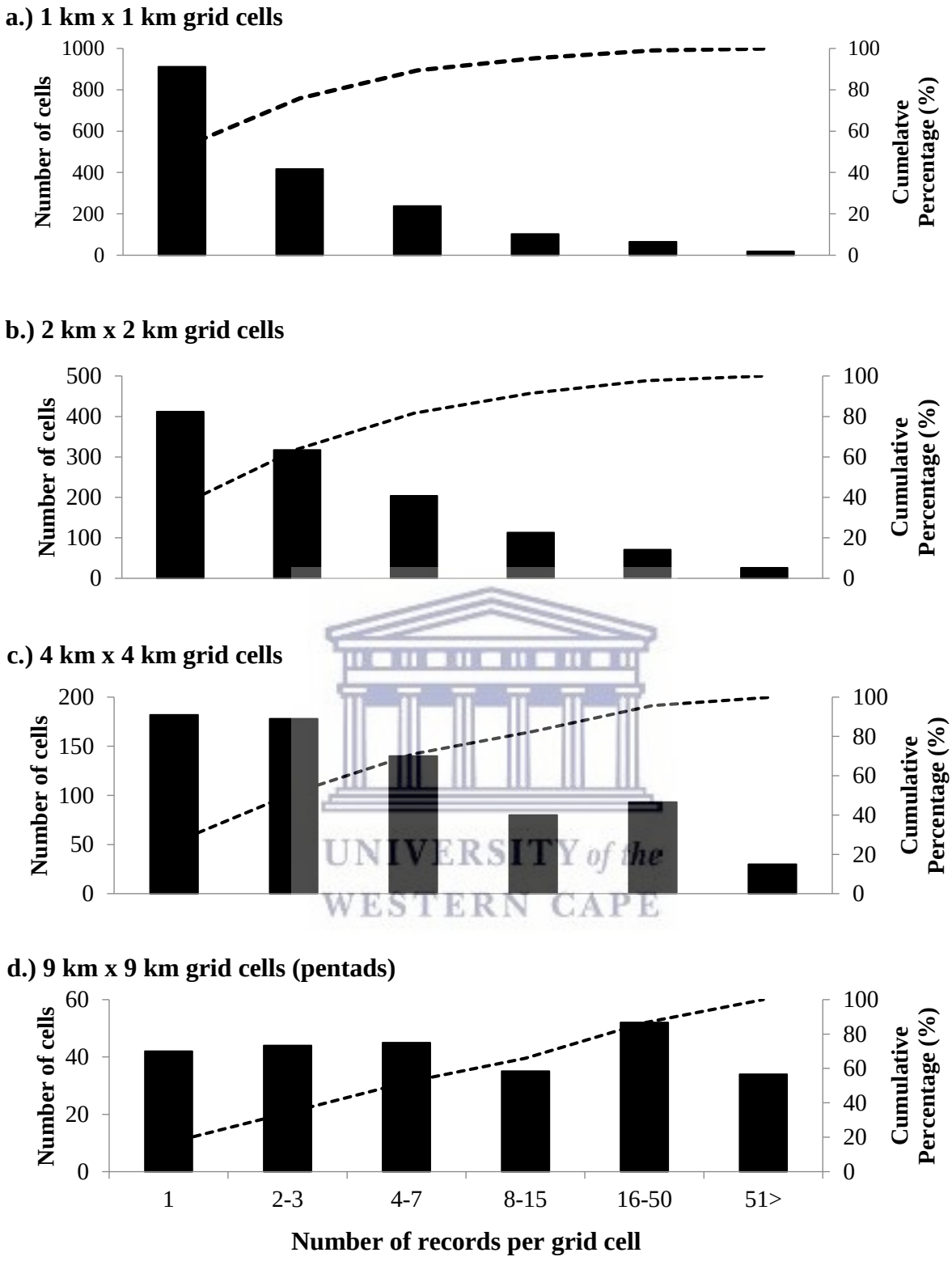


Figure 2.5: Distributions of the numbers of records of reptile occurrences within different sized grids cells within Kruger National Park: a.) 1 x 1 km, b.) 2 x 2 km, c.) 4 x 4 km, and d.) 9 x 9 km (pentad) grid cells.

The infrastructure network of KNP was not evenly distributed across the park. Of the 21761 grids cells making up KNP, 3048 (14 %) of these contains some form of human infrastructure. Based on a linear regression analysis I found that a relationship exists between the number of records of reptile occurrences per grid cell, and the proximity of each cell to the human infrastructure network ($F_{1, 1749} = 6.75$, $P < 0.01$; Figure 2.6), thereby confirming the hypothesis that occurrences were biased towards infrastructure. This relationship was relatively weak ($r = 0.07$), likely due to the fact that the majority of grid cells containing data possessed only a single recorded reptile occurrence, but the overall trend was clear. As distance to infrastructure increased, the number of recorded reptile occurrences per grid cell decreased, suggesting the presence of geographical sampling bias towards infrastructure. Additionally, I found that 28 % of all grid cells containing infrastructure possessed at least one recorded reptile occurrence, whereas only 4 % of grid cells not associated with infrastructure contained data, thus further providing evidence of this bias.

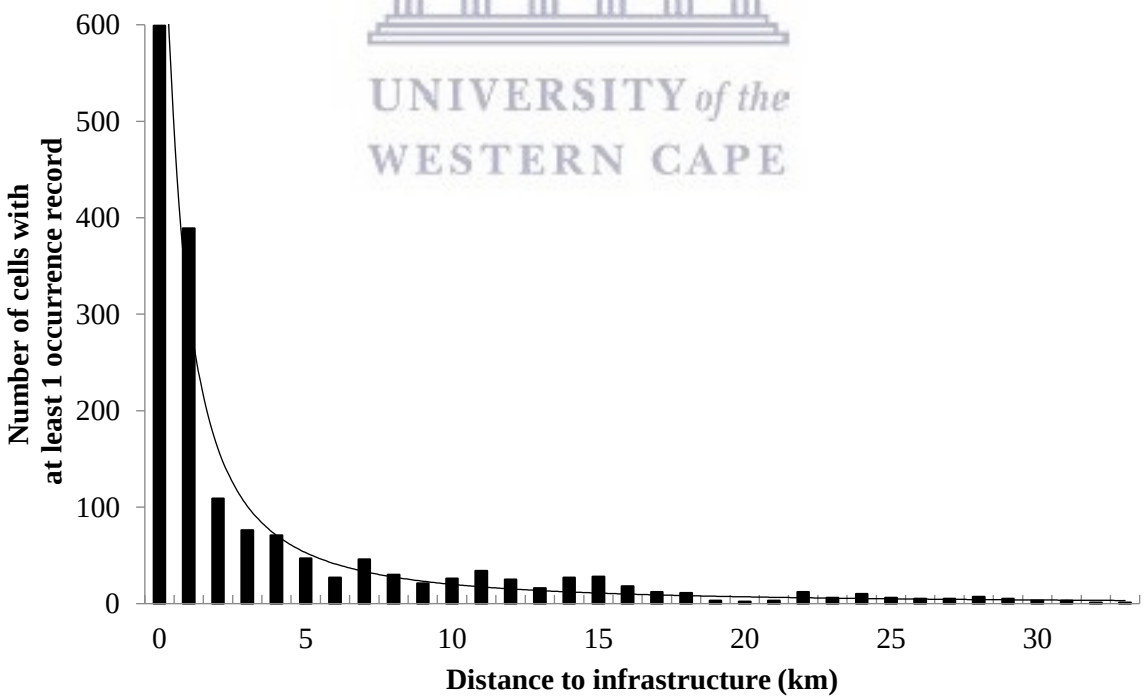


Figure 2.6: Distances of grid cells with reptile occurrences from grid cells containing human infrastructure.

2.3.3 Summary of climate and environmental data:

Using a principal component analysis, I produced 27 new, uncorrelated environmental predictor variables in the form of principal components, each consisting of different proportions of the original 27 variables (Appendix 2). Each of these principal components contributed independently towards overall environmental variability across KNP, and the first six of these cumulatively represented approximately 85 % of total environmental variability (Table 2.2). The remaining 21 components cumulatively represented only 15 % of environmental variability and I omitted these from further analyses.

Table 2.2: Contributions of variance per component for the first six components produced via principal component analysis, cumulatively representing over 85 % of environmental variance within Kruger National Park.

Principal component	**Proportion of variance**	**Cumulative proportion of variance**	**Standard deviation**	**Largest contributing variable**
Component 1	0.46	0.46	3.58	Annual mean temperature
Component 2	0.20	0.66	2.37	Isothermality
Component 3	0.07	0.72	1.35	Altitude
Component 4	0.05	0.77	1.17	Water: presence/absence
Component 5	0.04	0.81	1.07	Infrastructure: presence/absence
Component 6	0.04	0.85	1.02	Aspect

Within the six principal components included in my analyses, despite high percentages of overlap, I found evidence of geographical bias in the distributions of their climatic and environmental variability between the infrastructure network and the rest of the KNP (Figure 2.7). The results of Kolmogorov-Smirnoff tests showed that there were significant differences

in climatic variation between grid cells containing infrastructure and all grid cells excluding those within all six principal components ($P < 0.01$ in all cases; Table 2.3). This showed that heavily sampled areas within KNP (i.e. those in close proximity to infrastructure) were not representative of the entire park in terms of overall environmental and climatic variability. Overall, this suggests that the climatic and environmental conditions of KNP are biased against the human infrastructure network.

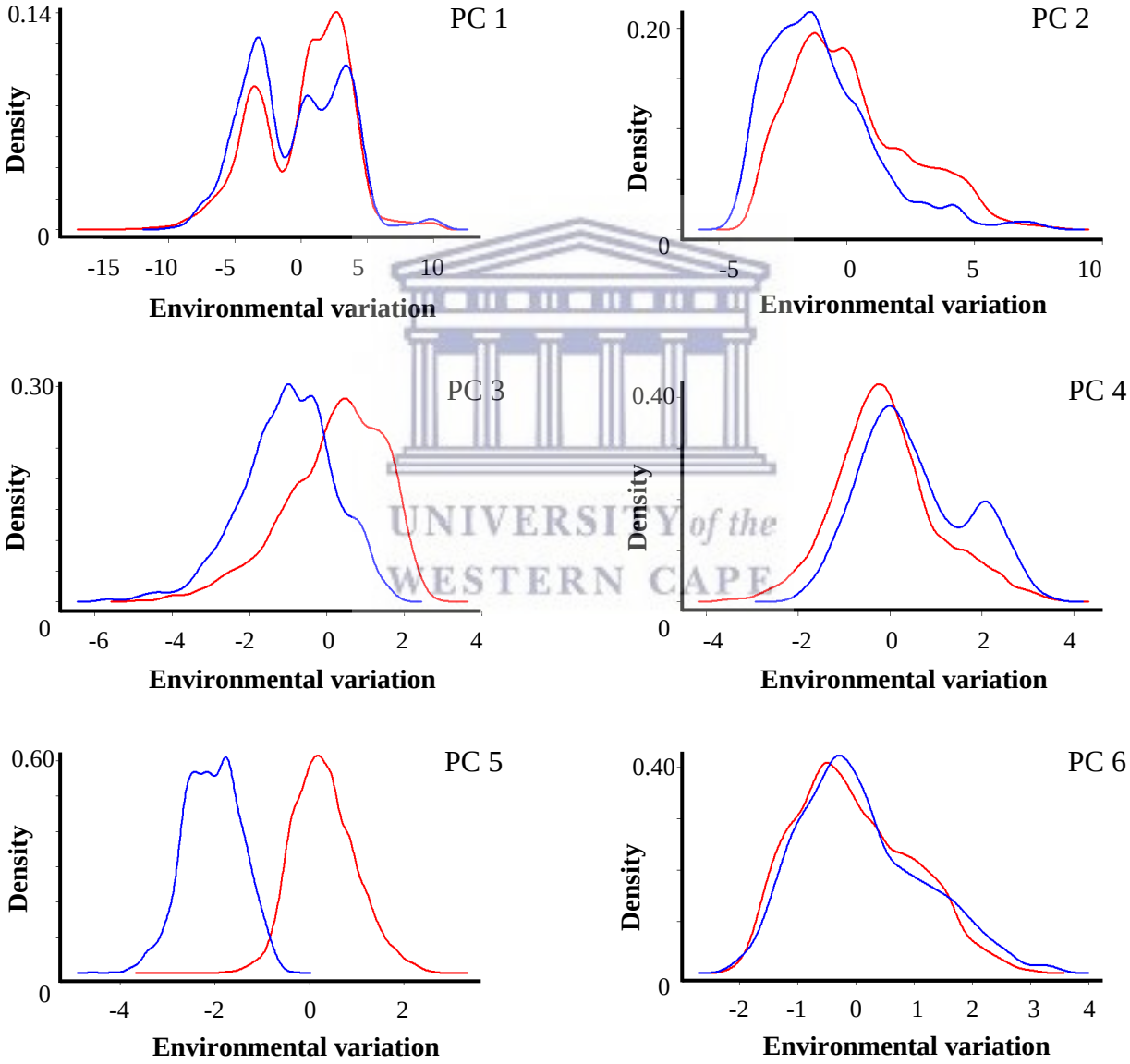


Figure 2.7: Kernel density plots showing differences in environmental variability within six principal components, comparing infrastructure (blue) and non-infrastructure (red).

Table 2.3: Geographical environmental bias of six principal components between infrastructural and non-infrastructural areas within KNP.

Principal component	Overlap percentage (%)	P value	D value
Component 1	87	< 0.01	0.22
Component 2	88	< 0.01	0.16
Component 3	62	< 0.01	0.47
Component 4	81	< 0.01	0.18
Component 5	24	< 0.01	0.96
Component 6	94	< 0.01	0.05

2.4 Discussion

My results showed that reptile occurrence data for KNP were taxonomically, spatially, and geographically biased. Reptile families with high detectability (for example: Scincidae) had notably higher representation within my dataset than those with typically low detection rates. As a result, some families had more records than expected whereas others were poorly represented due to overall data deficiency across the park. The majority of occurrence data were strongly linked towards the infrastructure network, which itself did not represent the overall environmental space of KNP. With approximately 68 % of all records having been in close proximity (< 2 km) to publically accessible areas, the severe bias in sampling associated with infrastructure meant that regions of the park comprised of unique environmental space were not represented in the majority of currently available data.

Bias in species occurrence data is a well-documented constraint of sampling. Human presence and infrastructure typically have a strong correlation with high densities of species occurrence records, particularly within museum collection data (Newbold 2010). The presence of these biases inhibits assessments of true biological distributions of species as infrastructural areas may not adequately represent the biological and climatic conditions that determines a species' range (Kadmon et al. 2004). For example, in assessing occurrence data of frogs within South Africa, Botts et al. (2011) encountered similar issues relating to data bias to those observed here. They concluded that due to sparse sampling efforts in areas away from infrastructure, and the complications that this may have had on determining real biological patterns (Reddy and Davalos 2003), that their database of records was unlikely to have been a true reflection of frog distributions across South Africa. In the same way, due to the biases present here I can conclude that my dataset of reptile occurrences was unlikely to truly reflect real-world distributions of those species across KNP.

Species occurrence records are often associated with human infrastructure, yet human infrastructure is unlikely to be distributed in a biological manner. Species distributions are determined by several factors, including environmental and climatic variation across the landscape (Tolley et al. 2016). Factors such as temperature, altitude, substrate, or vegetation type are likely to be the primary drivers that determine where species occur (Franklin 2010), and these conditions are unlikely to be restricted to human infrastructure. Kadmon et al. (2004) found that roadside bias in occurrence data can result in climatic bias of species distributions. In their study, Kadmon et al. (2004) found that the road network of Israel is biased in terms of climatic variation among some variables (such as precipitation) when compared to the rest of the country. This matches what I have observed here as there were significant differences in environmental variation between the infrastructure network and the rest of KNP.

Given the prevalence of biases present and the limitations of availability of occurrence data, producing distribution maps inferred solely from those data (for example, by drawing a polygon around those points on a map) would be ill advised. However, this dataset does lend itself towards species distribution modelling. This would have to be approached with caution however as the biases present will affect the accuracies of produced models if unaccounted for (Radosavljevic and Anderson 2014). Species distribution modelling is typically most effective when using unbiased data and usually requires an abundance of occurrences to be successful (Bahn and McGill 2013). Despite its limitations this dataset could form the basis of a predictive framework for mapping reptile distributions across KNP if the effects of the associated biases can be mitigated.

In order to successfully develop SDMs for reptiles within KNP using this dataset, the biases present will need to be accounted for. Species distribution models are fundamentally reliant on species occurrence data as these data, in conjunction with environmental or climate data,

forms the basis of predictions (Pineda and Lobo 2009). Here, where occurrence data for several species are relatively sparse, meaningful models for some of these may be impossible to produce with what is currently available as a minimum of at least four occurrences are required for most algorithms (Hernandez et al. 2006). Additionally, since the majority of occurrences are closely clustered together near infrastructural areas, the effects of spatial autocorrelation will have to be accounted for as well (Elith et al. 2006) but this can be dealt with by rarefying records to remove the effects of such clustering. By accounting for the biases and limitations present, an SDM approach using this dataset could help to fill in the gaps within data deficient areas of KNP and produce high-quality maps of where each reptile species occurs within the park.

Whilst overall bias may be unavoidable in the absence of greater sampling efforts, it is possible to work around these limitations. Most modern SDM applications offers a variety of options and settings and these will be pivotal in producing maps that reflect on species true distributions. This was explored further in chapter 3, where this prior knowledge of bias plays an important role.

Chapter 3: Modelling the distributions of reptiles in KNP

3.1 Introduction

KNP is an area of high reptile diversity within South Africa, hosting at least 120 different reptile taxa within its borders (Pienaar 1978; Branch 1998; Parr et al. 2009). Because of this, KNP is considered as a reptile diversity hotspot within South Africa albeit with relatively low levels of reptile endemism (Bates et al. 2014). The high reptile diversity present within KNP can be attributed to the wide variety of heterogeneous vegetation and microhabitat structures that are spatially distributed across the park, as well as suitable climatic conditions (Price et al. 2010). It is also influenced by the proximity of KNP to the tropical African biogeographic centre (Jetz et al. 2004). Whilst we have a relatively good understanding of which species occurs within KNP, and we broadly know that some species only occur within certain areas within the park (Bates et al. 2014), little is known about how these animals are spatially distributed across the entire landscape of KNP at fine spatial scales.

Understanding fine-scale reptile distributions across KNP is critically important for monitoring changes to these species' statuses with regards to TPCs but exact distributions of most of these animals are unknown. Although broad-scale attempts at mapping reptile distributions have been made (Branch 1998; Bates et al. 2014), these have limited management applications as broad scale data are not informative for the current monitoring protocols which largely focus on specific sites (Ferreira et al. 2011). Knowledge of fine scale reptile distributional patterns within KNP remains limited, and one method in which this can be quantified is via species distribution modelling.

Species distribution models are capable of producing probable distributions or ranges of a species based on occurrence data and environmental or climatic variables (Ferrier and Guisan 2006). Importantly, the likelihood of a species occurring in a particular grid cell on a map can

be predicted with some measure of statistical certainty on the basis of a range of environmental or climatic predictor variables. As suggested by Ferreira et al. (2011), a predictive framework could be used as an alternative to current monitoring protocols which involves on-site sampling. By using environmental variables to predict the occurrences of reptiles at relatively fine spatial scales this could provide informative data that is not limited by the difficulties associated with on-site sampling such as low detection and other logistical constraints.

Various species distribution modelling approaches are available via several software and GIS programs, with the most commonly used approach being that of ecological niche modelling (Raxworthy et al. 2003; Guisan and Thuiller 2005). Ecological niche modelling allows for the prediction of a species' fundamental niche/distribution within a given map, with the extents of recorded occurrences acting as the realized niche/distribution. Ideally, in conjunction with predictor variables, these models would be based on a combination of species' presence and absence data for a given area. However, since it is usually difficult to obtain reliable absence data (Pineda and Lobo 2009), most modern SDM applications only require presence data to produce hypothesized distributions of species that can be statistically and empirically tested.

One SDM application that has gained in popularity over the last decade is that of Maximum Entropy (MaxEnt; Phillips et al. 2006). MaxEnt requires species presence data, and environmental or climatic variables to model distributions based on machine learning algorithms. MaxEnt has fared well against similar programs (Elith et al. 2006; Hernandez et al. 2006) and is currently the preferred platform for many species' distribution studies (Merow et al. 2013), including studies of reptiles in South Africa (For example: Tolley et al. 2009; Barlow et al. 2013). In comparison to similar programs such as GARP, MaxEnt was found to outperform it in terms of predictive ability (Elith et al. 2006). Similarly, MaxEnt also outperforms more established methods such as BIOCLIM or ENSA. This is largely

because MaxEnt has a high level of flexibility in both its model building procedure and cross validation methods (Elith et al. 2006). For instance, MaxEnt offers various adjustable settings and parameters (for example, adjustable regularization) that allow users to limit the effects of biases in presence data, spatial autocorrelation, and model complexity (Phillips et al. 2006; Phillips and Dudik 2008) that its competitors do not.

Here, I aim to predict the distributions of all reptile species occurring within KNP at a relatively fine spatial scale of 1 km x 1 km using MaxEnt. In addition, I aim to predict reptile species richness across the park and evaluate the success of model predictions using typical model statistical evaluation methods. This could allow for the accurate depiction of reptile distributions across KNP which will be useful for monitoring purposes with regards to TPCS.

3.2 Methods

3.2.1 Species distribution modelling:

I predicted the distributions of 119 reptile species across KNP using MaxEnt version 3.3.4 at a spatial resolution of 1 km x 1 km. These predictions were based on presence-only occurrence data, and six principal component layers which acted as environmental predictor variables. These principal components were summarized layers that represented the overall environmental and climatic variability of KNP based on temperature, precipitation, vegetation, soil type, altitude, distance to water, and infrastructure variables (see chapter 2) and were obtained from Worldclim, SANParks, SANBI, and SOTER.

I tested if reptile occurrences across KNP were randomly clustered by calculating the Morans I value of the dataset within ArcGIS software. The Morans I value is a correlation coefficient that measures the overall similarity of clustering within a dataset (Eviritt and Hothorn 2011). Thereafter, to limit the effects of clustering, and thus spatial autocorrelation, within reptile occurrence records I rarefied my dataset by removing records within 1 km of each other on a species by species basis (i.e. within each 1 km x 1 km grid cell the same species would not be represented more than once). In carrying out the above mentioned process, I aimed to limit the effects of species clustering and sampling bias whilst still preserving broad coverage across KNP. Additionally, I omitted eight species (*Amplorhinus multimaculatus*, *Bitis caudalis*, *Duberria lutrix lutrix*, *Lamprophis guttatus*, *Leptotyphlops scutifrons conjunctus*, *Monopeltis leonhardi*, *Naja melanoleuca*, and *Psammophis trinasalis*) from modelling as I did not have sufficient records of occurrence for all those species (less than four records). My final tally of occurrences used to produce models was 5859 records, representing 82 % of my initial dataset.

I modelled the distributions of each species using the same set of parameters within MaxEnt to account for the similar biases present across all data. This meant that across the range of all produced models, the effects of shared biases were mitigated to the same degree for each species. However, because parameters were not species-specific it is possible that model performances of some species may have been inhibited (Hernandez et al. 2006). Since the numbers of occurrence records varied per species, I ran all models using linear, hinge, product, and quadratic features to account for these differences (Phillips et al. 2006). For all models, I used a random test percentage of 25 % of occurrences to act as a test of model performance within a maximum of 10000 background points. I also applied a regularization multiplier of three for each model to account for model over-fitting (Radosavljevic and Anderson 2014) and minimize model complexity (Galante et al. 2018).

To further account for geographical sampling biases in occurrence data, I included a bias file within the production of each model. Within MaxEnt, the inclusion of a bias file allows for species occurrences to be unequally weighted within the modelling algorithm in order to place greater emphasis on some occurrences depending on the bias selected. Since infrastructural bias was so prevalent in my occurrence data (see chapter 2), I included a bias file that was based on the distance of each occurrence record to grid cells containing infrastructure (created within ArcGIS; Figure 3.1). This allowed for records further away from well-sampled, infrastructural areas to receive a greater weighting than those in close proximity to infrastructure, resulting in the effects of sampling bias having less of an impact on model predictions and depicted distributions (Phillips and Dudik 2008). Finally, I ran all models using a bootstrapping approach, with 100 replicates for each species as recommended by Radosavljevic and Anderson (2014).

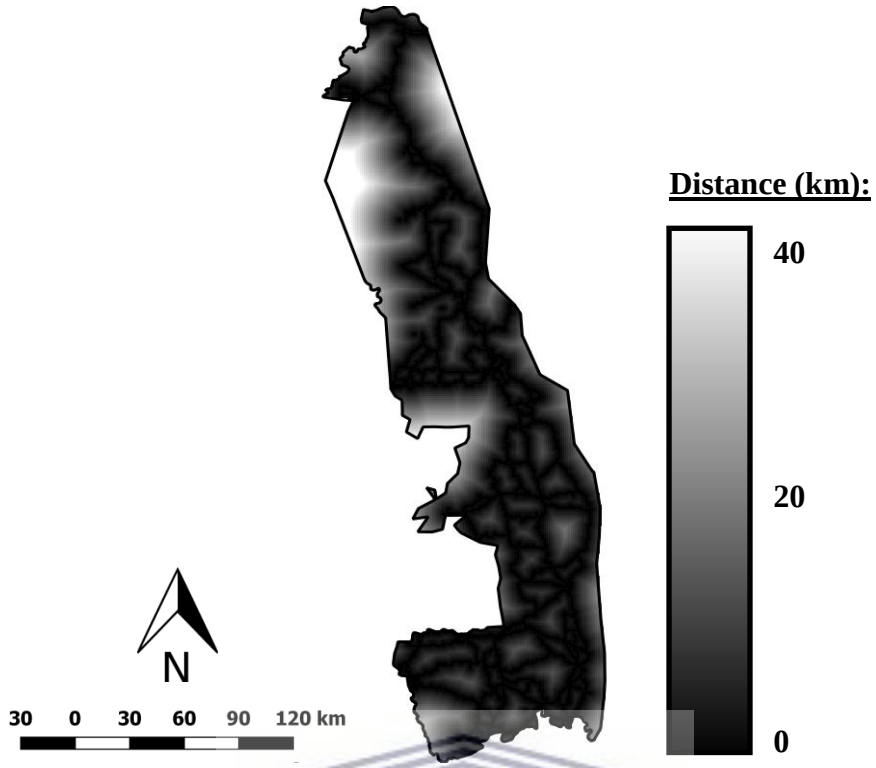


Figure 3.1: Distance of grid cells to infrastructure within KNP, used as a bias file within MaxEnt. Areas in close proximity to roads are black, with those far away being white.

3.2.2 Model selection and evaluation:

I evaluated the predictive strength of all models I produced based on area under the receiver operating curve (AUC) scores obtained for each. AUC scores represent the most widely used evaluation measure for SDM predictions based on presence-only data (Stockwell and Peters 1999; Pearce and Ferrier 2000). These scores offer a discriminative value of a fitted model's predictive performance, ranging from zero to one, with scores closer to one representing greater predictive power (Phillips et al. 2006). MaxEnt automatically provides AUC scores as one of several outputs when producing models, and I used those values as indicators of the performances of each of the models I produced. Since I produced 100 replicate models for each species, I took the average AUC score across replicates to obtain a measure for each of the 119 species. Similarly, to obtain final predictive distributions for each species I used the

average predicted distribution from all of the replicate models to eliminate biases in models in terms of predictive strength.

3.2.3 Model performance analyses:

In order to determine potential factors that influenced the predictive strength of each model at both the species and family levels, I compared species-specific input data to their respective model AUC scores. I tested the hypothesis that variances in the numbers of occurrence records would affect model performance by using a linear regression analysis to assess if a relationship exists between the numbers of occurrences used in model production and the respective AUC scores of the outputs. I also tested the hypothesis that a relationship exists between the extents of the areas surrounding occurrences (determined via the 'convex hull' tool within QGIS software version 3.2.3), and the AUC scores obtained per species. These areas surrounding occurrences represented estimates of the species range based on the locations of each point, analogous to IUCN's extent of occurrences. Lastly, I tested if a statistical relationship exists between the number of occurrences and the extents of areas around these occurrences per species. I repeated these comparisons at the family level, using the averages across species for all three variables.

3.2.4 Environmental variable importance:

To quantify which variables had the largest influence on the models produced for all species, I compared the percentage contributions and permutation importance of each of the six principal components across all taxa. For each model, MaxEnt employs a specific algorithm to obtain a final prediction, but multiple algorithms could result in the same final prediction (Phillips and Dudik 2008). The percentage contribution is a measure of how much each of the six principal components contributed towards the final prediction based on the specific algorithm used for that exact model (Phillips et al. 2006). Conversely, the permutation

importance depends on the final model rather than the algorithm and measures how heavily that model depended on each variable. Both measures offer insight into identifying which components were most important for each model, and I hypothesized that different reptile species and families would be associated with different principal components. To test this hypothesis, I used separate one-way ANOVA tests to determine if differences existed between percentage contributions, and permutation importance of the six principal components across all models. I carried out all analyses using SPSS software version 23.

3.2.5 Estimating species richness:

Models output from MaxEnt are continuous and required a cut-off threshold (a percentage or probability) to be converted into a dichotomous, binary classifications of a species' presence or absence within each grid cell. There is no strict rule for determining an appropriate cut-off threshold, but the selected value should not be arbitrarily chosen (Wilson et al. 2005; Hernandez et al. 2006). As recommended by Phillips et al. (2006) I used the 10^{th} percentile of training data for each model as estimated by MaxEnt as the cut-off threshold for each species. Cut-off thresholds were thus species-specific, and I used these values to convert all models into binary presence/absence maps representing the distributions of all 119 reptile species.

I overlaid all 119 predicted reptile presence/absence layers onto one another within QGIS software version 3.2.3 and used the 'point sampling tool' to extract their attribute information. Attribute information refers to the geographic features of each layer, which here was the presence or absence of each reptile species within each grid cell. The 'point sampling tool' allows for the collection of attribute information from multiple layers simultaneously, which allowed me to estimate the number of reptile species predicted to occur within each grid cell of the park and thus create a predicted species richness map of KNP at a 1 km x 1 km resolution.

3.3 Results

3.3.1 MaxEnt models:

My overall dataset showed evidence of clustering (Morans I = 0.17). Thus, using my rarefied occurrence records and six principal components I produced predicted distribution models for 119 reptile species across KNP. These models displayed the probability of each species' occurrence within each grid cell, ranging from zero to one, with one representing the highest probability and zero representing the lowest (Phillips et al. 2006). Using species-specific 10^{th} percentiles of training presences as cut-off thresholds, I converted all models into presence/absence maps, thereby predicting the ranges of reptile species present within KNP (Appendix 3). The predictive strength of these models varied per species with several models performing relatively well (< 0.90 AUC) and others decidedly poorly (> 0.60 AUC).

3.3.2 Model evaluation:

I evaluated the predictive strength of each model based on their respective AUC scores. I found that the average AUC score across all 119 species was approximately 0.75, indicating relatively good predicative performance that can be considered as informative (Phillips et al. 2006; Phillips and Dudik 2008). However, AUC scores varied per species with some models performing notably better than others (Appendix 4). Models that obtained higher AUC scores were likely to represent truer estimates of distributions for those species than those that performed poorly. For example, it would generally be considered that a model with an AUC score as low as 0.53, as was the case for *Prosymna lineata* (Figure 3.2a), is not much greater than a random prediction (Merow et al. 2013). Conversely, models that achieved higher AUC scores (such as that of *Varanus niloticus* or *Pachydactylus affinis*; Figures 3.2b and 3.2c respectively), could be considered as probable representations of those species' actual distributions within KNP.

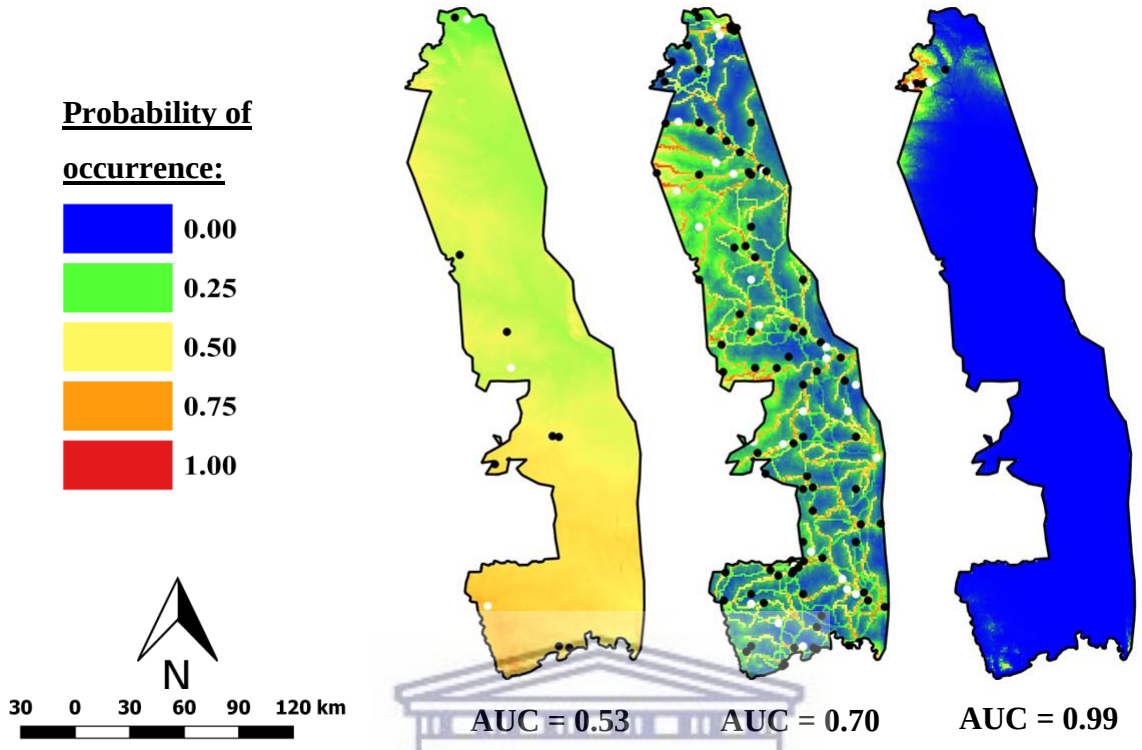


Figure 3.2: Predicted distribution maps of a) *Prosymna lineata*, b) *Varanus niloticus*, and c) *Pachydactylus affinis* representing MaxEnt models that obtained low, median, and high AUC scores respectively. Black dots represent rarefied species occurrences and white dots represent test data. Map colours represent probability of occurrence within each grid cell.

3.3.3 Model performance:

Model performance varied across the range of all taxa. I found that due to a combination of variances in the numbers of occurrence records, and the extents of areas encompassing those, models for each species obtained different AUC scores. I found a significant relationship exists between the AUC scores obtained for each model, and the number of occurrence records used to produce each model within MaxEnt. As the number of occurrence records used to create a model increased, the AUC score of the resulting model decreased (Figure 3.3a; Table 3.1). As such, species whose models performed best thus had very low numbers of occurrences, confirming the hypothesis that the number of occurrences used affected

model performance. However, in terms of biological interpretations, despite achieving high AUC scores it was unlikely that these models were depicting real-world distributions as the few occurrences used in model creation would not represent complete environmental space.

Similarly, I found that there was also a significant relationship between the AUC scores of each model and the extents of areas encompassing occurrence points of each species. I found that species which had smaller areas of occurrences produced models with higher AUC scores than those with relatively large areas (Figure 3.3b; Table 3.1). This suggested that rather than the number of occurrences used for model creation, the spread of these occurrences across the KNP was more influential in the strength of the resulting models.

Table 3.1: Results of four separate linear regression analyses comparing model performance across 119 reptile species.

Comparison	**F value**	**P value**	**df**	**R^2**
Log AUC score vs. log number of occurrences	89.19	< 0.001	1, 117	0.43
Log AUC score vs. log area of occurrences	201.37	< 0.001	1, 117	0.63
Log AUC score vs. log density of occurrences	85.70	< 0.001	1, 117	0.42

I also found that species that had low numbers of occurrences tended to cover smaller areas geographically across KNP and so densities of occurrences were high. Often, when occurrences were few, I found that they clustered closely together geographically, and this was typical for species with small distributional ranges within KNP. Conversely, widespread species tended to have higher numbers of occurrences available, and as a result, these covered larger areas across KNP. I found a significant relationship exists between AUC scores obtained for each model and the density of occurrences per area (Figure 3.3c; Table 3.1), suggesting that the spread of occurrences across geographical space rather than the numbers of occurrences used was more important in affecting AUC scores.

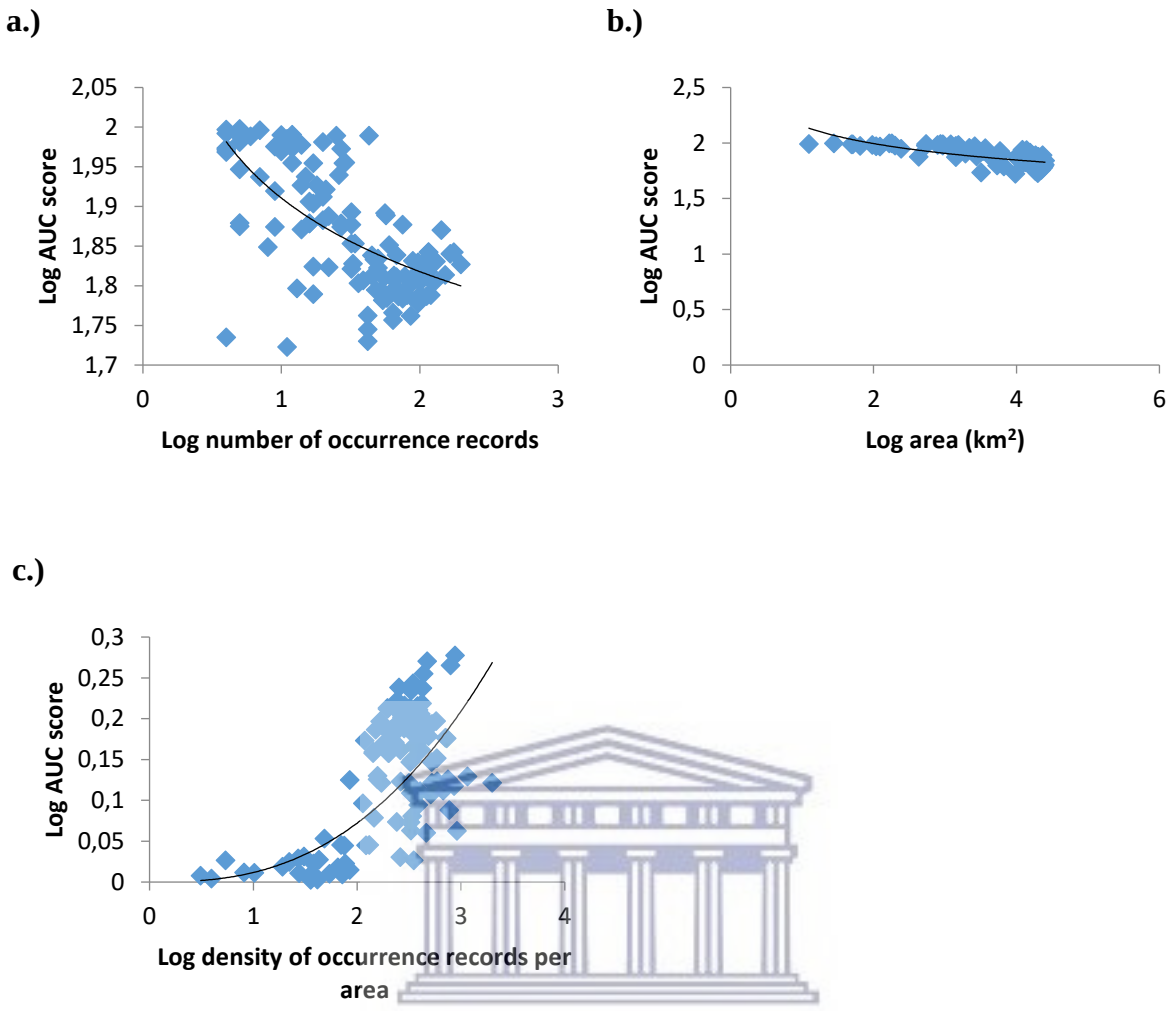


Figure 3.3: Scatter plots comparing input variables and model scores for 119 reptile species.

Further analyses revealed that the average predictive strength of distributional models produced via MaxEnt also varied at the family level (Table 3.2). The Amphisbaenidae family produced the strongest models, achieving an average AUC score of 0.88, across an average of 23 occurrence records per species (158 occurrences in total). The weakest performing family, the Viperidae, achieved an average AUC score of 0.61 with an average of 58 occurrence records per species. Differences in model performance across reptile families were clearly apparent.

Table 3.2: Summary of average model performance of reptile families within KNP.

Family	Number of species	Average number of occurrence records	Total number of occurrence records	Average test AUC
Agamidae	3	51	152	0.74
Amphisbaenidae	7	23	158	0.88
Chamaeleonidae	1	125	125	0.64
Colubridae	10	55	545	0.70
Cordylidae	9	19	170	0.84
Crocodylidae	1	143	143	0.74
Elapidae	6	57	339	0.70
Gekkonidae	14	40	558	0.85
Gerrhosauridae	4	73	291	0.64
Lacertidae	6	39	234	0.73
Lamprophiidae	25	42	1048	0.70
Leptotyphlopidae	4	49	197	0.71
Pelomedusidae	3	68	203	0.64
Pythonidae	1	112	112	0.68
Scincidae	14	58	818	0.78
Testudinidae	3	87	261	0.77
Typhlopidae	3	39	116	0.70
Varanidae	2	99	197	0.65
Viperidae	3	58	174	0.61

Similar to the comparisons made at species level, I found that families with higher numbers of occurrence records produced models with weaker predictive strength than models with relatively few occurrence points. The results of a linear regression analysis showed that a significant relationship existed between the average numbers of occurrences and average AUC scores across reptile families (Figure 3.4a; Table 3.3). I found a significant relationship between AUC values and the numbers of occurrence records at the family level.

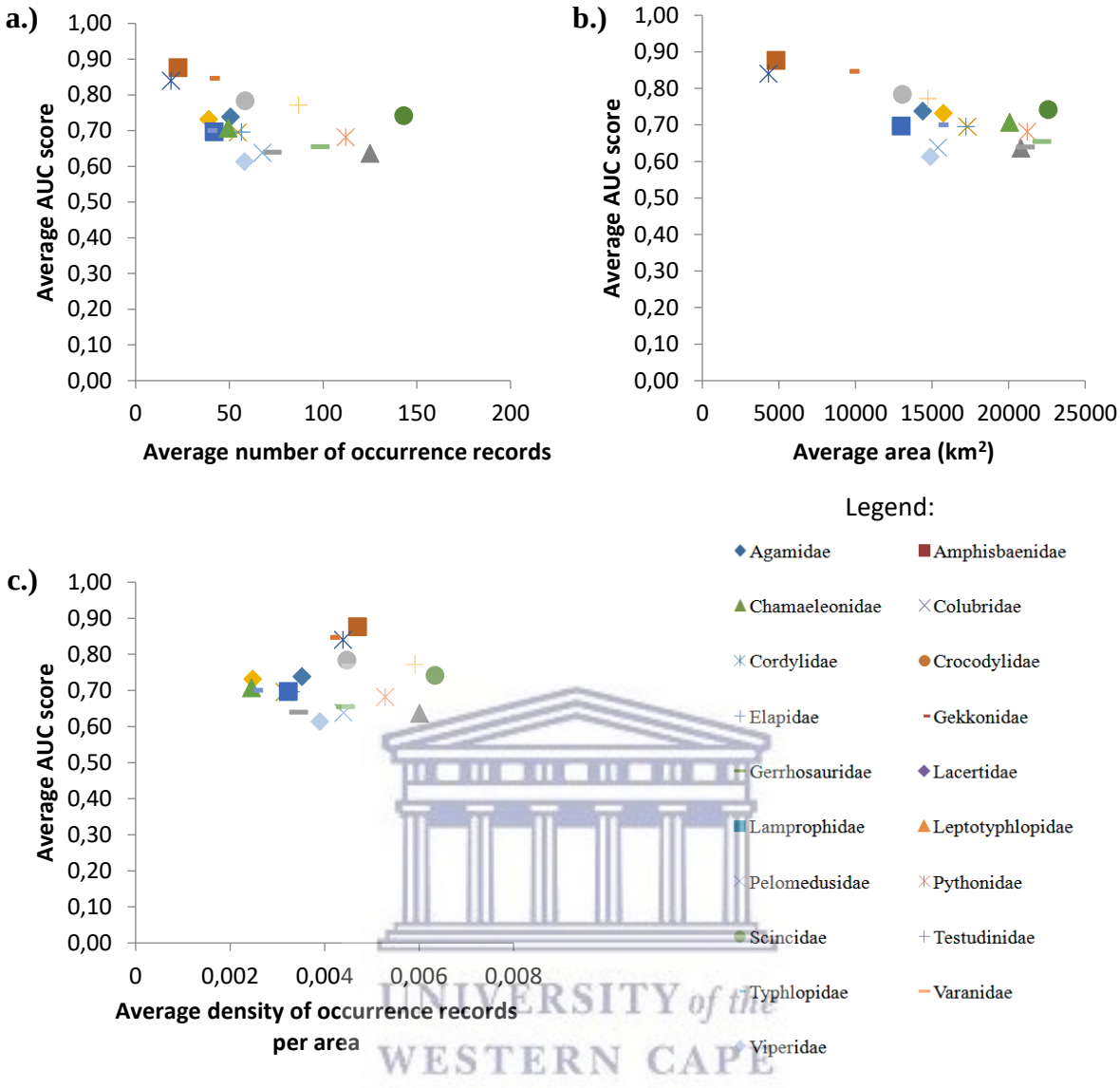


Figure 3.4: Scatter plots comparing input variables and model scores for 19 reptile families.

I found that the number of occurrences was not the only factor affecting AUC scores of reptile families, with the areas around those occurrences being important contributors. Via linear regression analysis I found that on average the test AUC scores of each reptile family was higher when the average area of occurrences for said family was lower. This was likely due to smaller areas of occurrence having less environmental and climatic variance, thus resulting in less complex and therefore stronger predictive models. This relationship was statistically significant (Figure 3.4b; Table 3.3). This showed that the average area encompassing occurrences also affected the AUC scores obtained at the family level.

However, I found that there was no significant relationship between average AUC scores and densities of records at the family level (Figure 3.4c; Table 3.3). This showed that similar patterns are true for reptile species and families when using input occurrence data, except when assessing densities of records.

Table 3.3: Results of four separate linear regression analyses comparing model performance across 19 reptile families.

Comparison	**F value**	**P value**	**df**	**R^2**
Average AUC score vs. average number of occurrences	5.47	0.03	1, 17	0.24
Average AUC score vs. average area of occurrences	21.91	< 0.001	1, 17	0.56
Average AUC score vs. average density of records	0.285	0.600	1, 17	0.02

3.3.4 Environmental variable importance:

Environmental predictor variables varied across taxa in both their contributions, and importance to models. Across all 119 species, in terms of percentage contribution towards model creation, I found a significant difference amongst the six principal components (one-way ANOVA test: $F_{5, 708} = 55.64$, $P < 0.01$). Similarly, the same differences were present for permutation importance across components (one-way ANOVA test: $F_{5, 708} = 39.51$, $P < 0.01$). On average, I found that the first and second principal components had significantly greater influence towards predicting distributions of reptile species across KNP than the remaining four components (Figure 3.5) and were the primary drivers determining suitable areas within the park for the majority of models.

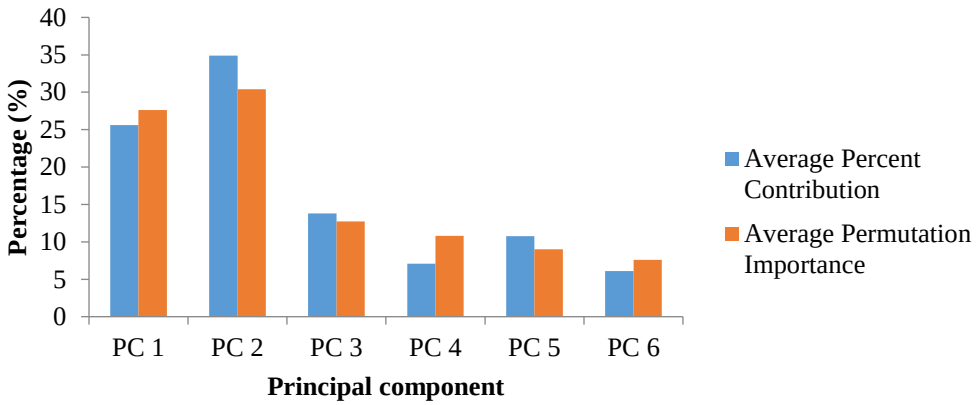


Figure 3.5: Average percentage contributions and permutation importance of six principal summarising overall environmental variability of KNP components towards distribution models of 119 reptile species across KNP.

At the family level, I found the same patterns to be present. I found that there was a significant difference between principal components in terms of both percentage contribution (one-way ANOVA test: $F_{5, 108} = 14.32$, $P < 0.01$), and permutation importance (one-way ANOVA test: $F_{5, 108} = 13.82$, $P < 0.01$) towards models. On average, at the family level I again found that the first and second principal components had the most influence across model production, with the fifth and sixth components contributing least (Figure 3.6).

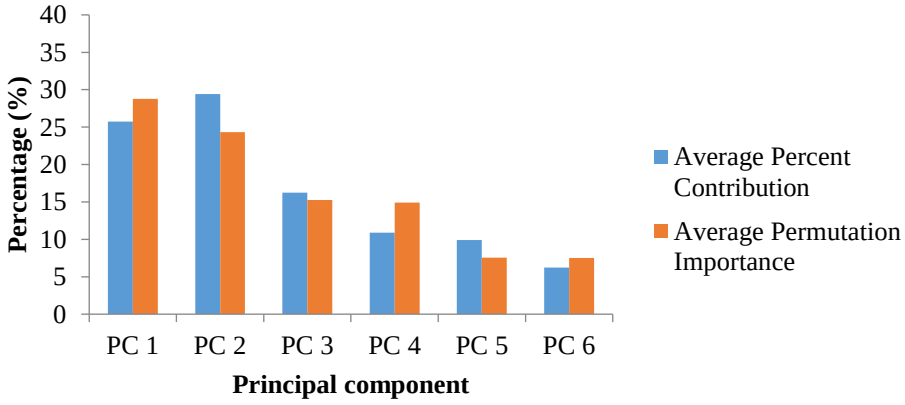


Figure 3.6: Average percentage contributions and permutation importance of six principal components summarising overall environmental variability of KNP towards distribution models of 19 reptile families across KNP.

Whilst on average the first two principal components had the largest impact on all models across families, for individual families this was not always the case. Some models were influenced more by the other four principal components (Figure 3.7), likely due to the unique ecologies of those particular species. For example, I found that the predicted distribution of the Crocodylidae was most influenced by principal components three and four. This made sense biologically as those two components were closely associated with the presence of, and distance to water across KNP (see chapter 2), which are undoubtedly the limiting environmental factors for aquatic species such as *Crocodylus niloticus*. The same can be seen within the Varanidae family, where principal components associated with water again contributed most towards model production. Overall, principal components one and two influenced models most, followed by components three and four, with components five and six having had very little influence across reptile families.

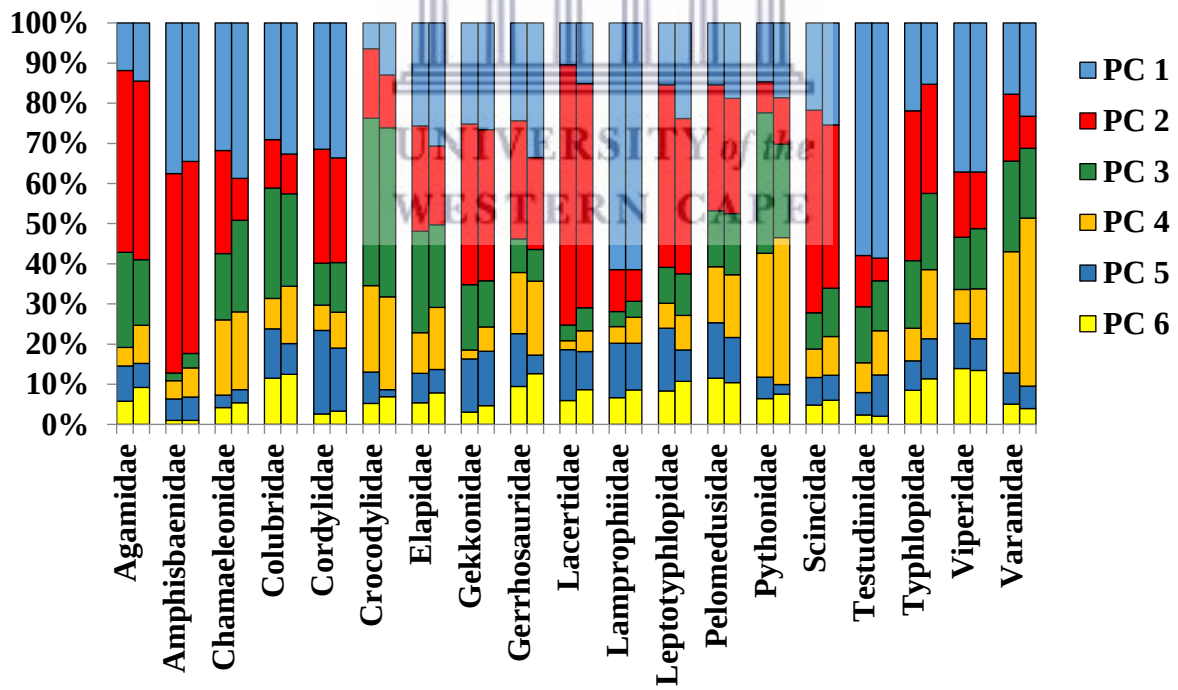


Figure 3.7: 100 percent stacked columns displaying average percentage contributions (left columns), and average permutation importance (right columns) of six principal components towards distribution models of 19 reptile families within KNP.

3.3.5 Predicted species richness:

Based on my modelled distributions of reptiles within KNP, I predicted reptile species richness across the landscape of the park. Species richness was unevenly spread across KNP, with several areas having considerably higher predicted richness than others (Figure 3.8). The northern most regions of the park, particularly along the eastern and western borders near Punda Maria and the Nyandu sandveld regions respectively appeared to have the highest predicted richness within the park. The central most region of KNP, such as those areas dominated by Mopaneveld, were predicted to have the lowest numbers of reptile species present (with the exception of the riverine regions that maintained relatively high species richness throughout KNP). Overall, I predicted that every grid cell within the KNP had a minimum of at least six species occupying it.

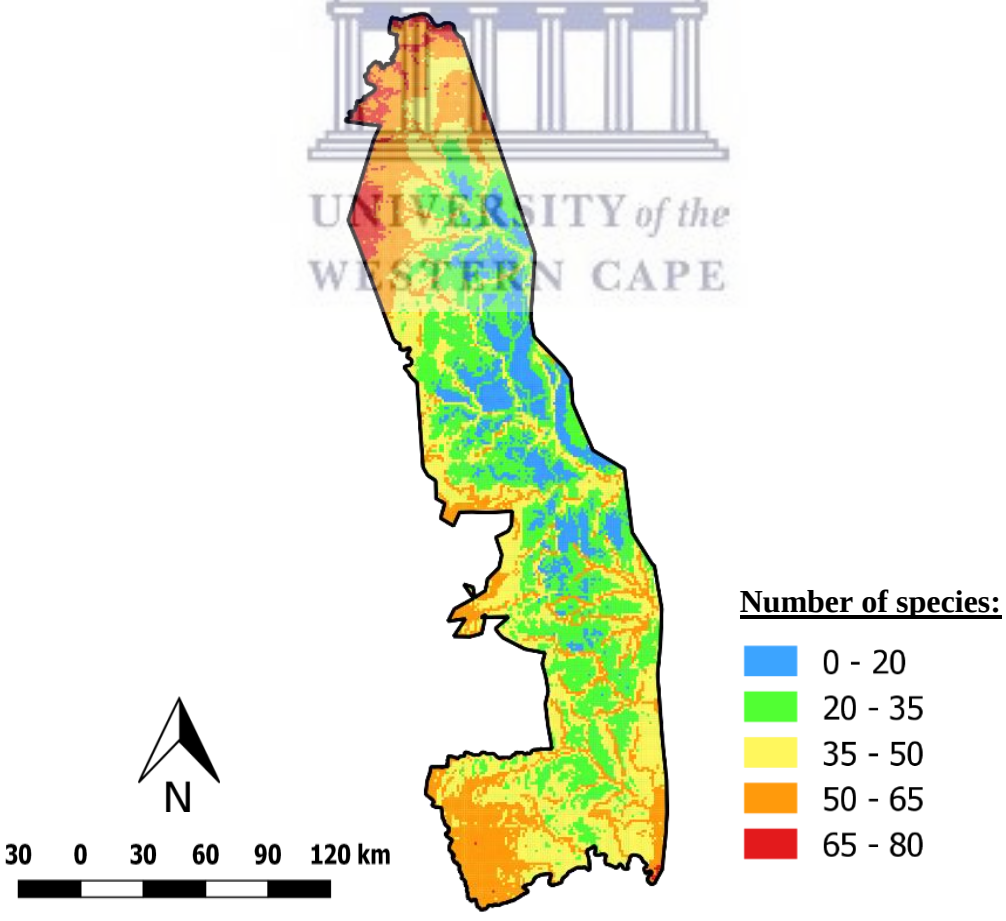


Figure 3.8: Predicted reptile species richness across KNP at a 1 km x 1 km spatial scale.

3.4 Discussion

Understanding species distributions is fundamentally important in assisting with their conservation. Here, I have provided the first fine-scale attempt at predicting distributions for 119 reptile species within KNP. Using an ecological niche modelling approach via MaxEnt machine learning, I predicted distributions for each of these species with varying degrees of model performance, achieving an average AUC score of 0.75 across all models. Instances of poor performance within some models highlighted the flaws within my predictions and most likely occurred as a result of the inherent biases within input occurrence data. As a result, models for some species may have been fatally flawed and were unlikely to accurately represent true distributions of those species across KNP. That said I have learnt that variances in model strength were strongly affected by differences in the numbers and densities of occurrence records used to produce each model, as well as the extents of the areas encompassing those occurrences.

Model performances, in the form of AUC scores, were affected by several factors. Several of the species that I aimed to model distributions for had very few occurrence data available. One of the major drawbacks associated with MaxEnt is that it is sensitive to sample size (Phillips et al. 2006; Anadon et al. 2012). For reptiles, particularly cryptic species, occurrence records are not always readily available (Pearson et al. 2007) and this was indeed the case here. In addition, the majority of occurrence records I obtained were from museum databases. Whilst extremely valuable due to the overall scarcity of reptile occurrence records, museum data are not particularly ideal in the use of SDMs as these collections are highly likely to be inherently biased towards infrastructural and other frequently sampled areas (Newbold 2010). Input occurrence data was thus both limited and biased, and undoubtedly inhibited model performance to some degree, but this was not the major limiting factor.

I found that the spatial extent and clustering of occurrences across geographical and environmental space within KNP was likely to have affected model performance the most. Species with few occurrences produced stronger performing models than those with greater numbers, matching the results of previous studies (Stockwell and Peterson 2002; Hernandez et al. 2006). This was however not due to low numbers of occurrences but rather due to the limited spread of the extents of these. In these instances, occurrences were so closely clustered together that there were little changes in environmental variability between them. Therefore, the machine-learning algorithm employed by MaxEnt would not be overly complex and this resulted in stronger models being produced (Phillips et al. 2006). Conversely, species which had higher numbers of available occurrence records tended to cover larger geographical spaces, causing an in increase in complexity as environmental variability was therefore greater and weaker models were produced as a result.

In addition to biases and limitations in occurrence data, my choice in spatial scale may also have affected model performance. A resolution of 1 km x 1 km offered the finer most resolution at which I could operate due to the limited availability of reliable climate and environmental data, although broader options were available. By opting to produce models at such a fine spatial scale across such a large area, and with limited occurrence data, there is little doubt that these choices will have affected model performance. Whilst many SDM focussed studies opt for a fine scale approach (Hurlbert and Jetz 2007), it may not always be the most effective choice. For the purposes of monitoring changes to reptile populations across KNP it made biological sense to use as fine a spatial scale as reasonably possible since monitoring is related to the scale at which those animals select microhabitats (Ferreira et al. 2011), but as evidenced here this approach may not be the most effective at yielding strong performing models.

The large variances in model performance amongst reptile species, and reptile families remain difficult to explain. Usually, differences in AUC scores obtained per species could be explained by some species being inherently better suited to SDM predictions than others (Elith et al. 2006; Radosavljevic and Anderson 2014). In the models I produced, since species had differing numbers of occurrences and these occurrences were directly linked to model performance, I cannot state that some reptile species or families had better modelability than others as there were clear biases present (Kery 2011). Species with lower numbers of occurrences produced the best models, largely because in those cases the densities of occurrences clustered closely together geographically (Stockwell and Peterson 2002; Pearson et al. 2007). This means that the environmental variability was likely to be relatively similar at each occurrence, and the resulting model thus performed well. This suggests that environmental variability across points rather than the number of, or area encompassing points is most likely the cause of varying AUC scores.

Alternatively, it may be that some of the reptile species within KNP are not suitable for climate modelling and required a greater number of environmental variables to produce meaningful models. Hernandez et al. (2006) found that in modelling the distributions of 18 taxa across California, variances in model performance was likely due to models for some species requiring additional variables relating to those species' actual distributional patterns. Hernandez et al. (2006) concluded that to produce stronger models they required additional environmental data rather than climate information. This may have been the case here in my study, where models for several species may have benefitted from the inclusion of additional, species-specific variables. Whilst this was included to some degree in the form of water layers, which undoubtedly benefited models for aquatic species (Pineda and Lobo 2009; Kery 2011), models for most species did not receive the same benefits. For example, individuals of *Broadleysaurus major* are known to sometimes inhabit unused termite mounds (Branch 1998;

Bates et al. 2014), and an inclusion of a variable representing this would likely have affected model performance for this species. However, since all species were modelled here using the same layers and parameters, species-specific benefits were largely absent from models.

Given that my aim was to produce distribution maps for all reptiles within KNP with the purpose of overlaying these, my approach here matched that of Overton et al. (2002) as a 'predict first, assemble later' strategy. This strategy allows for the observation of bigger picture patterns of reptile species richness despite potential inaccuracies within individual models (Feria and Peterson 2002). Some of my models did not obtain particularly high AUC scores and so the predictive ability for those species were unreliable (Merow et al. 2013). However, despite achieving less than ideal AUC scores, in several cases the predicted distributional ranges of species appeared to make reasonable biological sense. For example, the model for the primarily aquatic species *Varanus niloticus* achieved a relatively modest AUC score of only 0.70 (Baldwin 2009), but its predicted distribution closely followed riverine areas across KNP which is where this species would likely be present in reality (Pienaar 1976). As such, for some of my models AUC scores may not have been a meaningful measure of model performance or accuracy.

I found that principal components influenced models differently across reptile species and families. Whilst knowing which variables correlates most with the distributions of each reptile species and family is informative (Raxworthy et al. 2003), this knowledge in itself does not explain the variances in model performance across these taxa. Instead, these environmental correlates indicate the degree to which species are limited in their distributions (Guisan and Thuiller 2005; Brown et al. 2014), thereby influencing probabilities of occurrences within each grid cell rather than overall model performance. In reality, reptile species are limited in their ranges by different environmental factors (for example: crocodiles

are limited by the presences of rivers and other large water bodies; Branch 1998; Tolley et al. 2009) and my models have largely shown this despite less than ideal performance.

In the context of reptile management within KNP, distributional data is of considerable importance. Currently, TPC statuses of reptiles are assessed by monitoring changes to reptile populations at given sites every three years (Ferreira et al. 2011). Due to the low detectability of reptiles, on site-sampling is unlikely to yield true reptile richness patterns (Mazerolle et al. 2007), and as such the current system is inherently flawed. As suggested by Ferreira et al. (2011), a predictive inventory approach could better suit monitoring needs than on-site sampling for this group of animals. Here, my models explicitly predict where 119 reptile species are likely to occur within KNP at a 1 km x 1 km resolution. This is the first step towards testing the effectiveness of a predictive inventory approach, but it requires extensive ground-truthing to determine how these predictions fare within the real-world.

The approach of overlaying species distributions on top of each other to obtain species inventories and assemblage patterns has been successfully employed in previous studies. Feria and Petersen (2002) modelled and predicted the distributions of 89 bird species across south western Mexico using a similar 'predict first, assemble later' strategy as I used here. In the above mentioned study, Feria and Petersen (2002) empirically tested their predicted hypotheses via inventory sampling across several localities within their study site in order to compare their predictions with reality, and they found that their predictions were statistically accurate. This shows that using models to predict distributions for several species simultaneously is possible despite the difficulties associated with SDMS.

The predictions I have made here concerning the presences and absences of reptile species within and across KNP at a 1 km x 1 km resolution could potentially be used as an alternative means of monitoring to that of on-site sampling. However, model performance was variable

across species, and although AUC scores may have been a misleading measure in some cases, extensive testing of all models are required to properly and empirically assess their accuracies and potential use within the TPC and SAM framework.

Chapter 4: Ground-truthing model predictions

4.1 Introduction

The predicted distributions of reptiles occurring in KNP that I produced in chapter 3 were variable. The majority of models achieved passable AUC scores and would be deemed as meaningful attempts at accurately estimating the distributions of reptiles across KNP (Phillips et al. 2006; Phillips and Dudik 2008), but this was not the case for each model individually. Because there are always likely to be discrepancies between model predictions and reality (Stockman et al. 2006; Sarquis et al. 2018), even for those models that obtained AUC scores close to one (Lobo et al. 2008), it is important to test these predictions within the context of the real world (Bahn and McGill 2013). To empirically test the accuracies of my model predictions and assess how these translated into reality, ground-truthing efforts involving on-site trapping and sampling within KNP were necessary.

Effective trapping and sampling of reptiles can be done in several ways, the most efficient of which involves trap arrays combining pitfall traps, funnel traps, and drift fences (Maritz et al. 2007). This has become the standardized method for capturing lizards, snakes, and amphibians as it generally yields high rates of positive captures (Campbell and Christman 1982; Greene et al. 1999; Kuhnz et al. 2005). Used in conjunction with incidental and opportunistic observations of reptiles, this offers a relatively simple method of building up a dataset of reptile occurrences to test against predicted distributions.

In this chapter, I aim to empirically assess the accuracies of my model predictions of reptile distributions within KNP. Using Y-shaped trap arrays consisting of pitfall traps, funnel traps, and drift fences to trap reptiles, I built up a dataset in which to test my model predictions against. Whilst limited in terms of sampling time and the number of sample sites, this allowed for me to determine the degree of accuracy of some of my model predictions.

4.2 Methods

4.2.1 Site selection:

I trapped individual reptiles at six different sites within KNP. Each of these six sites were situated within the Skukuza granite super-site in the southern region of KNP (Figure 4.1) and consisted of upland and lowland locations. These sites were selected because: 1) they represented grid cells in which reptile occurrence data were absent or sparse (chapter 2), and so sampling here was not only useful for testing predictions but also for filling in a deficiency gap, 2) ease of accessibility in terms of safety and time constraints, and 3) permission to sample as required from SANParks. Ground-truthing were performed as a component of the Reptile Diversity in African Savannas field course (RDAS; www.studyafricanreptiles.org), with ethical clearance approved by the University of the Western Cape under permit number AR17/10/1. Ideally, ground-truthing should have taken place across a larger area of KNP but I was logistically limited and could not sample more extensively.

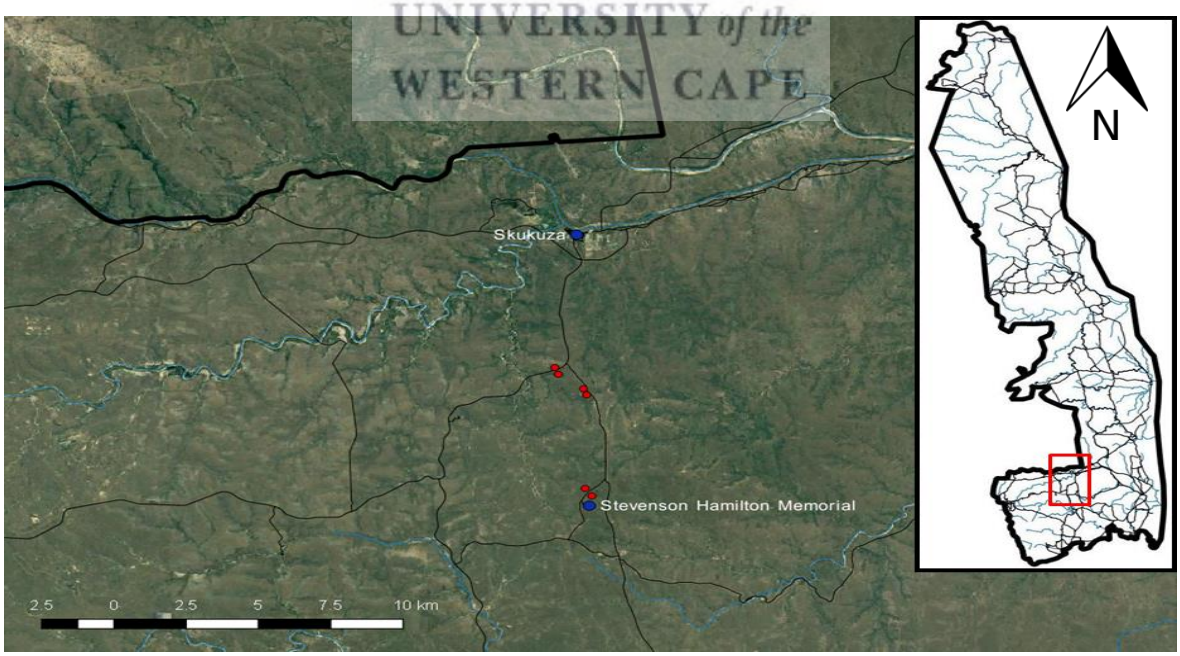


Figure 4.1: Locations of six ground-truthing sampling sites (indicated by red dots) within the greater Skukuza area of KNP.

4.2.2 Sampling design:

Together with a team of volunteers, at each site I installed a Y-shaped trap array. In similar fashion to Maritz et al. (2007), each of these trap arrays consisted of a combination of four pitfall traps, six mesh-entrance funnel traps, and three drift fences each 10 metres in length, arranged together in a 'Y-shape' (Figure 4.2). I constructed drift fences by stapling black plastic sheeting to wooden stakes that we buried into the ground at a standardized height of 300 mm. I stapled funnel traps to these wooden stakes towards the centre of the drift fences and I covered these with vegetation to provide shelter for potential individuals caught within. I also shaded each pitfall trap with black plastic sheeting and placed Petri dishes filled with water within each pitfall to limit the possibility of dehydration of individuals caught within. Together with volunteers, I checked traps twice a day from 04 December 2017 to 14 December 2017. In instances where we captured a reptile, we safely removed the individual from the trap before identifying and releasing it. In cases where other taxa were caught within traps, we safely removed and released those without recording.

4.2.3 Incidental searching:

In addition to sampling via traps, the volunteers and I also incidentally searched for reptiles within the greater Skukuza area throughout the same 10-day period. These searches consisted of game drives, on-foot searches within publically accessible areas such as campsites and picnic spots, and on-foot searching within typically non-accessible areas when game guards were available. I photographed all reptiles found and observed during these searches and identified individuals to the species level using field guides and expert opinion. In each case I recorded the GPS co-ordinates and locality where each individual reptile was observed.

4.2.4 Summarizing ground-truthed data:

For all analyses relating to ground-truthing efforts, I combined my trap data and incidental search data into one dataset. I categorized all reptiles found within the 10-day period in terms of group (chelonian, crocodylian, lizard, or snake), as well as their taxonomic family. In order to test whether I caught an even representation of reptiles across taxonomic groups, I compared the counts of species, and the counts of individuals caught based on reptile type and reptile family using four separate one-factor chi-square tests within SPSS software version 23.

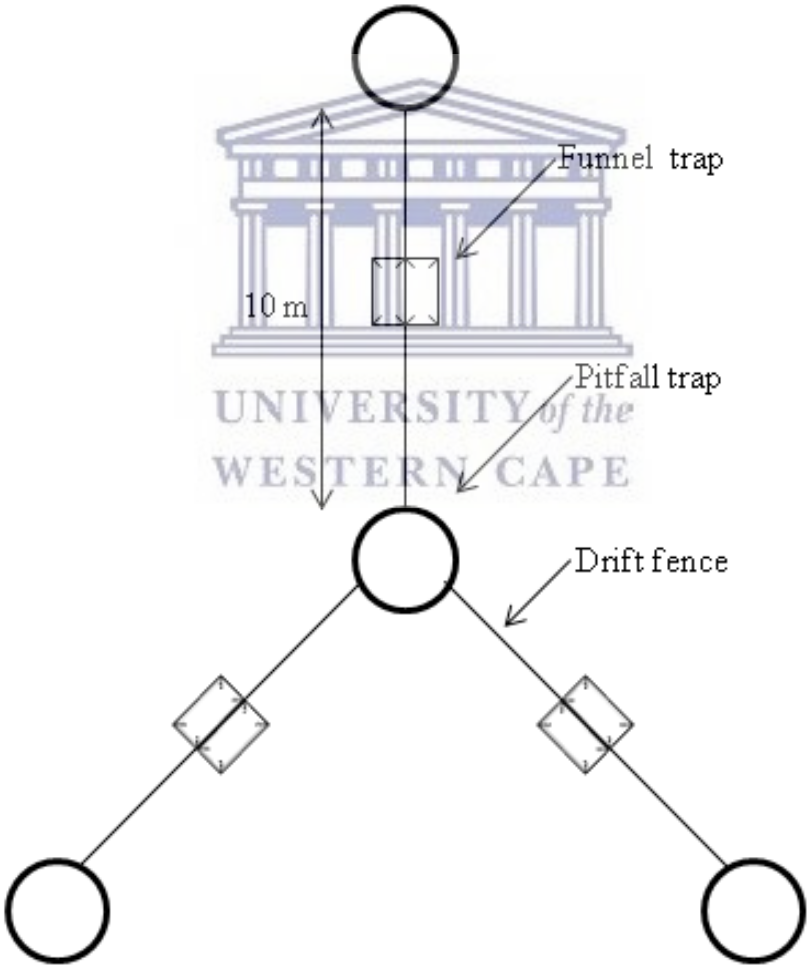


Figure 4.2: Y-shape trap array design consisting of funnel traps, pitfall traps, and drift fences used to capture reptiles.

4.2.5 Model testing:

I used my combined dataset of reptiles caught in trap arrays and reptiles observed incidentally (Figure 4.3) to evaluate the accuracies of model predictions. Given that this dataset consisted of presence-only observations, it was important to use an evaluation method that did not require absence data. Following Hernandez et al. (2006), I used prediction success as a measure of evaluating my model predictions. Prediction success was defined as the percentage of observations for each species that were correctly classified within models as being present within a grid cell. This evaluation method tests models for omission errors or false-absences, thereby giving an estimate of the number of true-presence predictions.

Since the evaluation dataset only contained observations of 36 reptile species, I only tested the models of those species. To test prediction success of each of these, I plotted all observations within the evaluation dataset (n = 151) within QGIS software version 3.2.3 and overlaid these upon their respective presence/absence maps (see chapter 3). For example, the locations of the four observations made of *Trachylepis striata* during trapping were overlaid onto the presence/absence map of this species. For each species I then counted the numbers of observations which occurred within grid cells where the species was predicted as present (true-presences), and the numbers of observations which occurred within grid cells where the species was predicted as absent (false-absences) and used these to estimate a percentage of prediction success. I then calculated average prediction success across all 36 species.

I then tested the hypothesis that models with higher AUC scores would have greater prediction success than those with lower AUC scores. To do this I used a linear regression analysis comparing AUC scores and prediction success of all 36 models. Similarly, I also performed a separate linear regression analysis to test the hypothesis that the numbers of observations used to evaluate models affected prediction success.

4.3 Results

4.3.1 Summary of ground-truthed data:

I recorded 151 occurrences of individual reptiles, comprising 36 species and 17 families (Table 4.1). These occurrences were not evenly distributed taxonomically and significantly differed amongst reptile groups, with the majority of observations being of individual lizards (Chi-Square test: $X^2_{df = 3} = 138.04$, $P < 0.01$; Figure 4.4a). The numbers of species observed per group also varied (Chi-Square test: $X^2_{df = 3} = 24.43$, $P < 0.01$; Figure 4.4b), with more lizard species being observed than species of any other reptile group. I also found family level taxonomic biases in the reptiles observed. I found significant differences in the numbers of observations of individuals per family (One-factor Chi-Square test: $X^2_{df = 11} = 126.11$, $P < 0.00$), with most observations belonging to members of the Testudinidae. Conversely, I found no significant differences present in the number of species observed per family (One-factor Chi-Square test: $X^2_{df = 5} = 5.67$, $P = 0.34$) as most families only had single representatives.

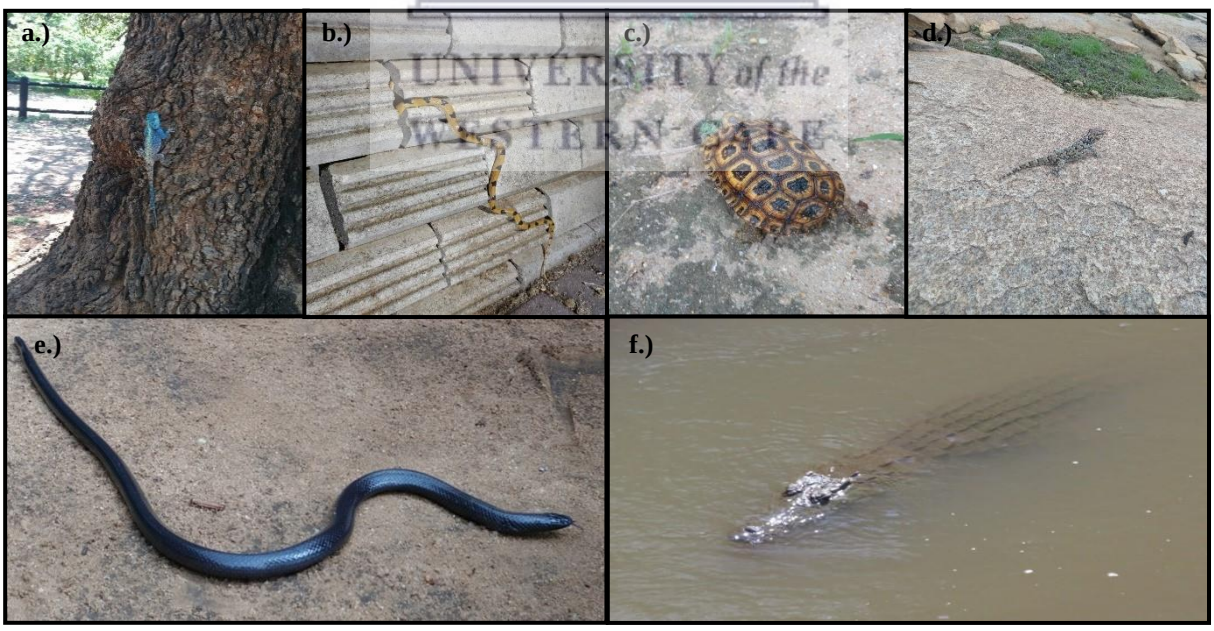


Figure 4.3: Examples of reptiles incidentally observed within the greater Skukuza area: a) *Acanthocercus atricollis atricollis*, b) *Telescopus semiannulatus semiannulatus*, c) *Kinixys spekii*, d) *Chondrodactylus turneri, e) Amblyodipsas polylepis*, and f) *Crocodylus niloticus*.

Table 4.1: Reptile species observed and caught in trap arrays during 10 days of sampling within the greater Skukuza area of KNP.

Species	Number of observations	Test AUC
Acanthocercus atricollis atricollis	3	0.64
Acontias plumbeus	1	0.69
Afrotyphlops schlegelii	4	0.68
Amblyodipsas polylepis polylepis	1	0.71
Bitis arietans arietans	2	0.61
Broadleysaurus major	2	0.65
Chamaeleo dilepis dilepis	3	0.64
Chondrodactylus turneri	3	0.65
Cordylus jonesii	1	0.54
Crocodylus niloticus	3	0.74
Dendroaspis polylepis	1	0.65
Dispholidus typus typus	1	0.61
Gerrhosaurus intermedius	5	0.69
Hemidactylus mabouia	2	0.66
Hemirhagerrhis nototaenia	1	0.58
Homopholis wahlbergii	1	0.64
Kinixys spekii	14	0.78
Leptotyphlops incognitus	1	0.77
Lygodactylus capensis capensis	5	0.68
Matobosaurus Validus	5	0.61
Mochlus sundevallii sundevallii	4	0.68
Nucras holubi	19	0.74
Panaspis wahlbergii	4	0.71
Pelusios sinuatus	3	0.65
Philothamnus semivariegatus	2	0.65
Prosymna stuhlmannii	1	0.65
Psammophis subtaeniatus	1	0.67
Psammophylax tritaeniatus	3	0.58
Stigmochelys pardalis	13	0.67
Telescopus semiannulatus semiannulatus	1	0.57
Thelotornis capensis	1	0.65
Trachylepis margaritifer	8	0.70
Trachylepis striata	4	0.70
Trachylepis varia	22	0.69
Varanus albigularis albigularis	4	0.61
Varanus niloticus	2	0.70
Average	**4**	**0.66**

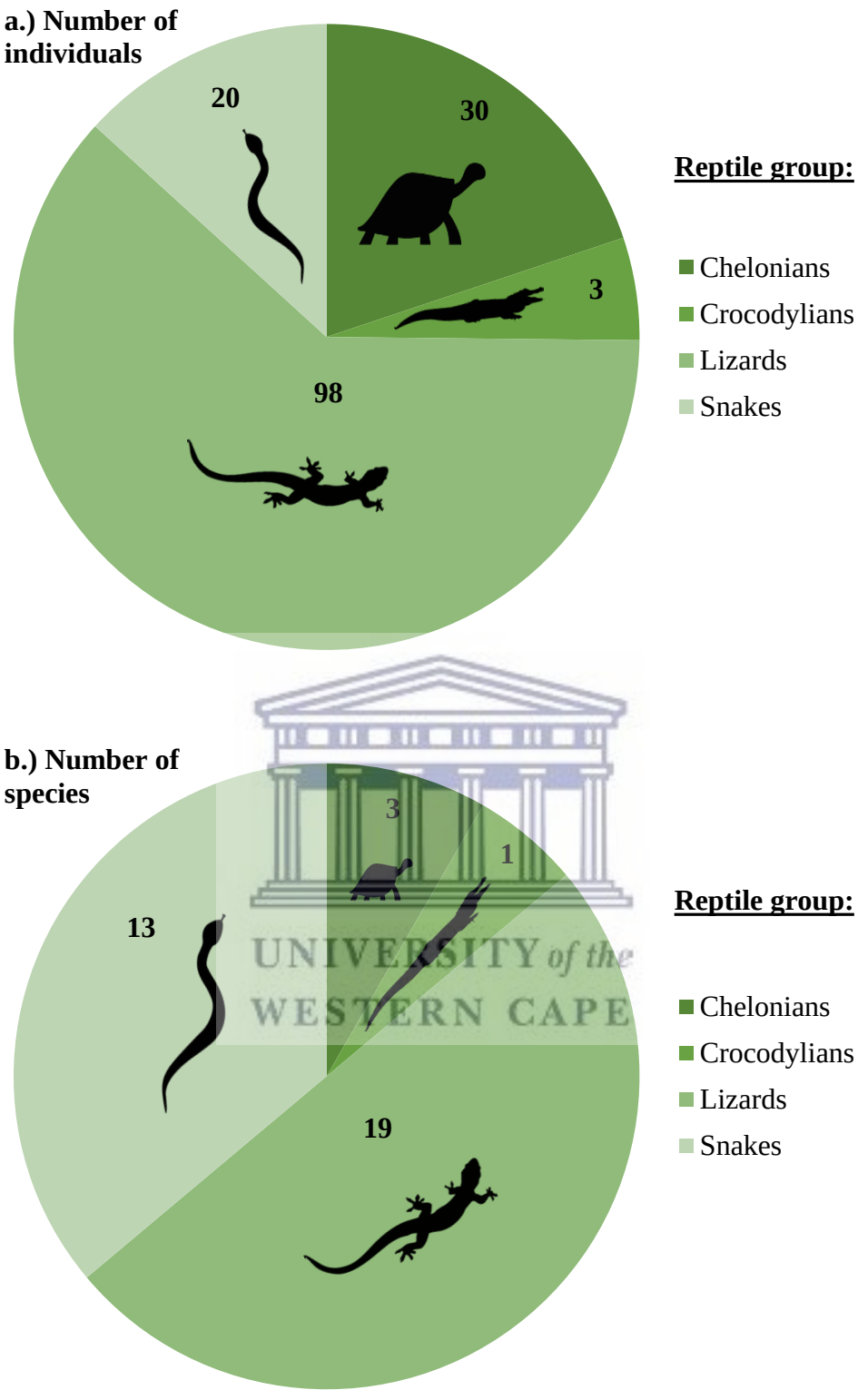


Figure 4.4: Numbers of observations of a) individual reptiles and b) reptile species per reptile group.

4.3.2 Model testing:

In empirically testing the accuracies of my model predictions, I found that the models I produced did not accurately predict species' presences and absences in all cases. In total, of the 151 observations of reptile occurrences, only 103 (68 %) of these occurred within grid cells where the models predicted those species should occur. Of the 36 models I specifically tested, 19 of these had 100 % prediction success but the remaining 17 models all had varying numbers of omission errors (Figure 4.5). Several reptiles were observed in reality within grid cells where I predicted them as being absent. For example, during sampling I observed three individuals of *Chamaeleo dilepis dilepis* at three different locations, but only two of those observations occurred within grid cells where my modelled distributions predicted the species as being present. These types of omission error rates varied across these 17 flawed models, with some having only a single false-absence whereas others consisted entirely of false-absences and having 0 % prediction success. I found that average prediction success across all 36 tested models was 65 %, with the remaining 35 % being attributed to false-absences.

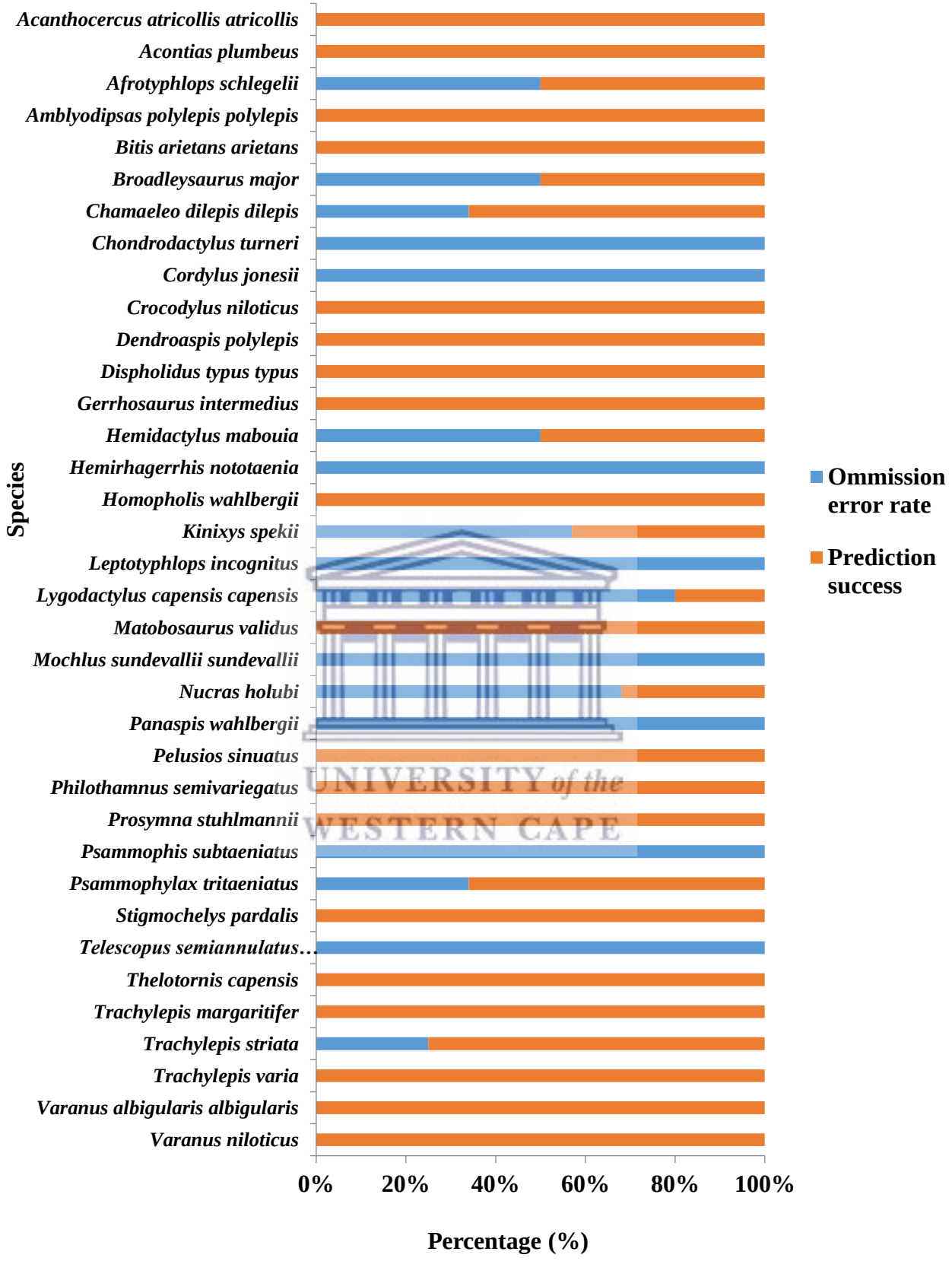


Figure 4.5: Prediction success of 36 reptile species distribution models.

In testing if there were any relationships between AUC scores and prediction success obtained for each model, I found that this relationship was not present within these 36 models ($F_{1, 35} = 0.02$, P = 0.88). Models that obtained higher AUC scores did not obtain higher prediction success and there was no apparent trend (Figure 4.6a). Similarly, I also found no relationship between the number of observations used to evaluate models and the prediction success obtained ($F_{1, 35} = 0.12$, P = 0.73). Again, there was no apparent trend as models which had higher numbers of observations to test against did not achieve statistically higher or lower prediction success than those with fewer observations (Figure 4.6b).

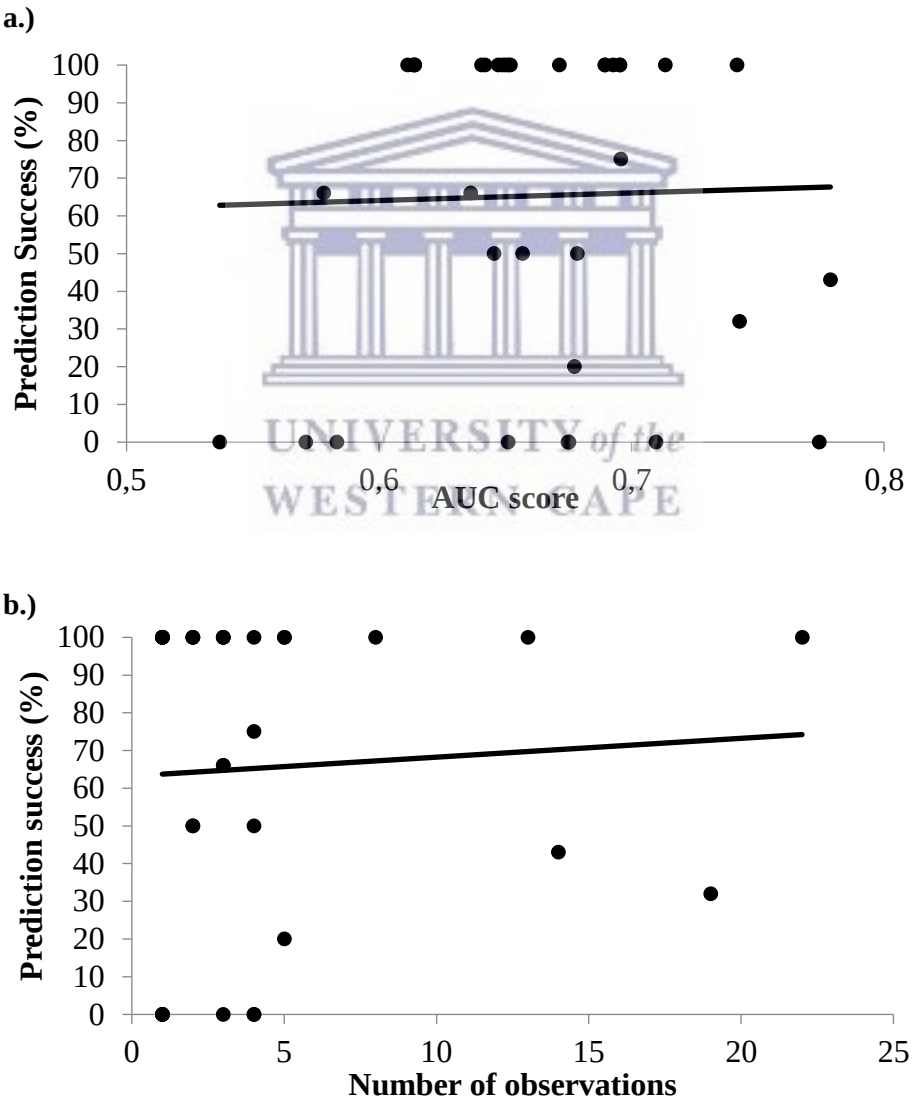


Figure 4.6: Comparisons between prediction success of 36 reptile species distribution models and a) AUC scores of each model, and b) numbers of observations used to test each model.

4.4 Discussion:

The predicted reptile species distribution models I produced in chapter 3 achieved good AUC scores (average AUC = 0.75), but in terms of their real world performance they had variable rates of prediction success. On-site trapping and ground-truthing within KNP allowed me to build up a dataset of 151 individual reptile observations to test against modelled distributions of 36 species, the results of which showed that these models had an average prediction success of approximately 65 %. The remaining 35 % was attributed to false absences (omission errors) in model predictions as I observed several individual reptiles in areas where those species were not predicted to occur.

As my ground-truthing efforts were limited in terms of both locations and time, there are several implications to consider. Firstly, all of my trapping took place within the Greater Skukuza area, which whilst spatially diverse is not a true representative of KNP as a whole. My trapping efforts were therefore more akin to an extensive sampling of this particular area as opposed to ground-truthing of KNP. It is possible that trapping success could vary across the landscape of KNP and so by limiting my trapping to one location this means that we still do not know how prediction success of my distribution models would fare across the rest of KNP. Secondly, because trapping was confined to so few days my trapping cannot be deemed as being a thorough measure of reptile diversity within those sites. Intensive sampling via trapping can take months or years to yield true reptile species richness patterns (Maritz et al. 2007). Because my sampling was constrained logistically it undoubtedly had an effect on the available data to test against model predictions.

Since my sampling was limited, I did not expect to obtain complete inventories of reptile species at each sampling site, and more broadly, the grid cells where sampling took place. Each trap array was limited in terms of placement as they required habitats which allowed for digging holes in the ground. As such, I did not place arrays in close proximity to several typically suitable reptile habitats such as rocky outcrops or riverine areas (Branch 1998;

Bates et al. 2014) and I was therefore unlikely to capture species associated with those habitats. Additionally, my modelled predictions were made at a spatial scale of 1 km x 1 km, but each trap array only covered a fraction of that area. Since all trap arrays were situated within separate grid cells, this meant that in each case the species caught would only represent a subset of reptile species richness within said grid cell. These limitations coupled with the low detectability of reptiles and limited sampling time meant that several other reptile species were undoubtedly present within those areas but remained undetected.

In terms of trapping, my approach matched that of Maritz et al. (2007) in which pitfall traps, funnel traps, and drift fences used in conjunction were employed in Y-shaped arrays to capture reptiles. This is generally accepted as the preferred method of reptile sampling (McDiarmid et al. 2012) and proved successful here where I caught 55 individual reptiles (in addition to a further 96 incidental observations) within a limited sampling period of only 10 days. Model testing was therefore limited to comparing the observed occurrences of reptiles against predicted distributions for only 36 species to assess omission errors (false-absences) rather than commission errors (false-presences) which I was unable to test. Nevertheless, each individual captured or observed offered useable information for testing my predictive hypotheses of reptile distributions.

Prediction success of tested models varied per species and was unconvincing overall. As discussed in chapter 3, model performances in terms of AUC scores were good but less than ideal, and this translated into reality as nearly half of all tested models contained omission errors. However, I found that for those 36 tested models there was no relationship between AUC scores and model performance, suggesting that for these models, and likely the untested models as well, AUC scores may not have been the correct indicators of model strength (Lobo et al. 2008).

Empirical testing of SDMs in previous studies have met greater success that what I observed here. In testing the prediction success of models representing 18 different taxa across California, Hernandez et al. (2006) reported an average success rate of approximately 90 % for MaxEnt models that they developed using 100 occurrence records for each species. However, they also produced models using only five occurrences per species and these models had an overall lower prediction success (approximate average = 60 %) that was more line with what I observed here. In a separate study, using an alternative approach to prediction success, Sarquis et al. (2018) tested the predicted distribution of *Bothrops alternatus* in Argentina by comparing SDMs against the known empirical distribution of this species. They found that each of the different models they produced either over-estimated or sub-estimated the distribution of this species, but that the MaxEnt models were among those that most closely resembled its empirical distribution. These examples show that SDMs can, and have, been reasonably accurate when they are empirically tested, but that even with high success rates they are unlikely to perfectly predict species' true distributions as there will always be elements of overestimation or under-estimation present.

Since prediction success varied across the 36 models I tested, it is reasonable to assume that it would also vary across the remaining, untested models and that several of those would also contain errors. As such, the usage of all 119 models in further analyses and applications should be employed with caution. Although the models of some species achieved 100 % prediction success, several others were undeniably flawed. However, all 119 models provide valuable baseline data as the errors found here were present at a 1 km x 1 km spatial scale and might not be present at broader resolutions. Overall, it is clear that the models I produced require refinement if they are to obtain high AUC scores (> 0.9), and high prediction success, but it is unclear if that requires higher quality input data or different modelling approaches as both of these could potentially solve the issues present here.

Chapter 5: Reptile assemblage structure of KNP

5.1 Introduction

A central aim of biogeography is the classification of organisms into meaningful groupings in the form of biogeographical units (Mackey et al. 2008; Kreft and Jetz 2010). These units are an important component of biogeography as they allow for analyses of the geographical organization of the world's biota (Linder et al. 2012), which in turn allows for the development of a spatially explicit framework (Moura et al. 2017) that can be used to answer ecological questions and assist in conservation management (Kreft and Jetz 2010). These questions include those relating to dispersals, and distributions of species (Carstensen et al. 2013). By delineating a region into biogeographic units on the basis of its unique biota, we can identify areas which are most in need of conservation. This highlights the need to delineate regions as this process can prove inform conservation planning.

Delineating a region into biogeographic units based on compositional dissimilarity is known as biogeographical regionalization (Kreft and Jetz 2010; Moura et al. 2017). This procedure aims to separate a geographical area into spatially segregated units in which species compositions are broadly similar within units, but significantly differs across them (Mackey et al. 2008). In this way, important community ecology and biogeographical questions can be answered relating to monitoring, managing and conservation of organisms within a given area (Morrone 2009; Brown et al. 2014). For example, the boundaries at which biogeographic units are separated may be related to underlying ecological factors and may be informative in identifying physical barriers or other features that restricts a species' distributional range. Identifying these boundaries could be an important consideration in affecting conservation decisions within protected areas but this is largely dependent on the extent of the area in question as well as the spatial scale at which biogeographical regionalization is undertaken.

Studies focussing on biogeographical regionalization have most often operated at considerably broad scales. The majority of studies aiming to delineate areas into spatially segregated biogeographic units have been performed at national, continental, and even global scales (For example: Minter et al. 2004; Linder et al. 2012; Moura et al. 2017). Whilst informative, studies such as those do not offer much value in terms of management and conservation decisions at the localized scale of a national park where this information would be beneficial (Whittaker et al. 2005). This brings about the need for biogeographical regionalization at smaller scales (Moura et al. 2017), particularly for those groups of organisms whose biogeographic patterns are largely unknown.

Although covered briefly by Pienaar (1978), the biogeography of reptiles within KNP has not been extensively studied. KNP could be delineated into several biogeographic units, including individual assemblages, based on the compositions of reptile species within grid cells across its landscape (Minter et al. 2004). Reptile monitoring within KNP is largely dependent on knowing where each species occurs within the park and subsequently observing changes to the distributions of populations over a given period (Ferreira et al. 2011). By defining assemblages across KNP, this could facilitate the identification of factors which drives or inhibits reptile distributions within and across the park (Mackey et al. 2008; Moura et al. 2017). This requires specific distributional information for each reptile species as all methods of delineation requiring knowledge of species' presences and absences (Kreft and Jetz 2010).

Several methods are available for delineating a region into biogeographic units. This generally involves the use of cluster analysis techniques, of which there are several, and species dissimilarity matrices (Kreft and Jetz 2010). A cluster analysis assesses compositional dissimilarity between sites and uses this to classify similar sites together into meaningful groupings (Everitt 1993). This is usually carried out via K-means partitioning or

agglomerative hierarchical clustering (Kaufman and Rousseeouw 1990; Legendre and Legendre 1998; Linder et al. 2012). The K-means partitioning method allows for a user to specify a number of desired clusters, and partitions data based on their means around a set of stopping points (Macqueen 1967; Linder et al. 2012). Conversely, agglomerative hierarchical clustering does not require the user to specify a desired number of clusters, but instead produces a hierarchy of clusters at various stopping points. As mentioned by Kreft and Jetz (2010), hierarchical clustering is usually the more informative choice for biogeographical regionalization as biogeographic regions tend to be hierarchically arranged in nature.

In this chapter I aim to perform biogeographic regionalization of KNP at a spatial resolution of 1 km x 1 km based on the distributions of 119 reptile species so as to delineate the park into spatially segregated reptile assemblages. I opted for the hierarchical clustering approach as this would allow for the development of meaningful groupings without *a priori* providing a desired number of clusters before analyses, which allowed me to infer groupings based on produced outputs. I hypothesize that based on predicted presences and absences of reptile species across KNP, several distinct and meaningful assemblages will be present, with numerous significant splits occurring at broader levels. At the finer most level I predict that species richness, diversity, and endemism across assemblages will statistically vary across biogeographic units.

5.2 Methods

5.2.1 Protocol:

The protocol of biogeographic regionalization requires several multivariate steps. These steps (Figure 5.1) are outlined further below where applicable as some have been completed within previous chapters.

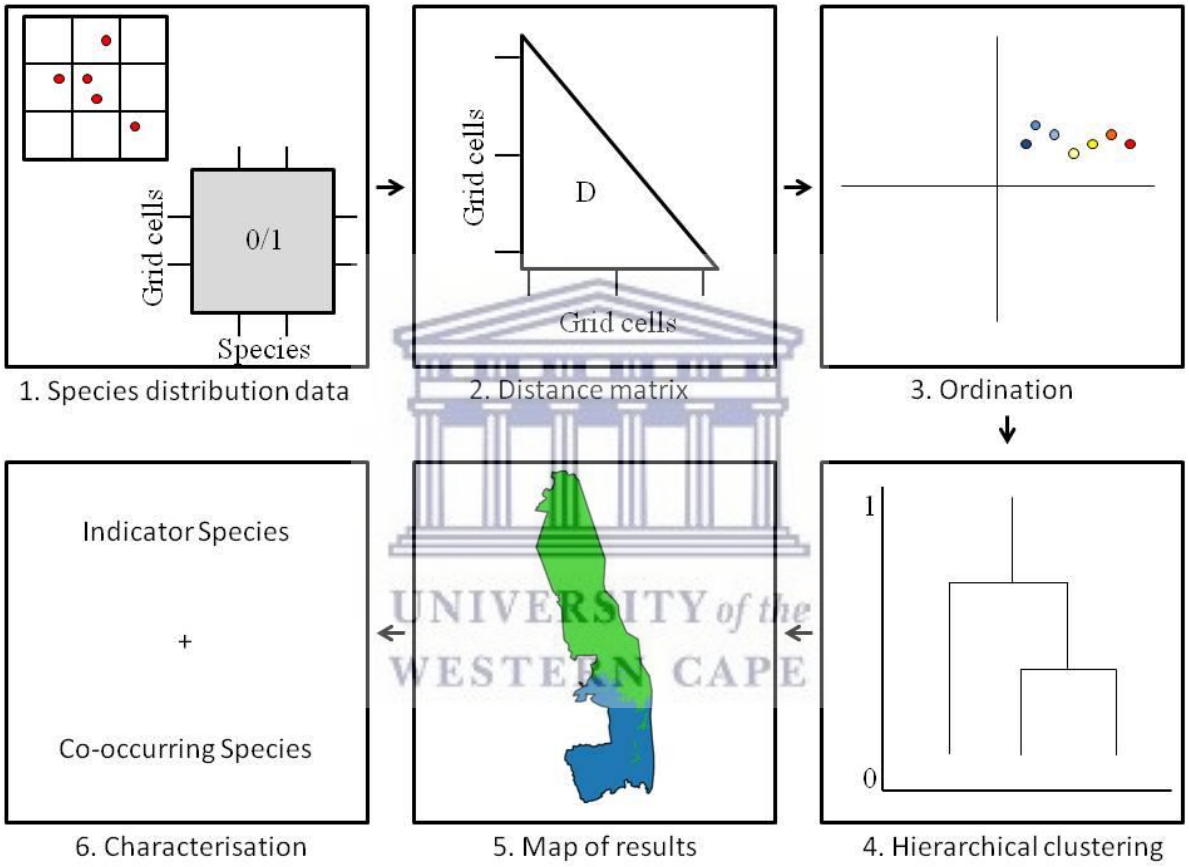


Figure 5.1: Conceptual diagram displaying the multivariate steps required for the protocol of delineating a region into biogeographic units.

5.2.2 Ordination:

A key component of assigning grid cells to biogeographic units is that of ordination. Ordination allows for the identification of clean breaks among units within geographical space or grid cells (Storch et al. 2003; Kent 2006) and can be used to show transitions

between them. To test the hypothesis that reptile species were not randomly assembled across KNP, I used non-metric multidimensional scaling (NMDS) ordination to produce projections of geographical relationships among grid cell compositions within two-dimensional space. NMDS ordination is widely regarded as the most appropriate method of ordination as it is relatively unconstrained (Minchin 1987; McCune et al. 2002; Kreft and Jetz 2010). I performed NMDS ordinations at both the species and family levels using the 'metaMDS' function of the 'vegan' package (Oksanen et al. 2007) in R software version 3.4. To find the best possible ordination solution, I used a Bray-Curtis index to calculate pairwise distances between grid cell compositions with 100 random starts to limit errors (Kreft and Jetz 2010). I then rescaled the ordination axes to fit between zero and one to better visualize the spread of data across two-dimensional space.

5.2.3 Cluster analysis:

To test the hypothesis that assemblages of reptile species across KNP were not random in geographical space, I performed an agglomerative hierarchical cluster analysis. To do this, I first compiled a matrix of reptile species presence/absence per grid cell for the entirety of the park (n = 21761; see section 3.3.5). I then used a Bray-Curtis similarity index to calculate pairwise differences between grid cells and create a similarity matrix. Using hierarchical agglomerative clustering I then performed a multivariate cluster analysis on this similarity matrix to cluster all 21761 grid cells into biogeographic units based on their species compositions. Essentially, each biogeographic unit would represent a different composition of reptile species occurrences over several grid cells across KNP (Minter et al. 2004). To do this, I used the R package 'cluster' (Maechler et al. 2016) to create the Bray-Curtis similarity matrix and produce a dendrogram showcasing compositional dissimilarity between grid cells, which allowed for comparisons between species lists at multiple levels. As recommended by

Legendre and Legendre (1998), I used Ward's minimum variance linkage method to prevent potential errors occurring within the multivariate analysis

Since hierarchical clustering produces clusters at several levels without any definition of when to cease (number of clusters = total number of grid cells - 1), I needed to assign a stopping point within the splitting of hierarchies where biogeographic clusters would be defined (Eviritt 1993). Currently, there is no standardized, objective method to determine which clusters are defined as the 'best set' (Minter et al. 2004). Despite this lack of a rule, it is also important to note that final clustering should make biogeographic sense and therefore any arbitrary stopping rules should be employed with caution (Everitt and Hothorn 2011).

To ensure my groupings would make sense, I used the 'dendextend' R package. This package allows for assemblage stopping points to be made based on either a user-input desired number of clusters or at a specific tree height (Galili 2015). Based on the visualization of the produced dendrogram, I used three separate stopping points to group grid cells into two, five, and nine clusters respectively as these appeared to make the most sense biogeographically. Further partitioning into a higher number of clusters ran the risk of producing unidentifiable biogeographic units, especially since there were limitations to my input data. I therefore did not opt to include additional stopping points.

5.2.4 Assemblage biogeography:

In order to visually map out the cluster arrangements, I used QGIS software version 3.2.3 to geographically arrange clusters into biogeographic regions across KNP. My hierarchical cluster analysis categorized each grid cell as belonging to a particular biogeographical unit at each stopping point. To represent this visually, I used these categorizations to assign a specific classification of each grid cell within QGIS. I did this using the 'join' function to

merge the categorizations with their respective grid cells at each stopping point to produce maps of reptile biogeographic units across KNP.

5.2.5 Indicator species:

At each stopping point, the various reptile biogeographic units present across KNP were comprised of spatially unique compositions of species predicted to occur in those locations. As such, each biogeographic unit could be exemplified by certain indicator species which acted as representatives of that particular unit. Usually, indicator species are those which are typical of a cluster with their distributions approximately matching the spatial boundaries of the assemblage as a whole (Minter et al. 2014). To find the indicator species of each predicted assemblage, I used the 'labdsv' package (Roberts 2016) in R software version 3.4 to identify which species best represented each unit. This package follows the methods suggested by Dufrene and Legendre (1997) in which each species is ranked according to an indicator value index. This index is expressed as a value between 0 – 100 % with higher values representing greater importance towards a group. This package calculates the mean abundance of each species per site compared to all other sites, by the relative frequency of the occurrence of each species in each site, and thereby classifies the best representative species.

5.2.6 Co-occurrence:

Species predicted as being indicators of a particular biogeographic unit may be those with low detection rates (for example, any of several burrowing species) and may prove to be unfavourable targets for empirical testing. To compensate for this, I performed a species co-occurrence analysis using the R package 'cooccur' (Griffith et al. 2016) to determine the probabilities of each species to occur within the same grid cell as each other. Those predicted to co-occur with indicator species could thus be treated as proxies, thereby aiding in sampling as these species could be more suitable as targets in instances where indicators were cryptic.

5.3 Results

5.3.1 Ordination:

My NMDS ordination analysis produced a projection of compositional dissimilarity between grid cells within two-dimensional space (Figure 5.2). I obtained a low stress value of 0.2327, indicating good, but not perfect representation. My ordination showed that grid cells with similar reptile species compositions were grouped closer together than those with vastly differing compositions, and showed clear, continuous transitions between them. The ordination showed clear separations between grid cell assemblages within two-dimentional space, and overall, the NMDS plot showed distinct relationships amongst grid cells which indicated the presence of various biogeographic units across KNP.

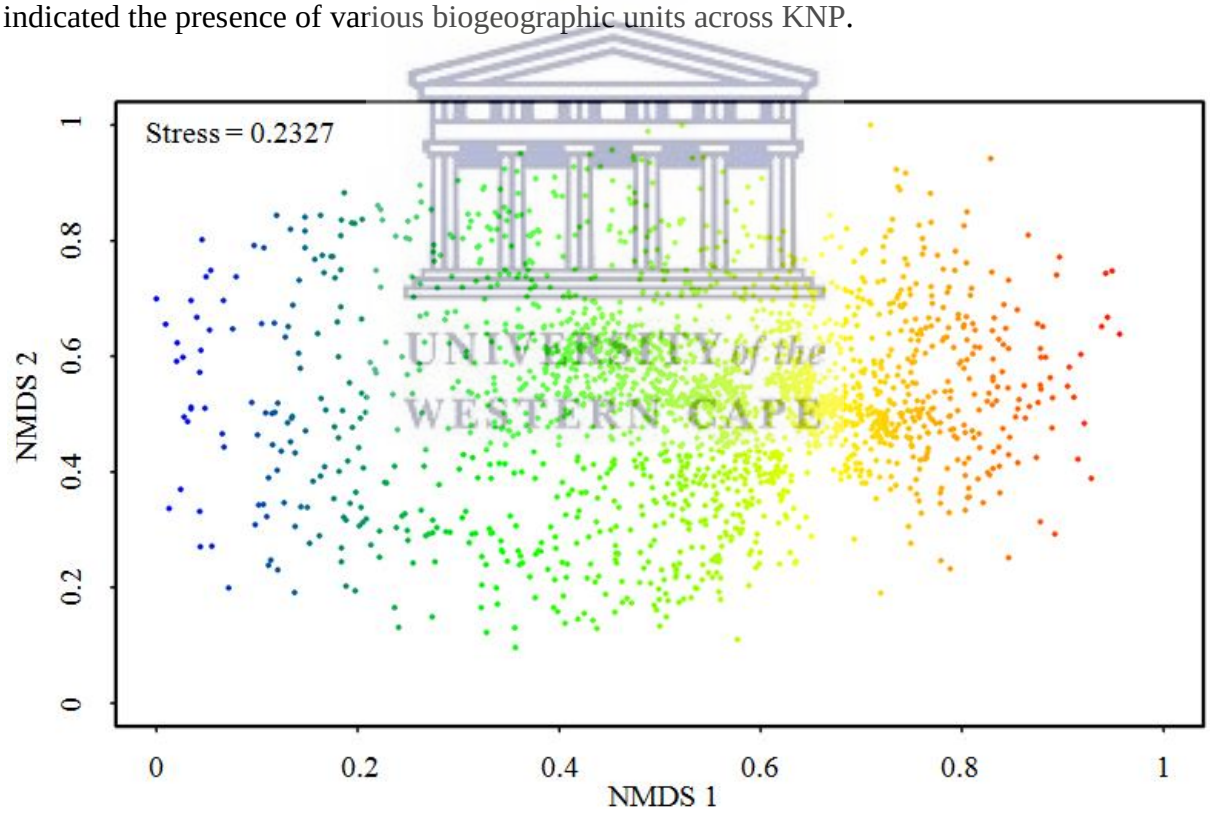


Figure 5.2: Non-metric multidimensional scaling (NMDS) of predicted reptile compositions within 1 km x 1 km grid cells across KNP based on a Bray-Curtis dissimilarity matrix. Each dot represents a specific grid cell, with colours indicating similarity.

5.3.2 Cluster analysis and assemblage biogeography:

In hierarchically clustering grid cell compositions of reptile species across KNP, I found that distinct biogeographic units were present at different stopping points along the produced dendrogram, culminating in a total of nine geographically distinct reptile assemblages. The initial split at the first stopping point produced two distinct clusters. These clusters were separated geographically between the northern and southern regions of KNP (Figure 5.3) and would best be described as being separate sub-regions of KNP as they were only 30 % dissimilar and had limited differentiation in species richness, diversity, and endemism (Table 5.1; One-factor Chi Square Test: $P > 0.05$ in all cases). Representatives of all 19 reptile families were present within both sub-regions, but the Northern sub-region had more endemic species than the Southern sub-region.

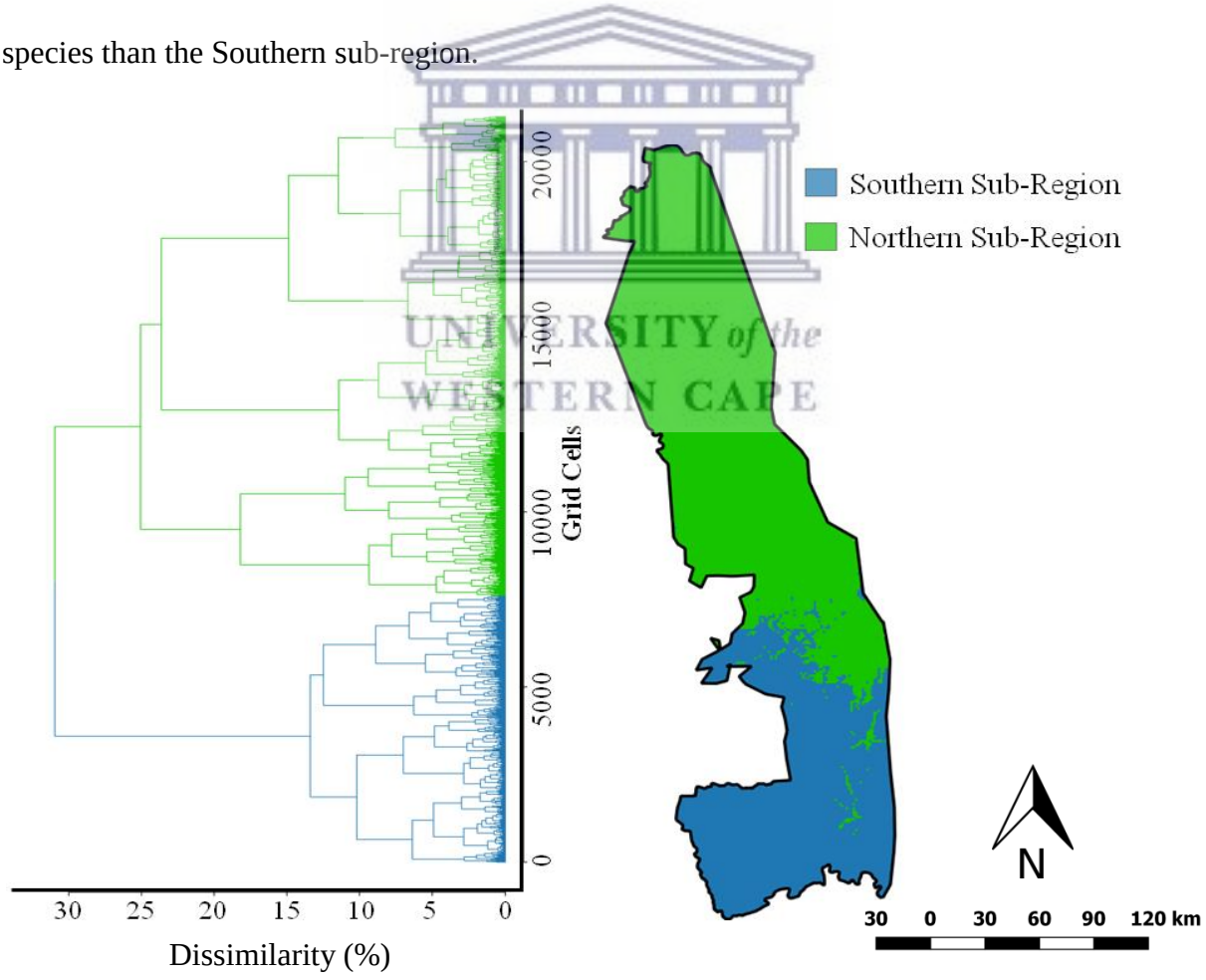


Figure 5.3: Dendrogram and map of predicted reptile sub-regions across KNP resulting from agglomerative hierarchical clustering based on Bray-Curtis dissimilarity.

The next major splits occurred within the Northern sub-region, which separated into four major districts before the Southern sub-region experienced any further splits. This resulted in a total of 5 sub-regions and districts (Figure 5.4). These biogeographic units were approximately 18 % dissimilar, with species diversity remaining relatively similar across the various districts (Table 5.1; One-factor Chi Square Test: $X^2_{df = 4} = 1.86$, P = 0.76). Species richness was also statistically similar across units (One-Factor Chi Square Test: $X^2_{df = 4} = 8.41$, P = 0.08), but had notably large differences (as high as 20 species) between them. The Northern sandveld district had the highest number of reptile species predicted to occur within that unit and the Mopani district had the least. Again I found that endemism was not significantly different across these units (One-Factor Chi Square Test: $X^2_{df = 4} = 0.33$, P = 0.56).

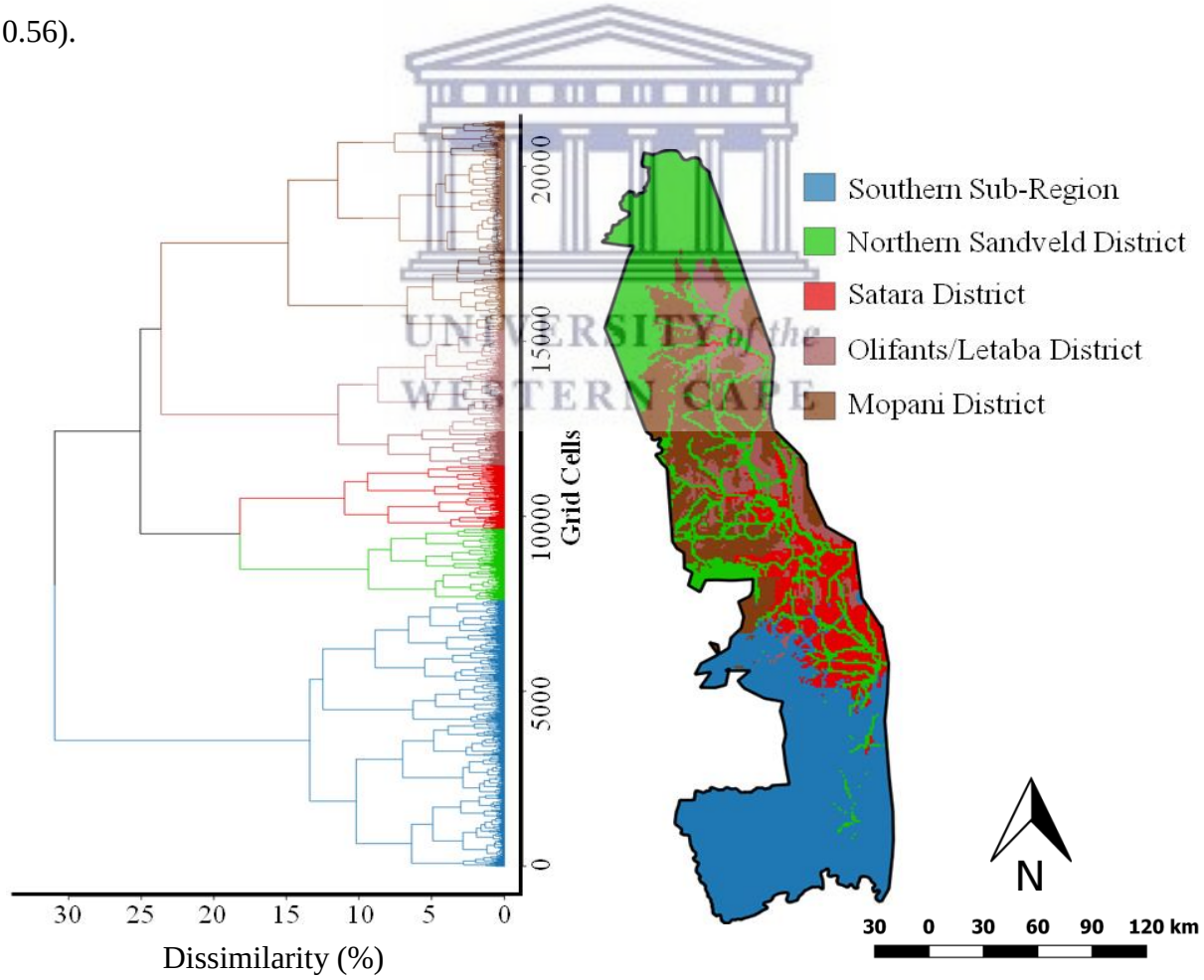


Figure 5.4: Dendrogram and map of predicted reptile districts across KNP resulting from agglomerative hierarchical clustering based on Bray-Curtis dissimilarity.

The final significant split within the dendrogram resulted in a total of nine geographical assemblages across KNP (Figure 5.5). The Southern sub-region split into three separate assemblages, namely the South Western assemblage, the Skukuza assemblage, and the Lower Sabie assemblage. Additiontally, the Northern sandveld district split into the North Eastern assemblage, the Riverine assemblage, and the Northern assemblage. The Satara, Olifants/Letaba, and Mopani districts all remained intact. These nine units were approximately 13 % dissimilar with species diversity (Table 5.1; One-Factor Chi Square test: $X^2_{df = 7} = 389.59$, $P < 0.01$) and species richness both significantly differing across assemblages (One-Factor Chi Square test: $X^2_{df = 7} = 114.81$, $P < 0.01$). I found that Endemism across assemblages did not differ statistically (One-Factor Chi Square test: $X^2_{df = 7} = 3.25$, $P = 0.19$), with only the northern assemblage having a comparitively high level of endemism.

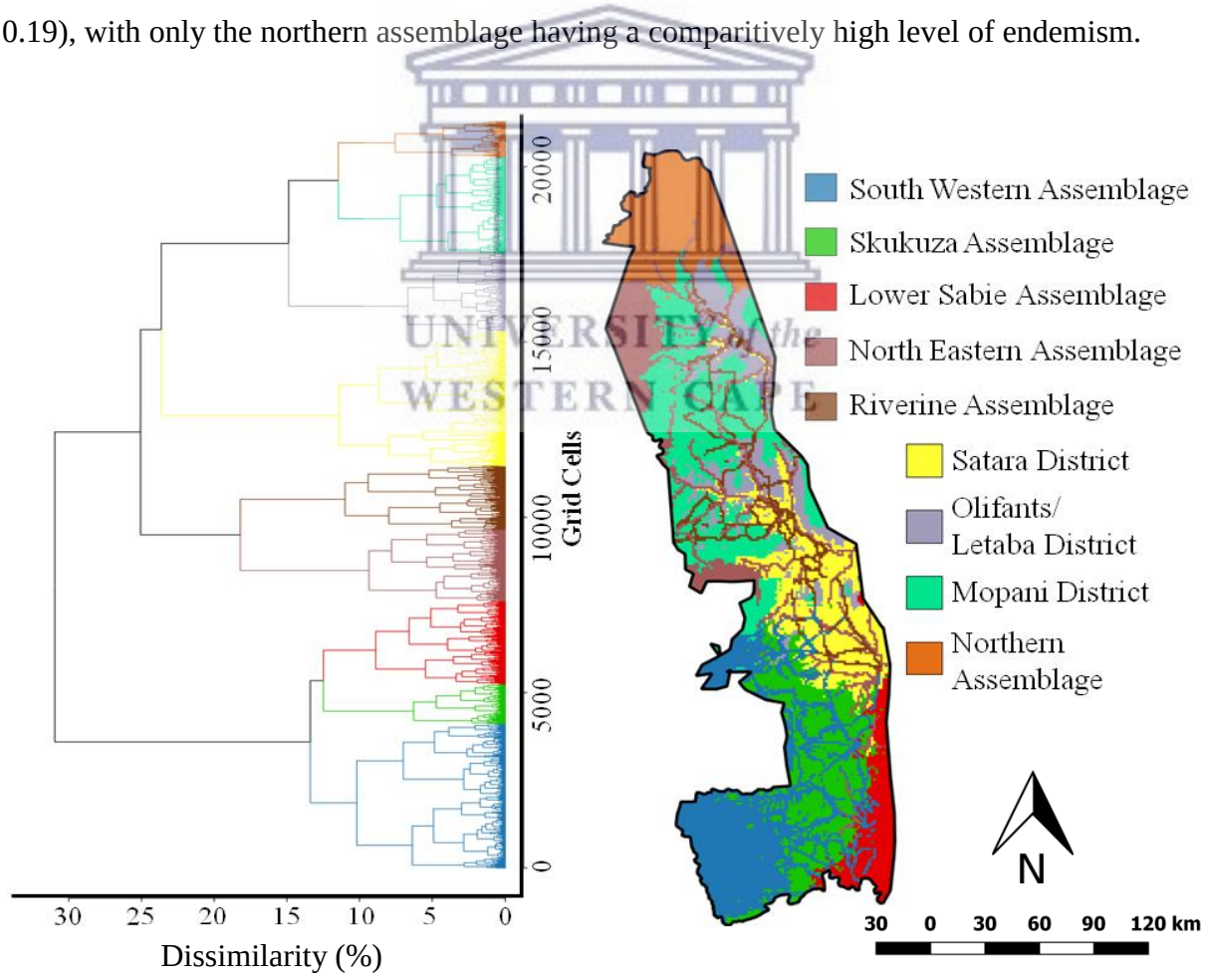


Figure 5.5: Dendrogram and map of predicted reptile assemblages across KNP resulting from agglomerative hierarchical clustering based on Bray-Curtis dissimilarity.

Table 5.1: Quantities of reptile species richness, endemism, and diversity across biogeographic units within KNP.

Biogeographic unit	Species richness	Endemic species	Species diversity (H')
Southern sub-region	**108**	**5**	**4.68**
South-western assemblage	96	2	4.56
Skukuza assemblage	86	0	4.45
Lower Sabie assemblage	101	1	4.61
Northern sub-region	**114**	**11**	**4.74**
Mopani district	85	0	4.44
Olifants/Letaba district	83	0	4.42
Satara district	87	0	4.47
Northern sandveld district	113	7	4.73
Northern assemblage	107	5	4.67
North-eastern assemblage	101	0	4.61
Riverine assemblage	77	0	4.34
Average	**91**	**1**	**4.51**

5.3.3 Indicator and co-occurring species:

In attempting to identify indicator species for each biogeographic unit I produced lists containing the three species which best represented each unit. However, these lists were dubious as the identified indicator species did not appear to make reasonable biological sense. For example, widespread, commonly occurring species that are present throughout KNP such as *Chondrodactylus turneri* or *Lygodactylus capensis capensis* were identified as being indicators for certain units. Commonly occurring species such as these could not feasibly be considered as indicators as they represent multiple units as opposed to specific, individual ones. As a result, I therefore opted not to include the results of this analysis. I also produced lists of species with the highest probabilities of co-occurring with identified indicator species, however, since I omitted those results, I therefore omitted the co-occurring species results as well.

5.4 Discussion

Global habitats, including those within protected areas may be spatially segregated or delineated and thus the distributions of species within them are unlikely to be uniform (Minter et al. 2004). Rare and endangered species may be linked to certain habitats or biogeographic units and so the identification of these units as well as which species occupies them could play a critical role in assisting with managing these species. Here, I have developed a protocol for delineating KNP into an assortment of biogeographic units at different hierarchies based on predicted reptile distributions. This process involved agglomerative clustering of predicted presences and absences of reptile species across within grid cells KNP into spatially segregated units within the park. At the finer-most level I delineated KNP into nine reptile assemblages, spatially and geographically arranged across the park.

The protocol I have developed here provided a good foundation for delineating KNP into different biogeographic units based on reptile assemblage structure. Using the predicted presences and absences of 119 reptile species as modelled by MaxEnt in previous chapters, I characterised a range of meaningful, spatially segregated biogeographic units within KNP at differing hierarchies. This included two sub-regions, five districts, and nine assemblages. These delineations appeared to make biological sense at broad scales, but at finer scales they became less reliable. This was evident via my attempt at identifying indicator species for each unit, where in some instances the species identified as being indicators represented more than one assemblage or unit, which should not be the case (Dufrene and Legendre 1997). Whilst the protocol was carried out as intended, the outputs were not always completely reliable.

My work here provided a novel view of reptile assemblage structure and distributional patterns within and across KNP. Whilst it was previously evident that several reptile species only occur within certain regions of KNP (Pienaar 1978; Bates et al. 2014), I was able to show here that there are clear patterns of spatial segregation across the landscape of the park. Despite equally high levels of species diversity and a large degree of overlap of species compositions between them, the northern and southern sub-regions of KNP were vastly dissimilar in their respective reptile species compositions. The same was true at finer divisions, with the various districts and assemblages also having large amounts of overlap in species present but with higher levels of endemism. Whilst indicator species were intended to represent each of these units at each hierarchical level, the identified indicators did not make biological sense, particularly at the assemblage level. Instead, each of these units may better be characterized by rare or endemic species.

The patterns I observed within these delineated biogeographic units match what I observed when predicting reptile species richness across KNP. As seen in chapter 3, reptile species richness was predicted as being highest within the northern sandveld regions of the park. Here, I found that the northern most region of KNP was indeed spatially segregated into its own assemblage (i.e. the northern assemblage) that had high levels of predicted species richness and endemism. Species richness thus appeared to be linked to segregations within biogeographic units, which makes biological sense as these segregations matches up with changes in the landscape of KNP. Additionally, I found that species richness and diversity significantly differed amongst assemblages, which is what would be expected given the heterogeneity of landscape features across KNP.

Similar studies which have identified biogeographic units within specific areas have been largely successful within the literature. Within southern Africa, a similar approach to the protocol I employed here was used to delineate southern Africa into spatially segregated units

based on frog species distributions (Minter et al. 2004). In their study they operated at a much broader spatial scale and used larger sized grid cells (15 km by 15 km). This approach proved successful as they were able to identify and define several frog assemblages across southern Africa, albeit with some limitations since in their case they did not use predicted presences or absences of species but instead worked directly with actual frog occurrence data. As a consequence, there were gaps in their results, but these were attributed to sampling biases rather than poor predictions which were likely to have been the cause of the issues present here.

One recurring problem throughout this study here has been that of spatial scale and the sizes of my grid cells, which has proven to be challenging with regards to the geographical extent of KNP. Finding the optimal size grid cells in which to operate remains an issue as this decision is inherently linked to predicting the presences and absences of reptiles at specific locations, which forms the basis of the delineation protocol. As a result, the biogeographic clusters I produced here may not perfectly represent reptile assemblages within reality at such a fine spatial scale, however, these could be reasonably accurate at a broader resolution. The various biogeographic units I defined appeared to make biological and geographical sense, but due to inconsistencies with my model predictions there remains an element of uncertainty within these predictions.

For future attempts at quantifying reptile assemblage structure within KNP I would recommend several adjustments to the protocol applied here. For many species of reptile, a spatial scale of 1 km x 1 km is too coarse to accurately represent their distributions. Therefore, should additional occurrence data become available I would advise that finer grid cells of 250 m x 250 m be used. However, should no significant additions to the numbers of reptile occurrences become available I would then suggest performing analyses at a broader scale. This will ensure that there are fewer errors in predictions and that reptile

presence/absence matrices are more accurate (Moura et al. 2017), thus limiting the effects of biases within input data and the cascading effects thereof. Ideally, these matrices would be based on actual occurrences rather than predictions so as to obtain better insight into real world patterns but in the absence of sufficient data, adjustments to spatial scale would provide one means of an alternative solution.

In classifying reptile assemblages within KNP, I have successfully provided a proof of concept and protocol for determining biogeographic patterns of reptile distributions across the park. I am confident that at broad scales the patterns produced here are representative of real-world reptile assemblage patterns across KNP. At finer scales, individual grid cells may have been incorrectly labelled as belonging to certain assemblage groupings and requires empirical testing. Overall, the patterns and hypotheses produced here offers a foundation for future studies.

Chapter 6: Main Conclusions

6.1 Data collection biases and limitations

In collating and quantifying available reptile locality data for KNP, I learnt several aspects relating to the inherent bias present within this dataset. For species such as reptiles which generally have low detectability, not only are records of their occurrences within KNP limited, they are also subject to multiple, unavoidable sampling biases. These biases included 1) taxonomical bias in that representation of taxa was not evenly spread amongst occurrence records, with some species and families having significantly more representation than others, 2) geographical bias in that most of KNP was data deficient with the majority of 1 km x 1 km grid cells lacking even a single occurrence record, and 3) spatial bias in that most occurrence records were strongly associated with the presence of human infrastructure.

The presence of these biases stemmed from the inherent limitations of data collection in KNP. The majority of occurrence records were obtained from museum collection databases and we know from previous studies that museum data are generally limited and do not adequately reflect on species' true occurrences within a given area. This was indeed the case here. Additionally, several occurrence records were lacking in information and did not contain accurate GPS co-ordinates. Instead, several of these merely contained centroid positions or locality descriptions without latitude or longitude positions and required estimates of their true positions. These uncertainties further added to a general lack of accurate occurrence information and compounded on the inherent biases already present. To alleviate some of these biases, sampling should be undertaken within data deficient areas as well as those where reptile species richness is predicted to be high but available data is lacking (Figure 6.1).

These biases highlight several gaps within current understanding of fine-scale reptile species distributions across the park. Moreover, these gaps present challenges towards monitoring the statuses of these species within the context of TPCs that requires this missing information for maximum efficiency. The recommended approach of a predictive framework offers a promising potential solution but may not be feasible for every species given the data limitations.

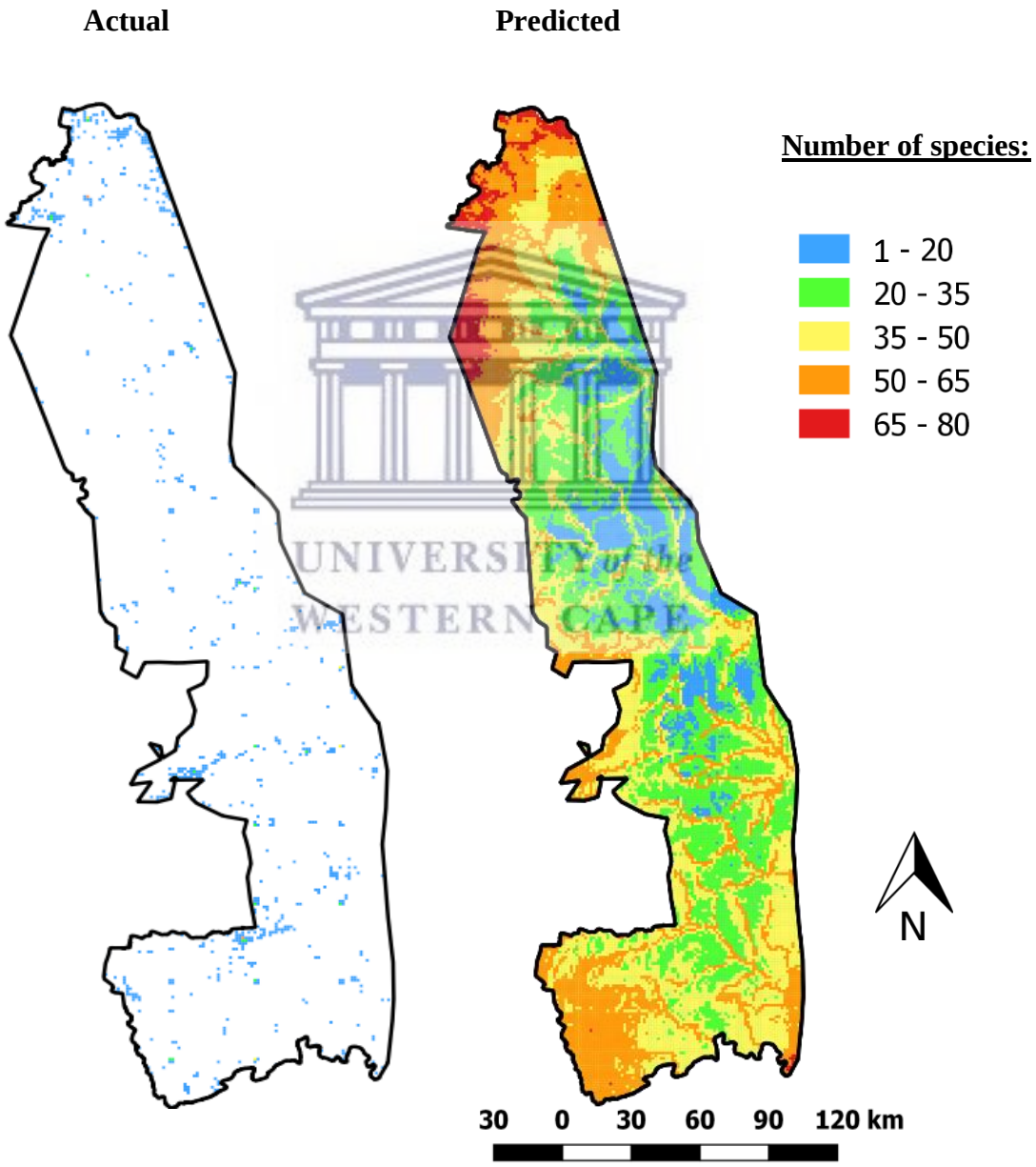


Figure 6.1: Actual detected reptile species richness vs predicted reptile species richness across KNP within 1 km x 1 km grid cells. White spaces represent grid cells with no data.

6.2 SDM and ground-truthing

Predicting distributions of reptiles at a spatial scale of 1 km x 1 km within as large an area as KNP produced variable and unreliable models. For most species, the distribution models I produced performed well in terms of their predictive strength via AUC scores, but several performed noticeably poorly. The inherent biases present within input reptile occurrence data and issues regarding the spatial scale at which I produced models severely affected the predicted ranges of these species. A clear mismatch between model performance and spatial scale was present, resulting in models performing poorly at fine spatial resolutions. However, AUC scores were not always a good indicator of model strength as several models appeared to make sense despite obtaining less than ideal scores. To achieve more reliable models, it is clear that additional occurrence data are needed for several species. This was a major limiting factor here, to the extent that eight species were excluded from modelling due to insufficient data. Nevertheless, despite variable model performance and limited occurrence data I was able to depict distributions of 119 reptile species across KNP thereby providing testable hypotheses.

Ground-truthing of my model predictions showed that within the context of the real-world, my predictions were accurate to some degree but were lacking overall. Whilst limited, my on-site sampling efforts within the greater Skukuza area of KNP revealed several flaws within some of my model predictions in the form of falsely predicting species absences. Nearly half of all tested models contained omission errors as several reptiles were observed in areas where they were not predicted to occur, thus suggesting that these models were flawed. Overall, my models require refinement in terms of spatial scale selection and predictor variables to obtain better prediction success and predictive strength. Predicting at a broader spatial scale, along with including species specific environmental variables, could produce more reliable models that contain fewer omission errors and greater prediction success.

6.3 Reptile biogeography and assemblage structure

Defining reptile biogeography and delineating KNP into spatially segregated biogeographic units based on their distributions offers valuable insight into spatial ecology and structure within the park. Whilst my underlying matrix of reptile species presences and absences across KNP was doubtful in its accuracy, I nevertheless identified meaningful patterns of reptile clusters across the landscape of KNP with clear distinctions being present between biogeographic units. My groupings remain to be empirically tested, but they appeared to make biological sense as several species associated with specific environmental conditions were associated with groupings occurring in those habitats. For example, aquatic species were grouped together within biogeographical units associated with riverine areas. As a concept these biogeographic groupings have provided evidence that it is indeed possible to identify reptile assemblages across KNP. The issue of spatial scale again played a major limiting role here, as at broad scales my delineations appeared reasonable but were less clear at finer scales. This was also clearly observable within identified indicator species which did not appear to make sense, suggesting that refinement to the selection process is required.

6.4 In the context of TPCs

Monitoring changes in reptile distributions within KNP as a measure of species performance remains challenging. On-site sampling is unlikely to yield complete inventories of reptile species present within a given area due to the difficulties associated with these animals in terms of detection. As recommended, a predictive framework could potentially offer an alternative approach to such sampling, but as demonstrated here in chapters 3 and 4, it is currently unfeasible. Simply put, given the large area of KNP and the extremely fine spatial scale at which predictions are required to be produced, currently available data are insufficient. However, predictive modelling remains a promising solution as a measure of

quantifying reptile species distributions for use within TPC monitoring, but resolutions to the issues I faced here are first required before such an approach is implemented within the SAM strategy. Ideally, on-site sampling and predictive measures should be used in conjunction with each other to measure changes in reptile populations as thoroughly as possible, but this will only be an option once predictions become more reliable.

6.5 Take-home messages

Using a predictive approach to estimate reptile distributions across KNP offers a viable alternative to on-site sampling as a mechanism for monitoring TPCs and remains promising, but it is not without its challenges. Here, I have successfully laid the foundations of a protocol for such a predictive framework. Whilst viable, this process was hindered by unresolved issues relating to spatial scale, as well as biases and limitations in species occurrence data. A mismatch between real-world reptile communities assembling at fine-scales but predictive models working best at broad scales highlighted the need for a refinement of the process as well as additional data. Whilst progress was made, additional attempts are required before implementing this framework into the current monitoring system employed by SANParks.

Importantly, several refinements to this approach can be made. Limitations in species occurrence data is a major constraint towards producing strong performing distribution models at a fine spatial scale across such a large area. While there is little to be done about a lack of data, apart from extensively sampling, several avenues remain available towards improving model performance. Different approaches towards predictor variables should be encouraged, with a particular focus being on attempting to produce models on a species-by-species basis. Alternative climate and environmental variables could offer more towards models than the variables I employed here. For instance, a variable representing solar radiation may offer advantages to those representing temperature.

6.6 Recommendations

Future attempts at predicting reptile distributions across KNP should be performed at a spatial scale more in line with available data. The finest possible spatial resolution should produce the most accurate results, but this is only possible if occurrence data supplements this option. Ideally, I would recommend a spatial scale of 250 m x 250 m, but in the absence of sufficient occurrence data further attempts would best be carried out at a broader scale to ensure meaningful results. Additionally, modelling parameters should operate on a species by species basis as opposed to here where each species' distribution was modelled identically. In addition, empirical testing and ground-truthing should aim to be less restricted (for example, limited number of sites) and incorporate additional sampling techniques such as artificial cover board arrays to enhance reptile capture success.

References

Anadón JD, Giménez A, Graciá E, Pérez I, Ferrández M, Fahd S, Mouden HE, Kalboussi M, Jdeidi T, Larbes S, Rouag, R, Slimani T, Znari M, Fritz U (2012). Distribution of *Testudo graeca* in the western Mediterranean according to climatic factors. *Amphibia-Reptilia* 33(2): 285 – 296.

Atkinson PM, Tate NJ (2000). Spatial scale problems and geostatistical solutions: a review. *The Professional Geographer* 52(4): 607 – 623.

Baldwin RA (2009). Use of maximum entropy modelling in wildlife research. *Entropy* 11(4): 854 – 866.

Bahn V, McGill BJ (2013). Testing the predictive performance of distribution models. *Oikos*, 122(3): 321 – 331.

Barbosa AM, Fontaneto D, Marini L, Pautasso M (2010). Is the human population a large-scale indicator of the species richness of ground beetles? *Animal Conservation* 13: 432 – 441.

Barlow A, Baker K, Hendry CR, Peppin L, Phelps T, Tolley KA, Wuester CE, Wuester W (2013). Phylogeography of the widespread African puff adder (*Bitis arietans*) reveals multiple Pleistocene refugia in southern Africa. *Molecular Ecology* 22: 1134 – 1157.

Bates MF, Branch WR, Bauer AM, Burger M, Marias J, Alexander GJ, De Villiers MS (2014). *Surricata 1: Atlas and red list of the reptiles of South Africa, Lesotho, and Swaziland*. South Africa National Biodiversity Institute, Pretoria. 504 p.

Böhm M, Collen B, Baillie JE, Bowles P, Chanson J, Cox N, et al. (2013). The conservation status of the world's reptiles. *Biological Conservation* 157: 372 – 385.

Botts EA, Erasmus BFN, Alexander GJ (2011). Geographic sampling bias in the South African frog atlas project: implications for conservation planning. *Biodiversity conservation* 20: 119 – 139.

Branch WR (1998). *Field guide to snakes and other reptiles of southern Africa*. Struik, Cape Town. 399 p.

Brown JL, Cameron A, Yoder AD, Vences M (2014). A necessarily complex model to explain the biogeography of the amphibians and reptiles of Madagascar. *Nature Communications* 5: 1 – 10.

Burgman MA, Fox JC (2002). Bias in species range estimations from minimum convex polygons: implications for conservation and options for improved planning. Animal Conservation 6(1): 19 – 28.

Butchart SH, Walpole M, Collen B, Van Strien A, Scharlemann JP, Almond RE, Bailli, JE, Bomhard B, Brown C, Bruno J, Carpenter KE (2010). Global biodiversity: indicators of recent declines. *Science* 2010: 1187512.

Campbell HW, Christman SP (1982). Field techniques for herpetofaunal community analysis. *Herpetological Communities* 1: 193 – 200.

Cardinale BJ, Duffy JE, Gonzalez A, Hooper DU, Perrings C, Venail P, et al. (2012). Biodiversity loss and its impact on humanity. *Nature* 486(7401): 59.

Carstensen DW, Lessard JP, Holt BG, Borregaard MK, Rahbek C (2013). Introducing the biogeographic species pool. *Ecography* 36: 1310 – 1318.

Chape S, Harrison J, Spalding M, Lysenko I (2005). Measuring the extent and effectiveness of protected areas as an indicator for meeting global biodiversity targets.

Philosophical Transactions of the Royal Society of London B: Biological Sciences 360(1454): 443 – 455.

Demsar U, Harris P, Brunsdon C, Fotheringham AS, McLoone S (2013). Principal component analysis on spatial data: an overview. *Annals of the Association of American Geographers* 103(1): 106 – 128.

Du Toit JT, Biggs HC, Rogers KH (2003). *The Kruger Experience: Ecology and management of savanna heterogeneity*. Island Press, Washington. 536 p.

Dufrêne M, Legendre P (1997). Species assemblages and indicator species: the need for a flexible asymmetrical approach. *Ecological Monographs* 67(3): 345 – 366.

Durso AM, Wilson JD, Winne CT (2011). Needles in haystacks: estimating detection probability and occupancy of rare and cryptic snakes. *Biological Conservation* 144: 1508 – 1515.

Elith J, Graham CH, Anderson RP, Dudík M, Ferrier S, Guisan A, Hijmans RJ, Huettmann F, Leathwick JR, Lehmann A, Li J, Lohmann L, Loiselle B, Manion G, Moritz C, Nakamura M, Nakazawa Y, Overton JMC, Peterson AT, Phillips SJ, Richardson KS, Scachetti-Pereira R, Schapire RE, Soberón J, Williams S, Wisz MS, Zimmermann NE (2006). Novel methods to improve prediction of species' distributions from occurrence data. *Ecography* 29: 129 – 151.

Everitt B (1993*). Cluster analysis: third edition*. Heinemann Educational Publishers, New York. 346 p.

Everitt B, Hothorn T (2011). *An introduction to applied multivariate analysis with R*. Springer Science and Business Media, New York. 274 p.

Feria TPA, Peterson AT (2002). Prediction of bird community composition based on point-occurrence data and inferential algorithms: a valuable tool in biodiversity assessments. *Diversity and Distributions* 8(2): 49 – 56.

Ferreira S, Deacon A, Sithole H, Bezuidenhout H, Daemane M, Herbst M (2011). From numbers to ecosystems and biodiversity: A mechanistic approach to monitoring. *Koedoe* 53(2): 1 – 12.

Ferrier S, Guisan A (2006). Spatial modelling of biodiversity at the community level. *Journal of Applied Ecology* 43(3) 393 – 404.

Franklin J (2010). *Mapping species distributions: spatial inference and prediction*. Cambridge University Press, Cambridge. 320 p.

Galante PJ, Alade B, Muscarella R, Jansa SA, Goodman SM, Anderson RP (2018). The challenge of modelling niches and distributions for data-poor species: a comprehensive approach to model complexity. *Ecography* 41(5): 726 – 736.

Galili T (2015). Dendextend: An R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* 31(22): 3718 – 3720.

Gaston KJ, He F (2011). *Biological Diversity: Frontiers in measurement and assessment*. Oxford University Press, Oxford. 368 p.

Gillson L, Duffin KI (2007). Thresholds of potential concern as benchmarks in the management of African savannahs. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 362(1478): 309 – 319.

Greene BD, Dixon JR, Whiting MJ, Mueller JM (1999). Reproductive ecology of the Concho water snake, *Nerodia harteri paucimaculata*. *Copeia* 1999(2): 701 – 709.

Griffith DM, Veech JA, Marsh CJ (2016). Co-occur: Probabilistic species co-occurrence analysis in R. *Journal of Statistical Software* 69(2): 1 – 17.

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

Harrison JA, Allan DG, Underhill LG, Herremans M, Tree AJ, Parker V, Brown CJ (1997). *The atlas of southern African birds – volume 1: non-passerine & volume 2: passerines*. BirdLife South Africa, Johannesburg. 1600 p.

Hernandez PA, Graham CH, Master LL, Albert DL (2006). The effect of sample size and species characteristics on performance of different species distribution modelling methods. *Ecography* 29(5): 773 – 785.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005): Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* 25 (15): 1965 – 1978.

Hijmans RJ, van Etten J (2016). Raster: Geographic analysis and modelling with raster data. R package version 2.2.3-1.

Hurlbert AH, Jetz W (2007). Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. *Proceedings of the National Academy of Sciences* 104(33): 13384 – 13389.

Jackson DA (1993). Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology* 74(8): 2204 – 2214.

Jackson SF, Gaston KJ (2008). Land use change and the dependence of national priority species on protected areas. *Global Change Biology* 14(9): 2132 – 2138.

Jetz W, Rahbek C, Colwell RK (2004). The coincidence of rarity and richness and the potential signature of history in centres of endemism. *Ecology Letters* 7(12): 1180 – 1191.

Jetz W, Freckleton RP (2015). Towards a general framework for predicting threat status of data-deficient species from phylogenetic, spatial and environmental information. *Philosophical Transactions of the Royal Society B* 370(1662): 20140016.

Kadmon R, Farber O, Danin A (2004). Effects of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications* 14(2): 401 – 413.

Kaufman L, Rousseeuw PJ (1990). *Finding groups in data: an introduction to cluster analysis*. John Wiley and Sons, New York. 342 p.

Kent M (2006). Numerical classification and ordination methods in biogeography. *Progress in Physical Geography* 30(3): 399 – 408.

Kéry M (2011). Towards the modelling of true species distributions. *Journal of Biogeography* 38(4): 617 – 618.

Kreft H, Jetz W (2010). A framework for delineating biogeographical regions based on species distributions. *Journal of Biogeography* 37(11): 2029 – 2053.

Kuhnz LA, Burton RK, Slattery PN, Oakden JM (2005). Microhabitats and population densities of California legless lizards, with comments on effectiveness of various techniques for estimating numbers of fossorial reptiles. *Journal of Herpetology* 39: 395 – 402.

Kulmatiski A, Sprouse SR, Beard KH (2017). Soil type more than precipitation determines fine-root abundance in savannas of Kruger National Park, South Africa. *Plant and Soil* 417(1): 523 – 533.

Lee TM, Jetz W (2010). Unravelling the structure of species extinction risk for predictive conservation science. *Proceedings of the Royal Society B:* 278(1710): 1329 – 1338.

Legendre P, Legendre L (1998). *Numerical ecology*. Elsevier, New York. 852 p.

Leutner B, Horning N (2016). RStoolbox: Tools for remote sensing data analysis**.** R package version 0.1.4.

Linder HP, de Klerk HM, Born J, Burgess ND, Fjeldså J, Rahbek C. (2012). The partitioning of Africa: statistically defined biogeographical regions in sub-Saharan Africa. *Journal of Biogeography* 39(7): 1189 – 1205.

Lobo JM, Jiménez-Valverde A, Real R (2008). AUC: a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography* 17(2): 145 –151.

Mackey BG, Berry SL, Brown T (2008). Reconciling approaches to biogeographic regionalization: a systematic and generic framework examined with a case study of the Australian continent. Journal of Biogeography 35: 213 – 229.

MacQueen J (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1(14): 281 – 297.

Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik, K (2016). Cluster: cluster analysis basics and extensions. R package version 2.9.7-1.

Maritz B, Masterson G, Mackay D, Alexander G (2007). The effect of funnel trap type and size of pitfall trap on trap success: implications for ecological field studies. *Amphibia-Reptilia* 28(3): 321 – 328.

Martin J, Runge M.C, Nichols J.D, Lubow BC, Kendall WL (2009). Structured decision making as a conceptual framework to identify thresholds for conservation and management. *Ecological Applications* 19(5): 1079 – 1090.

Mazerolle MJ, Bailey LL, Kendall WL, Royle AJ, Converse SJ, Nichols JD (2007). Making great leaps forward: accounting for detectability in herpetological field studies. *Journal of Herpetology* 41(4): 672 – 689.

McCune B, Grace JB, Urban DL (2002). *Analysis of ecological communities*. MjM software design, Oregon. 300 p.

McDiarmid RW, Foster MS, Guyer C, Chernoff N, Gibbons JW (2012). *Reptile biodiversity: standard methods for inventory and monitoring*. University of California Press, California. 424 p.

McGrath T, Guillera-Arroita G, Lahoz-Monfort JJ, Osborne W, Hunter D, Sarre SD (2015). Accounting for detectability when surveying for rare or declining reptiles: turning rocks to find the grassland earless dragon in Australia. *Biological Conservation* 182: 53 – 62.

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

Minchin PR (1987). An evaluation of the relative robustness of techniques for ecological ordination. *Theory and Models in Vegetation Science* 69: 89 – 107.

Minter LR, Burger M, Harrison JA, Braack HH, Bishop PJ, Kloepfer D (2004). *Atlas and red data book of the frogs of South Africa, Lesotho and Swaziland*. Smithsonian Institution, Washington. 410 p.

Morrone JJ (2009). *Evolutionary biogeography: an integrative approach with case studies*. Columbia University Press, New York. 304 p.

Moura MR, Argôlo AJ, Costa HC (2017). Historical and contemporary correlates of snake biogeographical subregions in the Atlantic Forest hotspot. *Journal of Biogeography* 44(3): 640 – 650.

Myers N, Mittermeier RA, Mittermeier CG, Da Fonseca GA, Kent J (2000). Biodiversity hotspots for conservation priorities. *Nature* 403(6772): 853.

Newbold T (2010). Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. *Progress in Physical Geography* 34(1): 3 – 22.

Oksanen J, Kindt R, Legendre P, O'Hara B, Stevens MHH, Oksanen MJ (2007). The vegan package. *Community Ecology Package* 10: 631 – 637.

Overton JM, Stephens RT, Leathwick JR, Lehmann A (2002). Information pyramids for informed biodiversity conservation. *Biodiversity Conservation* 11(12): 2093 – 2116.

Parr CL, Woinarski JCZ, Pienaar DJ (2009). Cornerstones of biodiversity conservation? Comparing the management effectiveness of Kruger and Kakadu National Parks, two key savanna reserves. *Biodiversity Conservation* 18(1): 3643 – 3662.

Pearce J, Ferrier S (2000). Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling* 133(3): 225 – 245.

Pearson RG, Raxworthy CJ, Nakamura M, and Peterson AT (2007). Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *Journal of Biogeography* 34(1): 102 – 117.

Phillips SJ, Anderson RP, Schapire RE (2006). Maximum entropy modelling of species geographic distributions. *Ecological Modelling* 190(3): 231 – 259.

Phillips SJ, Dudík M (2008). Modelling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography* 31(2): 161 – 175.

Pienaar UDV (1978). *The reptile fauna of the Kruger National Park*. Sigma Press, Pretoria. 223 p.

Pineda E, Lobo JM (2009). Assessing the accuracy of species distribution models to predict amphibian species richness patterns. *Journal of Animal Ecology* 78(1): 182 – 190.

Powney GD, Grenyer R, Orme CDL, Owens IPF, Meiri S (2010). Hot, dry and different: Australian lizard richness is unlike that of mammals, amphibians and birds. *Global Ecology and Biogeography* 19(3): 386 – 396.

Price B, Kutt AS, McAlpine CA (2010). The importance of fine-scale savanna heterogeneity for reptiles and small mammals. *Biological Conservation* 143(11): 2504 – 2513.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Radosavljevic A, Anderson RP (2014). Making better MaxEnt models of species distributions: complexity, overfitting and evaluation. *Journal of Biogeography* 41(4): 629 – 643.

Rahbek C (2005). The role of spatial scale and the perception of large-scale species-richness patterns. *Ecology letters* 8(2): 224 – 239.

Raxworthy CJ, Martinez-Meyer E, Horning N, Nussbaum RA, Schneider GE, Ortega-Huerta MA, Peterson AT (2003). Predicting distributions of known and unknown reptile species in Madagascar. *Nature* 426(1): 837 – 841.

Reddy S, Davalos L (2003). Geographic sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography* 30: 1719 – 1727.

Roberts DW (2016). Labdsv: ordination and multivariate analysis for ecology. R package version 1.0.

Roux DJ, Foxcroft LC (2011). The development and application of strategic adaptive management within South African National Parks. *Koedoe* 53(2): 01 – 05.

Sahlean TC, Gherghel I, Papeş M, Strugariu A, Zamfirescu ŞR (2014). Refining climate change projections for organisms with low dispersal abilities: a case study of the Caspian whip snake. *PloS one* 9(3): e91994.

Sarda-Palomera F, Brotons L, Villero D, Sierdsema H, Newson SE, Jiguet F (2012). Mapping from heterogeneous biodiversity monitoring data sources. *Biodiversity and Conservation* 21: 2927 – 2948.

Sarquis JA, Cristaldi MA, Arzamendia V, Bellini G, Giraudo AR (2018). Species distribution models and empirical test: Comparing predictions with well-understood geographical distribution of *Bothrops alternatus* in Argentina. *Ecology and Evolution* 2018: 1 – 13.

Scholtz R, Kiker GA, Smit IPJ, Venter FJ (2014). Identifying drivers that influence the spatial distribution of woody vegetation in Kruger National Park, South Africa. *Ecosphere* 5(6): 1 – 12.

Skelton PH (2001). *A complete guide to the freshwater fishes of southern Africa*. Struik, Cape Town. 395 p.

Stockman AK, Beamer DA, Bond JE (2006). An evaluation of a GARP model as an approach to predicting the spatial distribution of non-vagile invertebrate species. *Diversity and Distributions* 12(1): 81 – 89.

Stockwell D, Peters D (1999). The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science* 13(2): 143 – 158.

Stockwell D, Peterson AT (2002). Effects of sample size on accuracy of species distribution models. *Ecological modelling* 148(1): 1 – 13.

Storch D, Konvicka M, Benes J, Martinková J, Gaston KJ (2003). Distribution patterns in butterflies and birds of the Czech Republic: separating effects of habitat and geographical position. *Journal of Biogeography* 30(8): 1195 – 1205.

Thomas L, Buckland ST, Rexstad EA, Laake JL, Strindberg S, Hedley SL, et al. (2010). Distance software: design and analysis of distance sampling surveys for estimating population size. *Journal of Applied Ecology* 47(1): 5 – 14.

Tolley KA, Makokha JS, Houniet DT, Swart BL, Matthee CA (2009). The potential for predicted climate shifts to impact genetic landscapes of lizards in the South African Cape Floristic Region. *Molecular Phylogenetics and Evolution* 51(1): 120 – 130.

Tolley KA, Alexander GJ, Branch WR, Bowles P, Maritz B (2016). Conservation status and threats for African reptiles. *Biological Conservation* 204: 63 – 71.

Venter FJ, Naiman RJ, Biggs HC, Pienaar DJ (2008). The evolution of conservation management philosophy: Science, environmental change and social adjustments in Kruger National Park. *Ecosystems* 11(2): 173 – 192.

Underhill LG, Brooks M (2016). Displaying changes in bird distributions between SABAP1 and SABAP2. *Biodiversity Observations* 1: 1 – 13.

Whittaker RJ, Araújo MB, Jepson P, Ladle RJ, Watson JE, Willis KJ (2005). Conservation biogeography: assessment and prospect. *Diversity and Distributions* 11(1): 3 – 23.

Wilson KA, Westphal MI, Possingham HP, Elith J (2005). Sensitivity of conservation planning to different approaches to using predicted species distribution data. *Biological Conservation* 122(1): 99 – 112.

Appendices

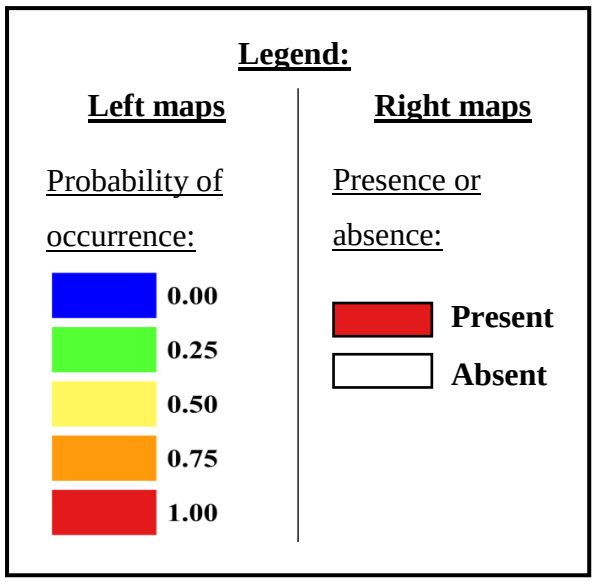
Appendix 1: List of 27 predictor variables and their sources.

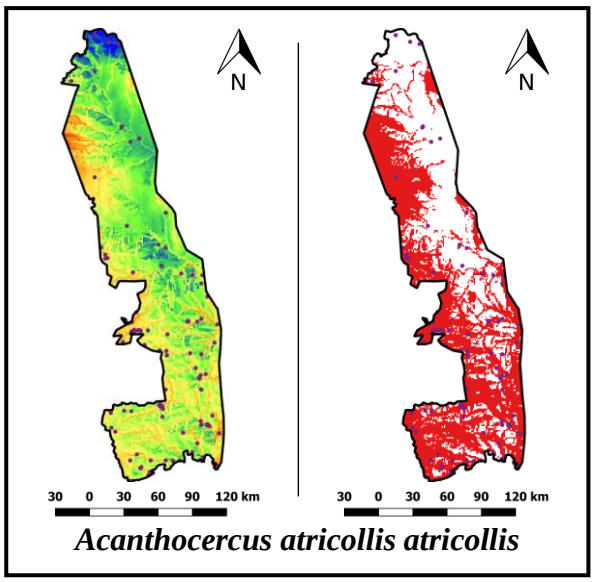
Variable	Source	Type
Altitude	Wordlclim	Continuous
Annual Mean Temperature	Wordlclim	Continuous
Annual Precipitation	Wordlclim	Continuous
Aspect	Calculated from Altitude	Continuous
Distance to Water	Calculated from Water p/a	Continuous
Infrastructure presence/absence (p/a)	SANParks	Categorical
Isothermality	Wordlclim	Continuous
Max Temperature of Warmest Month	Wordlclim	Continuous
Mean Diurnal Range	Wordlclim	Continuous
Mean Temperature of Coldest Quarter	Wordlclim	Continuous
Mean Temperature of Driest Quarter	Wordlclim	Continuous
Mean Temperature of Warmest Quarter	Wordlclim	Continuous
Mean Temperature of Wettest Quarter	Wordlclim	Continuous
Min Temperature of Coldest Month	Wordlclim	Continuous
Precipitation of Coldest Quarter	Wordlclim	Continuous
Precipitation of Driest Month	Wordlclim	Continuous
Precipitation of Driest Quarter	Wordlclim	Continuous
Precipitation of Warmest Quarter	Wordlclim	Continuous
Precipitation of Wettest Month	Wordlclim	Continuous
Precipitation of Wettest Quarter	Wordlclim	Continuous
Precipitation Seasonality	Wordlclim	Continuous
Slope	Calculated from Altitude	Continuous
Soils	SOTER	Categorical
Temperature Annual Range	Wordlclim	Continuous
Temperature Seasonality	Wordlclim	Continuous
Vegetation	SANBI	Categorical
Water presence/absence (p/a)	SANParks	Categorical

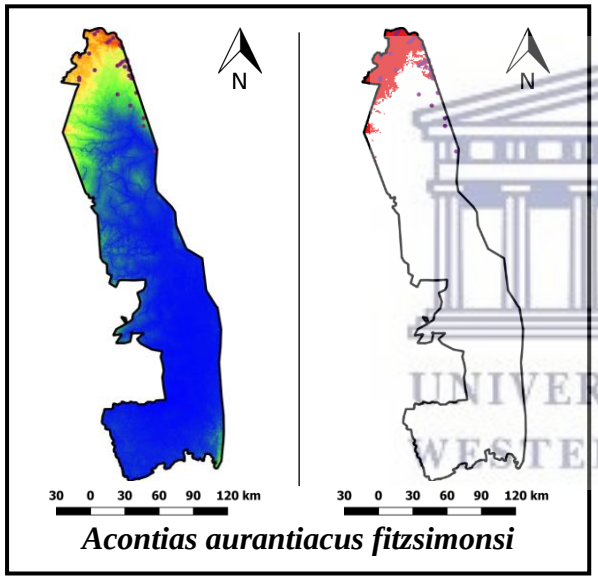
Appendix 2: Percentages of variable contributions towards compositions of six principal components.

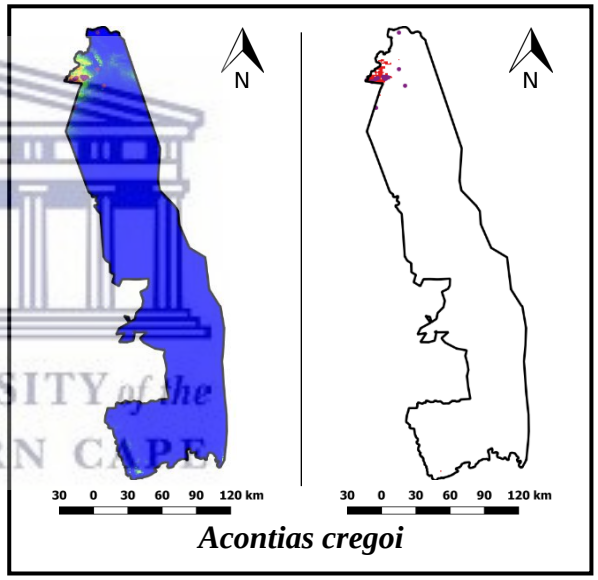
Variable	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6
Altitude	1.47	6.77	7.58	2.69	10.70	4.48
Annual Mean Temperature	5.78	0.74	4.51	1.68	2.12	2.55
Annual Precipitation	5.93	0.81	3.29	1.95	1.18	0.83
Aspect	0.22	0.38	0.68	7.94	2.70	35.84
Distance to Water	0.35	0.97	6.92	16.88	5.51	0.60
Infrastructure presence/absence (p/a)	1.27	6.07	5.68	5.26	21.99	1.71
Isothermality	0.79	8.55	4.75	2.89	0.56	0.66
Max Temperature of Warmest Month	5.78	2.21	2.19	1.17	1.83	1.59
Mean Diurnal Range	5.24	2.50	3.96	3.93	0.90	0.84
Mean Temperature of Coldest Quarter	4.73	5.07	5.80	1.21	1.81	2.51
Mean Temperature of Driest Quarter	4.73	5.07	5.74	1.21	1.81	2.51
Mean Temperature of Warmest Quarter	5.61	3.15	2.76	1.24	2.78	2.16
Mean Temperature of Wettest Quarter	5.60	3.18	2.82	1.28	2.70	2.15
Min Temperature of Coldest Month	2.46	8.34	2.34	0.52	0.88	1.87
Precipitation of Coldest Quarter	5.27	3.21	4.58	4.10	3.05	0.33
Precipitation of Driest Month	5.54	2.03	4.33	2.98	3.13	0.94
Precipitation of Driest Quarter	5.27	3.21	4.52	4.10	3.05	0.33
Precipitation of Warmest Quarter	5.41	2.88	4.06	1.47	0.95	2.47
Precipitation of Wettest Month	5.26	3.48	3.65	0.21	1.99	3.24
Precipitation of Wettest Quarter	5.41	2.88	4.06	1.47	0.95	2.47
Precipitation Seasonality	0.41	0.31	0.46	4.44	15.66	2.05
Slope	4.93	4.68	1.98	2.92	1.75	1.43
Soils	2.20	2.51	2.91	6.11	0.40	12.02
Temperature Annual Range	3.76	7.36	0.53	1.41	1.13	0.28
Temperature Seasonality	2.32	8.72	1.31	0.47	0.94	0.13
Vegetation	4.09	3.78	2.43	4.10	6.92	0.28
Water presence/absence (p/a)	0.16	1.12	6.18	16.37	2.61	13.72
∑	100	100	100	100	100	100

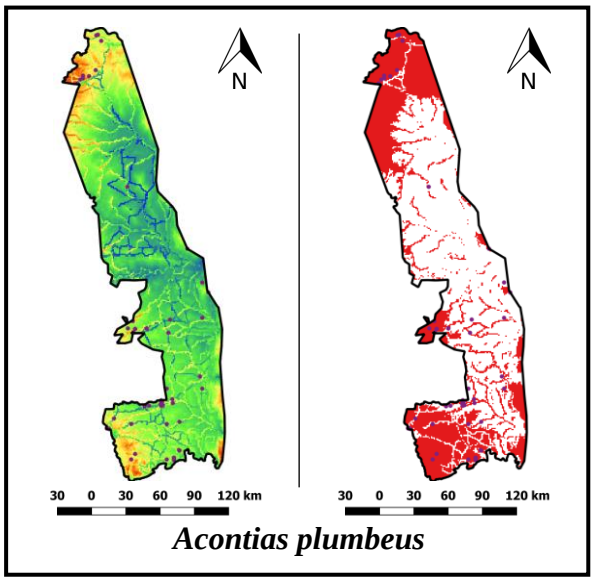
Appendix 3: Predicted distributions maps of 119 reptile species across KNP.

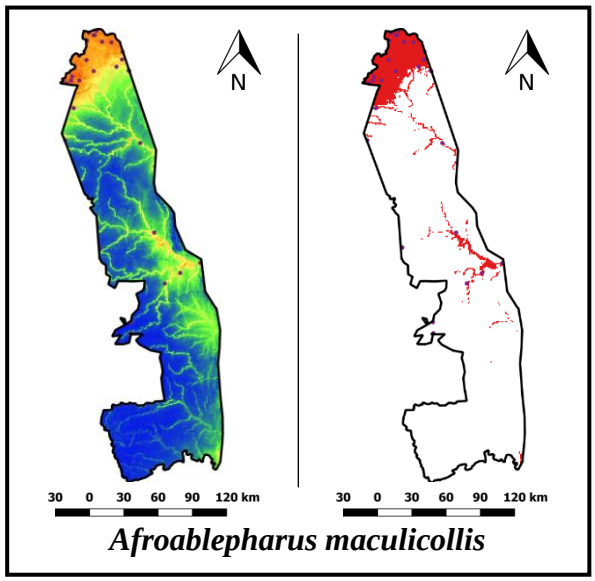


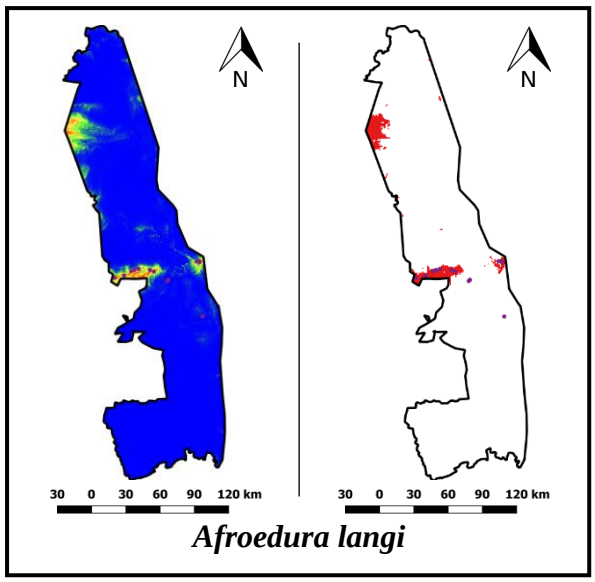




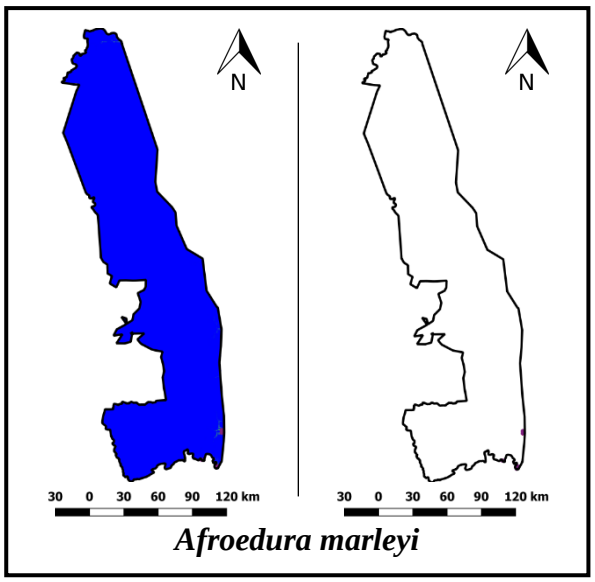




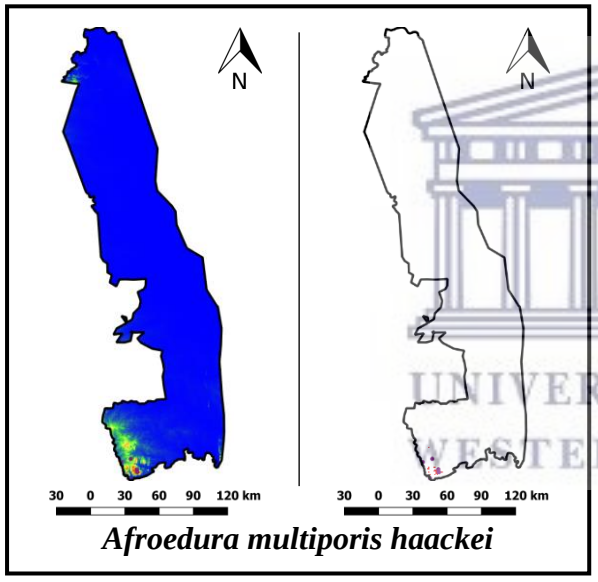




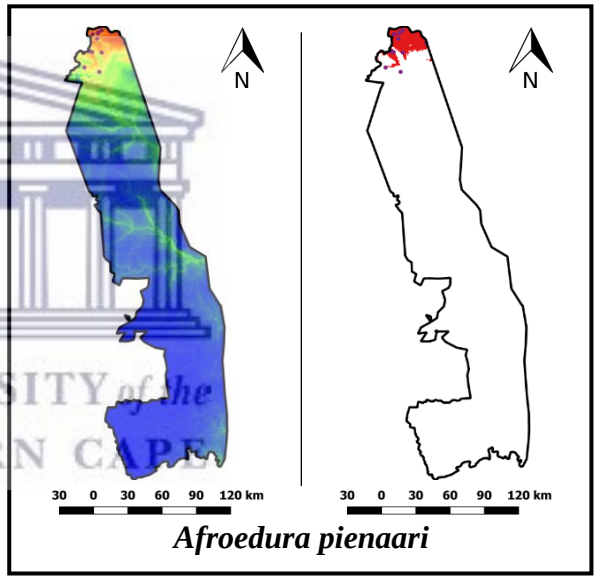
Afroedura langi



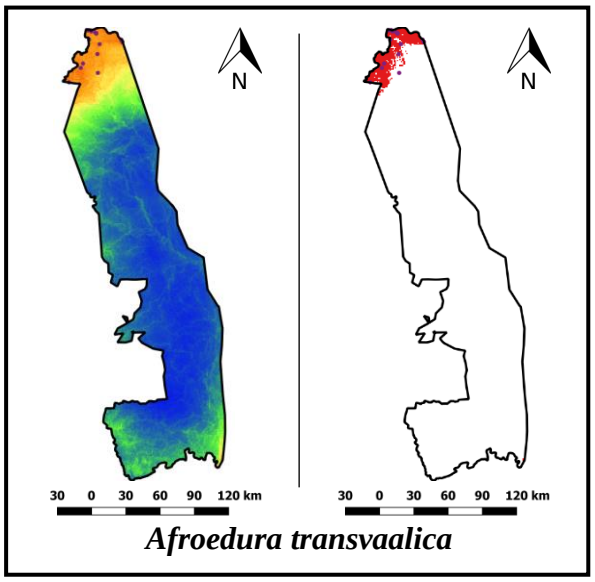
Afroedura marleyi



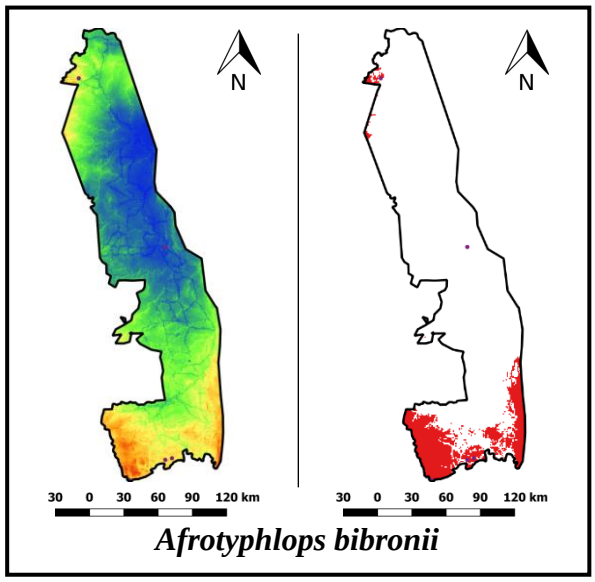
Afroedura multiporis haackei



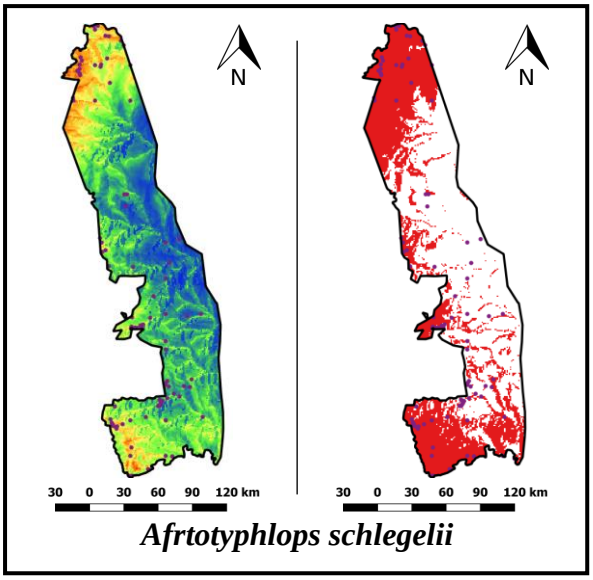
Afroedura pienaari



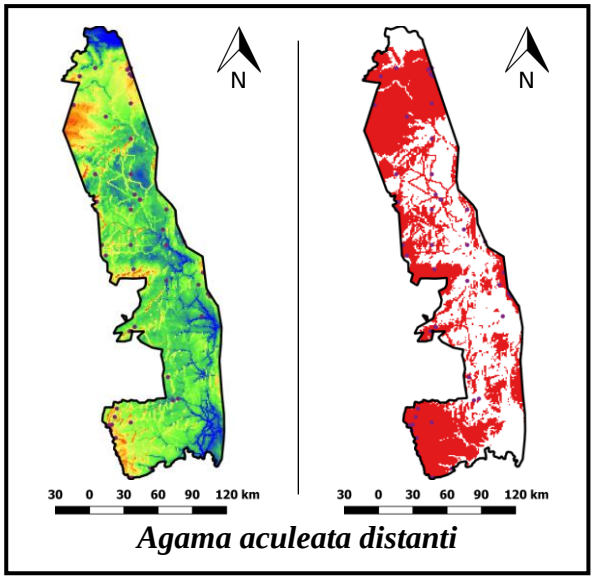
Afroedura transvaalica



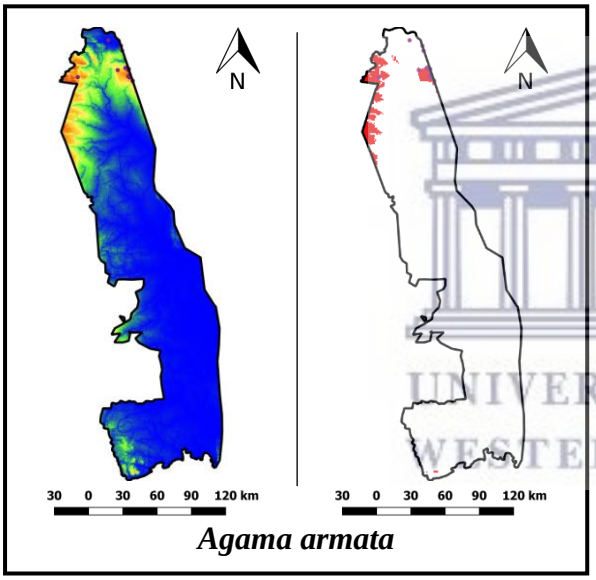
Afrotyphlops bibronii



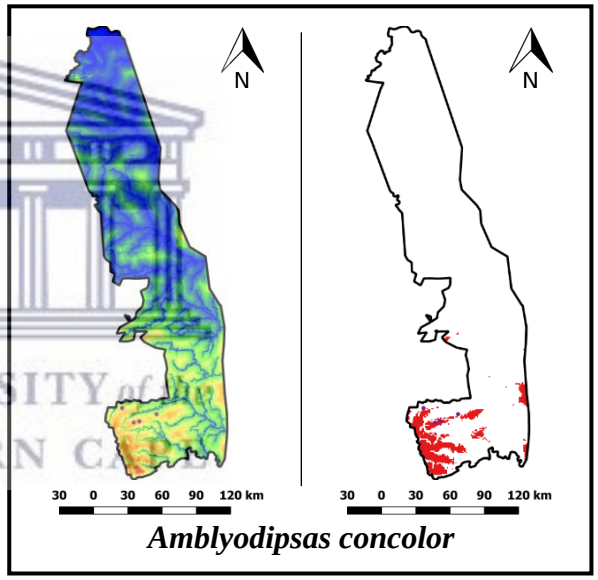
Afrtotyphlops schlegelii



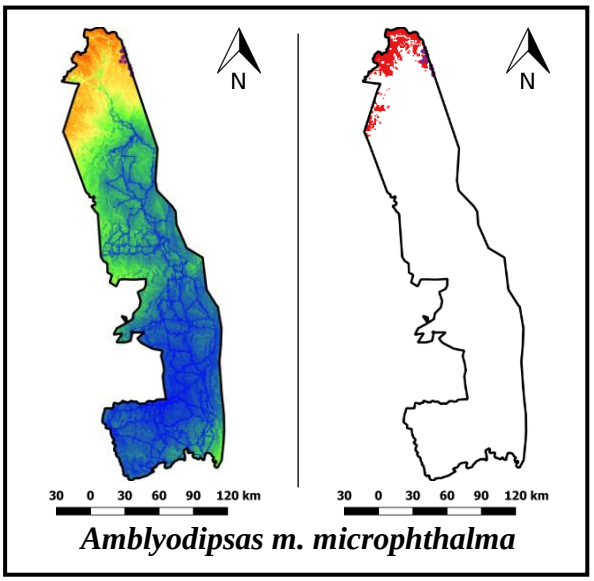
Agama aculeata distanti



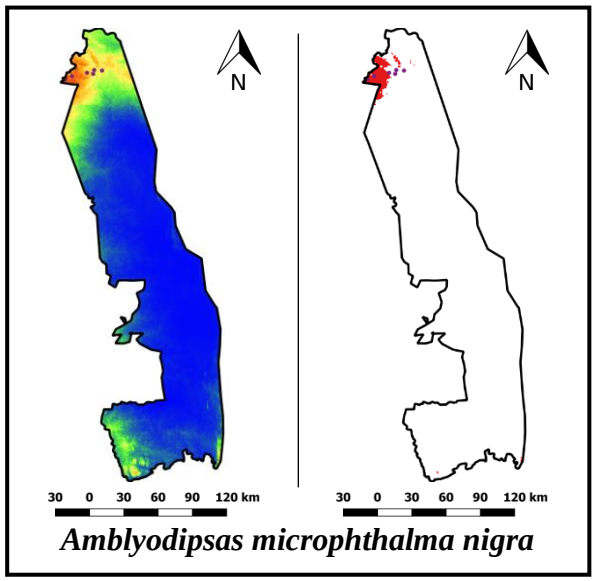
Agama armata



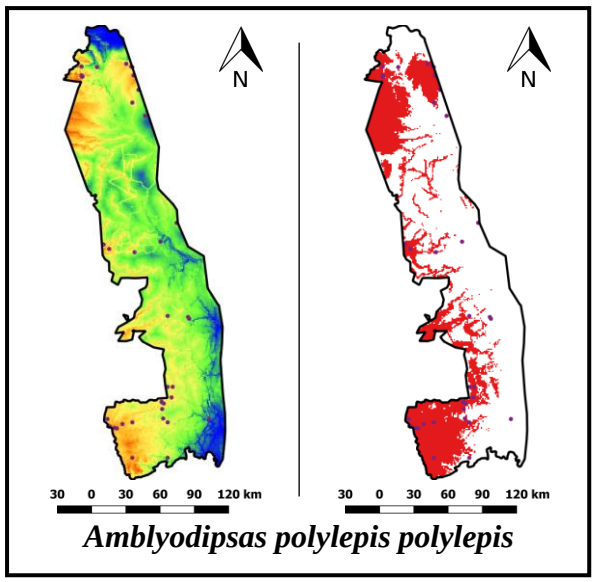
Amblyodipsas concolor



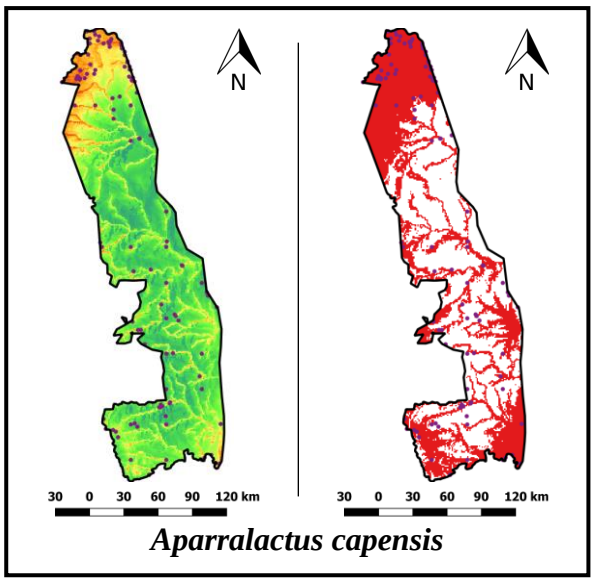
Amblyodipsas m. microphthalma



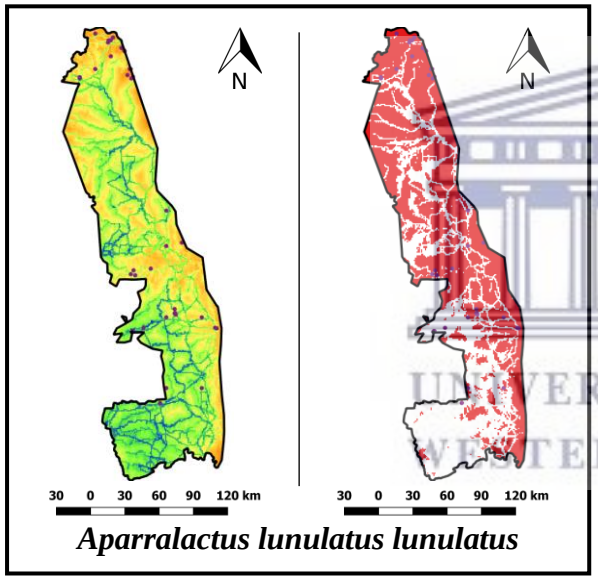
Amblyodipsas microphthalma nigra



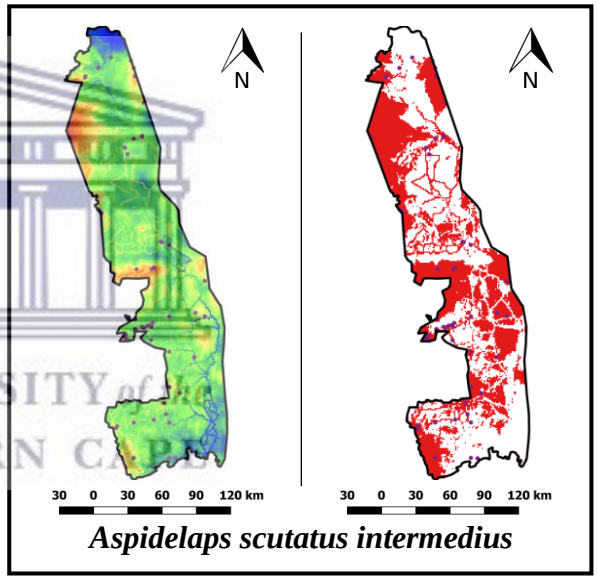
Amblyodipsas polylepis polylepis



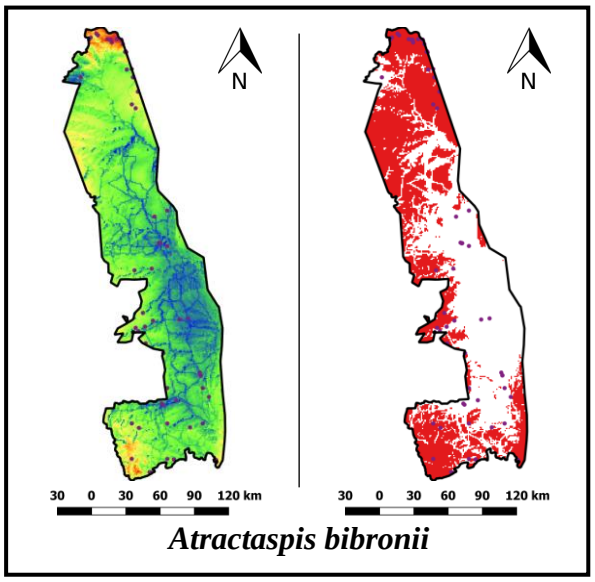
Aparralactus capensis



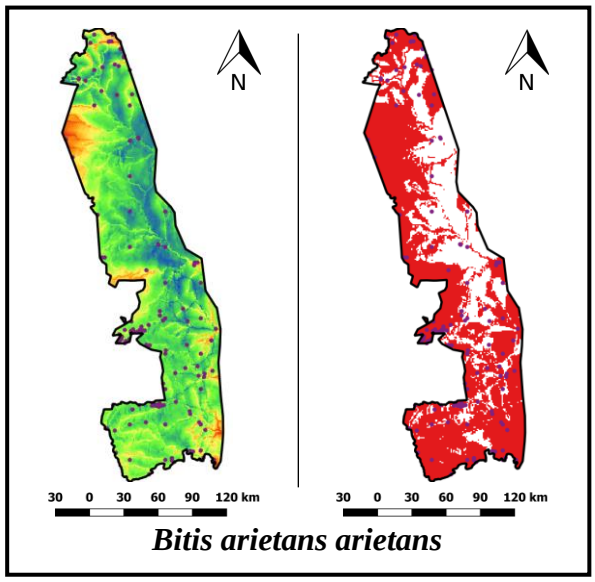
Aparralactus lunulatus lunulatus



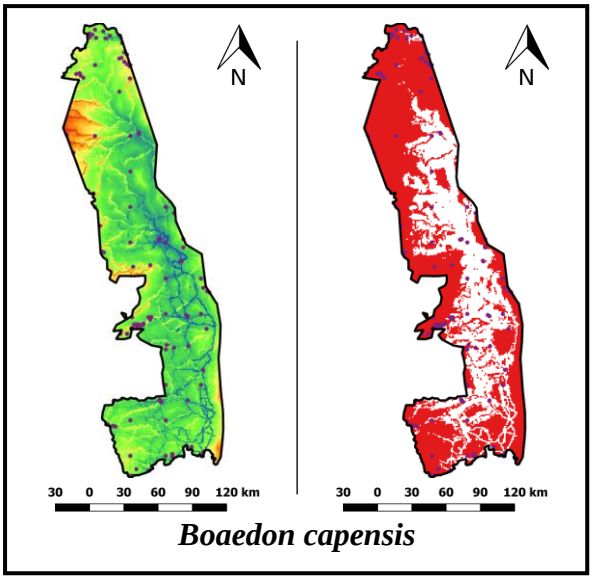
Aspidelaps scutatus intermedius



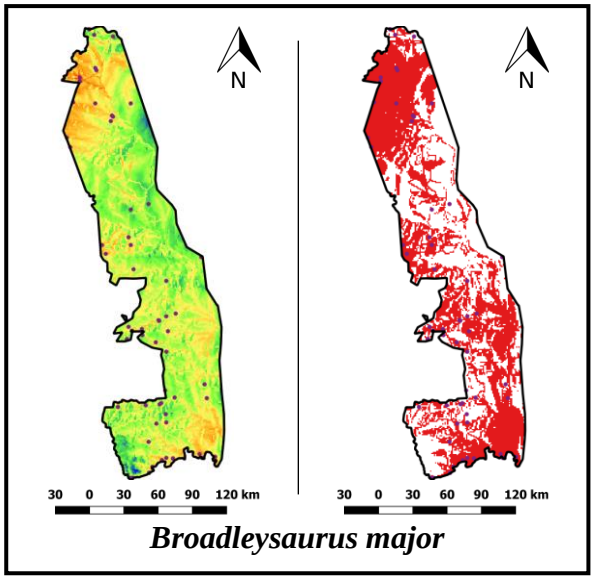
Atractaspis bibronii



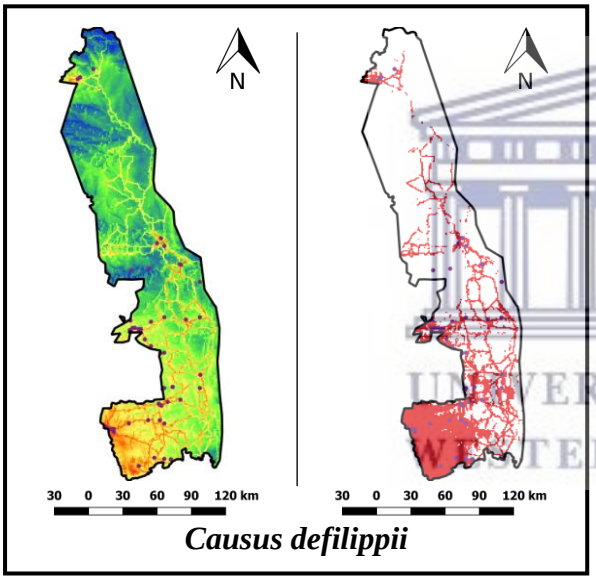
Bitis arietans arietans



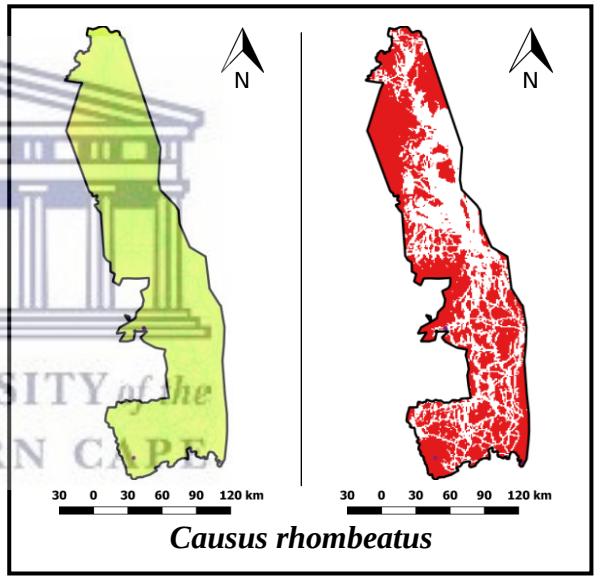
Boaedon capensis



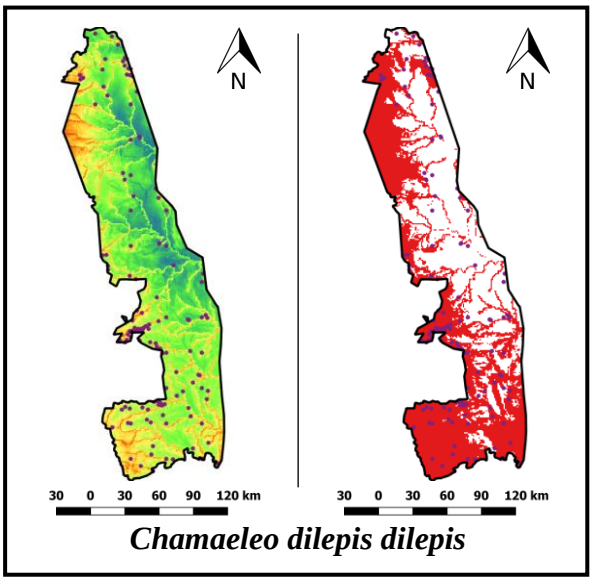
Broadleysaurus major



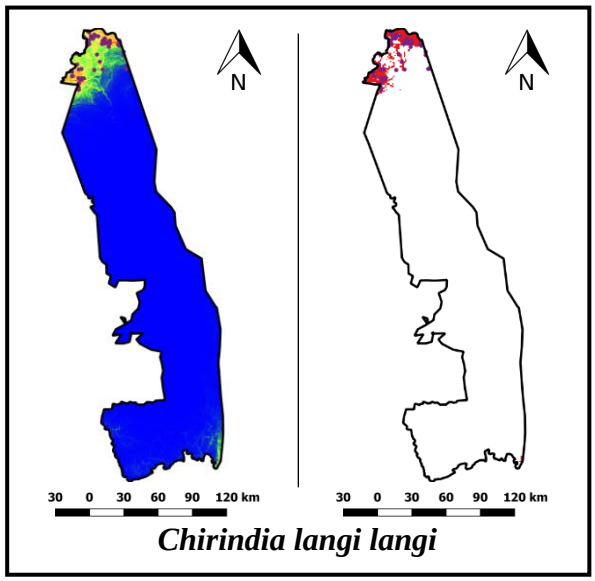
Causus defilippii



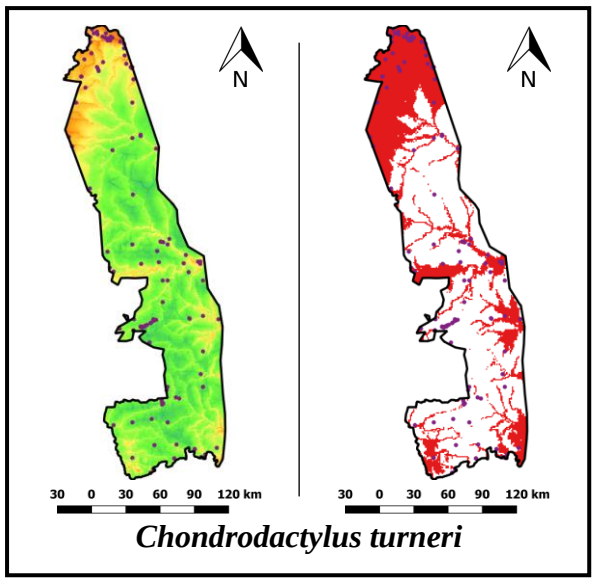
Causus rhombeatus



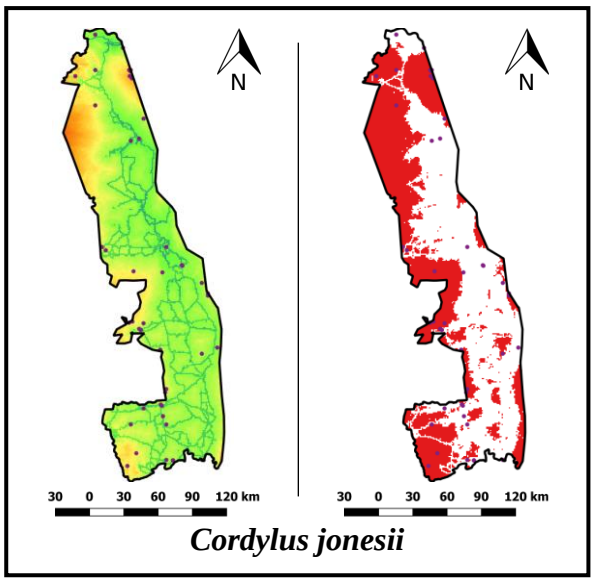
Chamaeleo dilepis dilepis



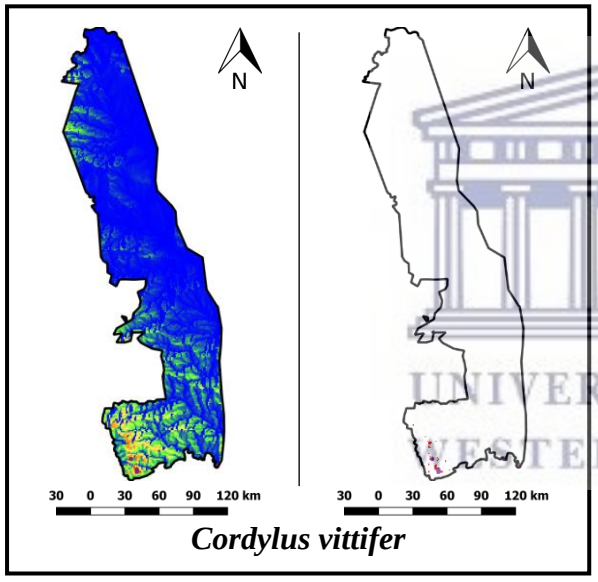
Chirindia langi langi



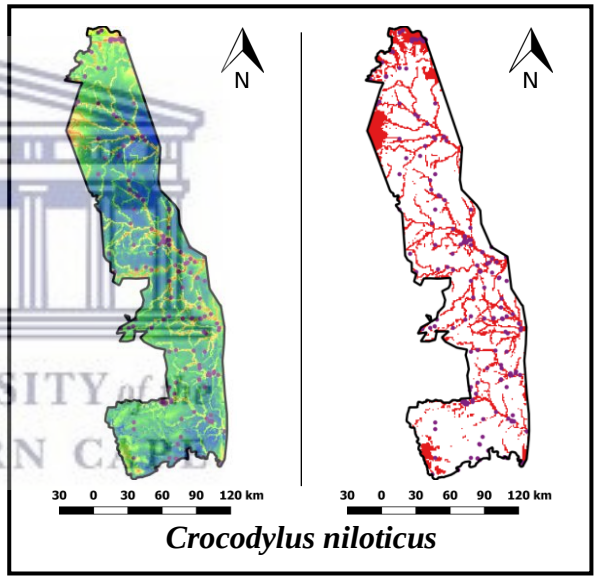
Chondrodactylus turneri



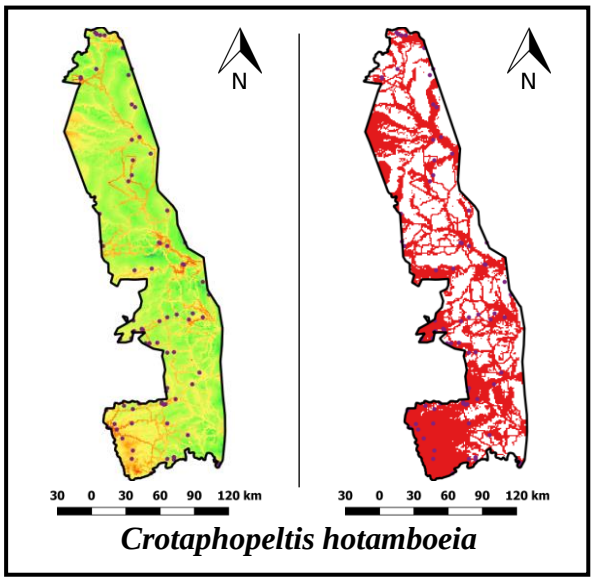
Cordylus jonesii



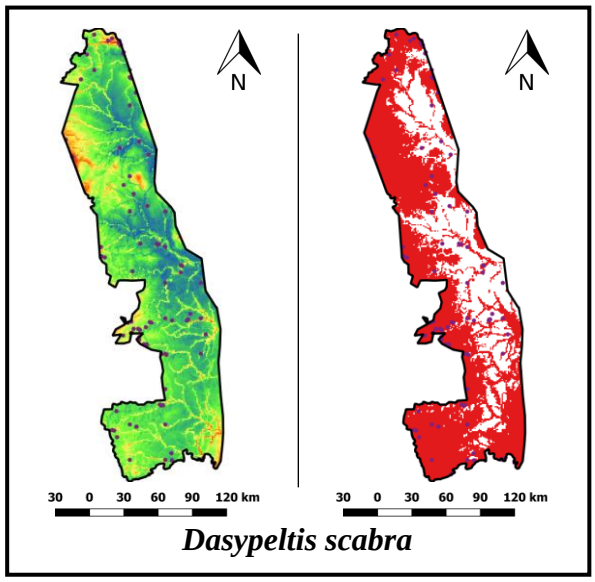
Cordylus vittifer



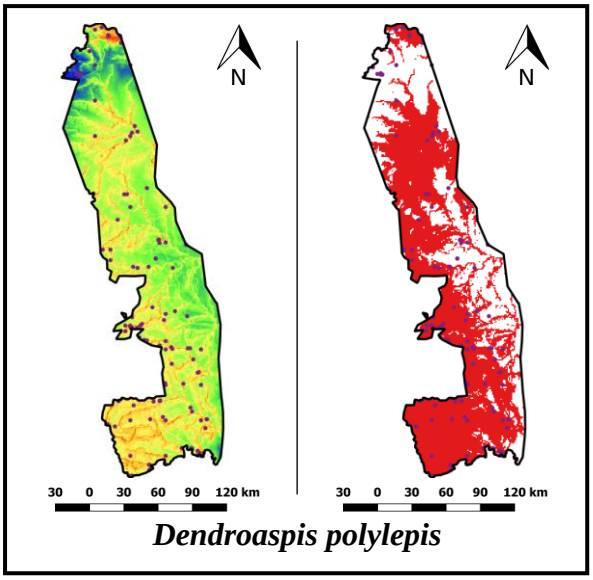
Crocodylus niloticus



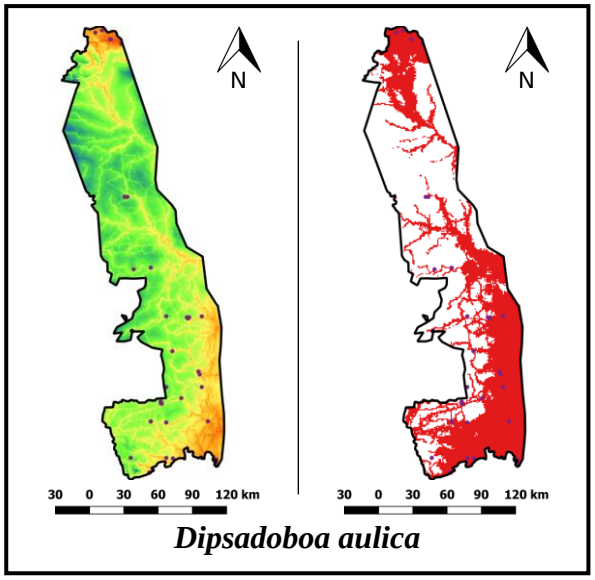
Crotaphopeltis hotamboeia



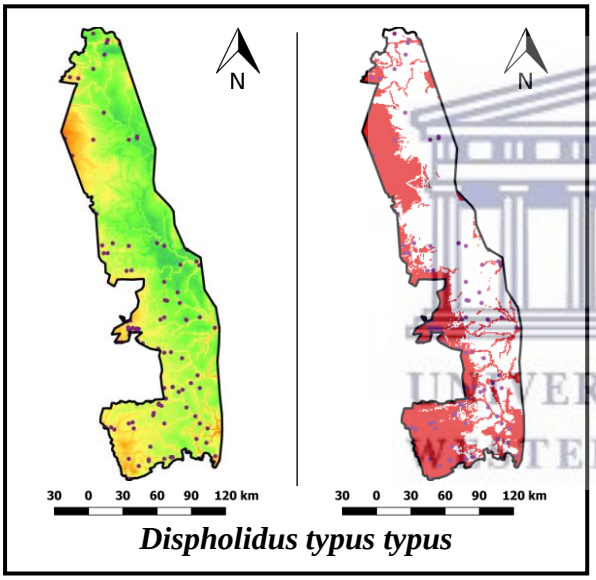
Dasypeltis scabra



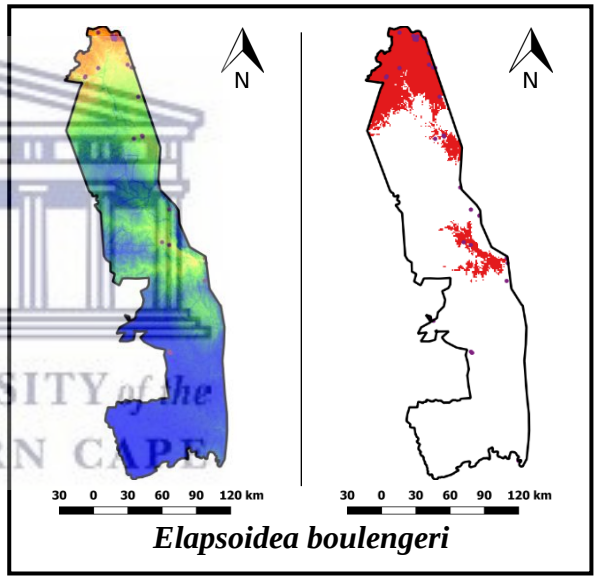
Dendroaspis polylepis



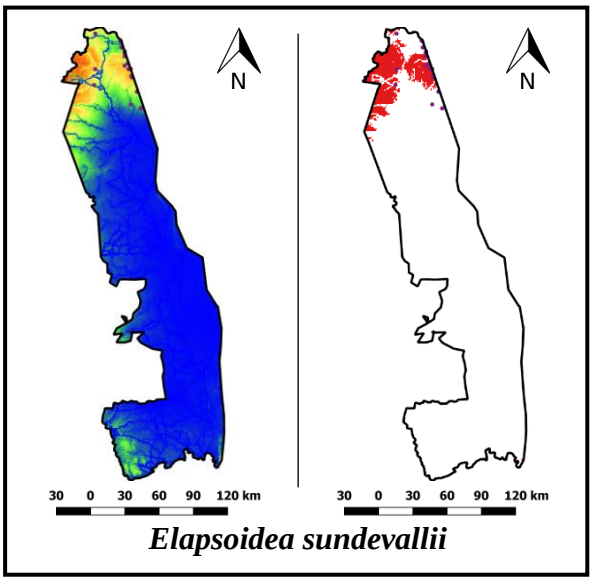
Dipsadoboa aulica



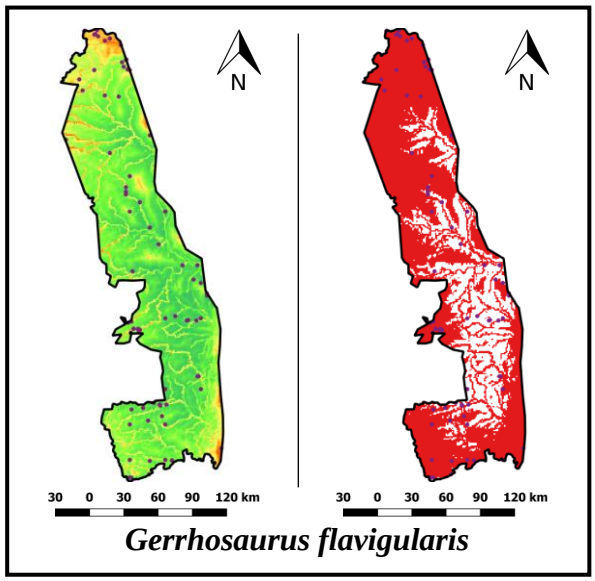
Dispholidus typus typus



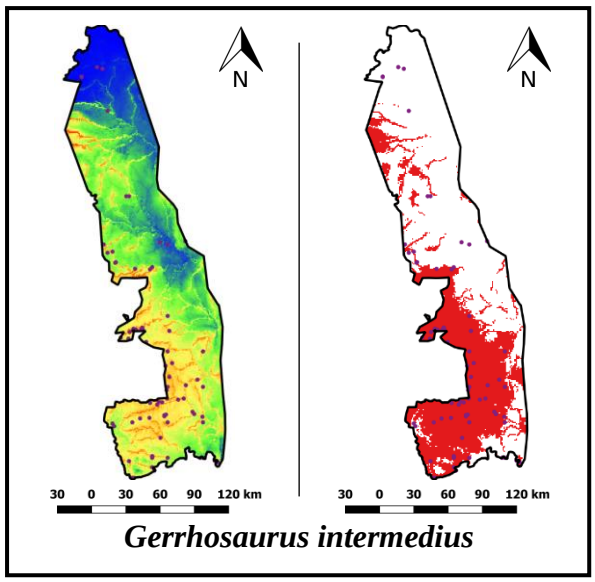
Elapsoidea boulengeri



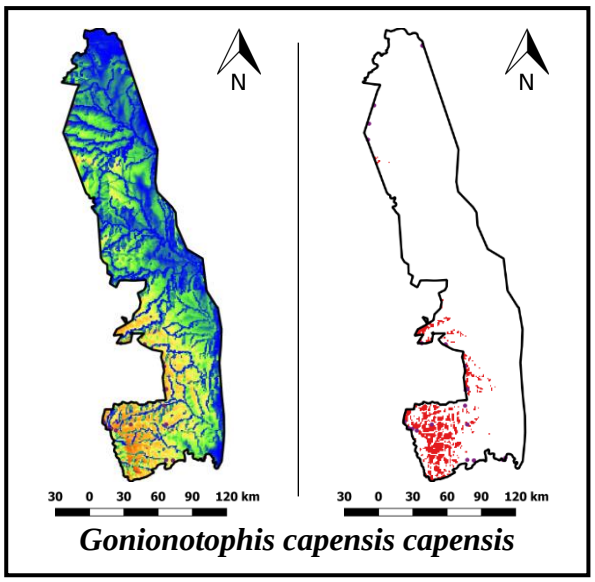
Elapsoidea sundevallii



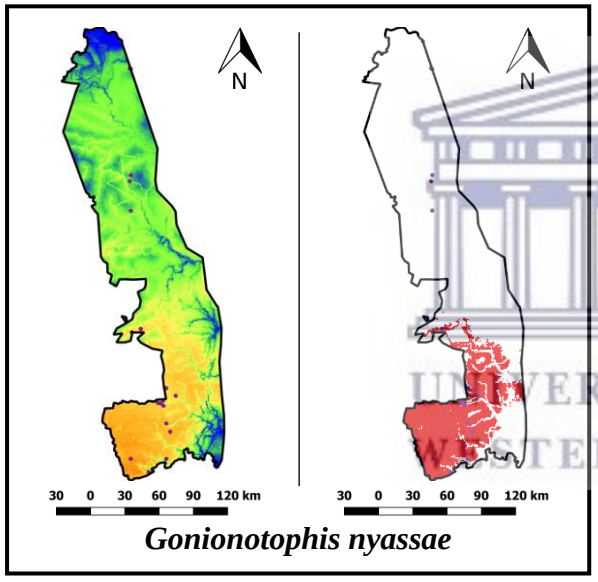
Gerrhosaurus flavigularis



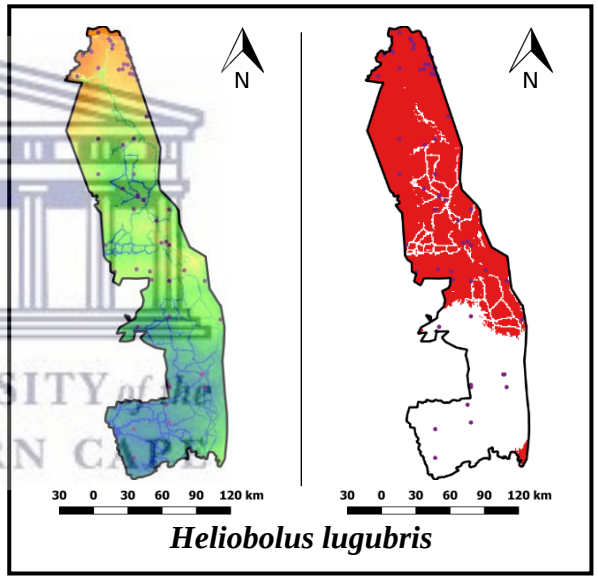
Gerrhosaurus intermedius



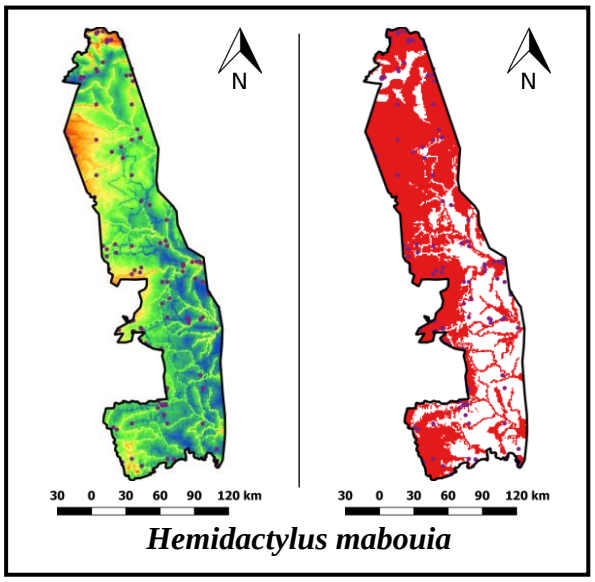
Gonionotophis capensis capensis



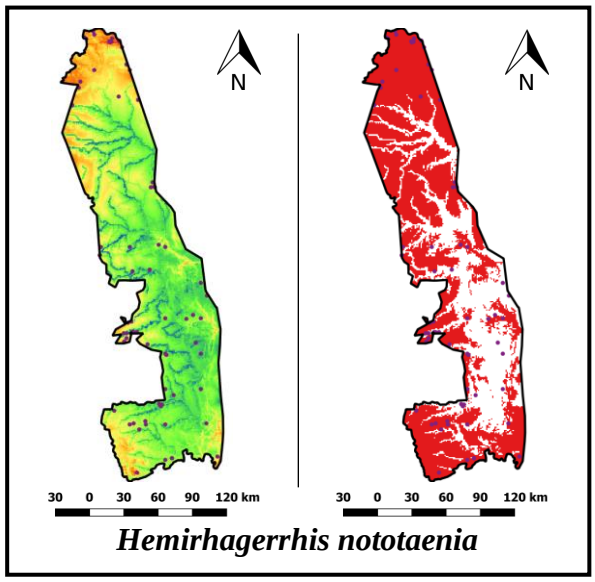
Gonionotophis nyassae



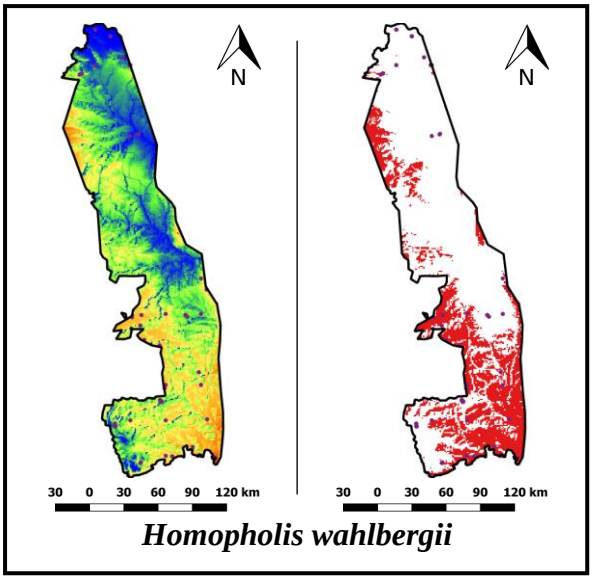
Heliobolus lugubris



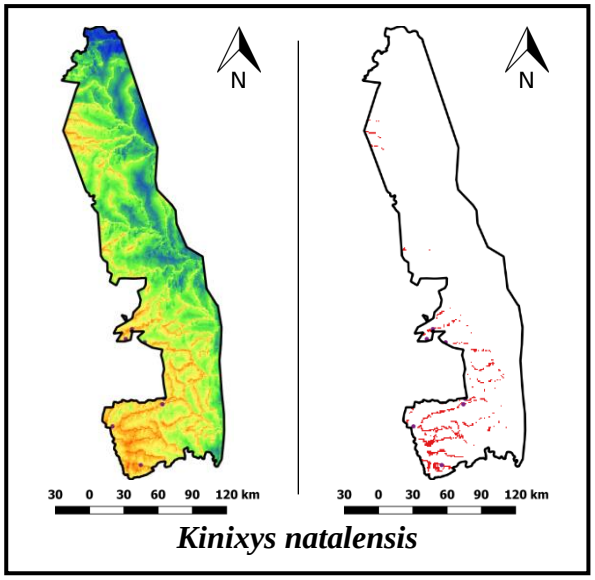
Hemidactylus mabouia



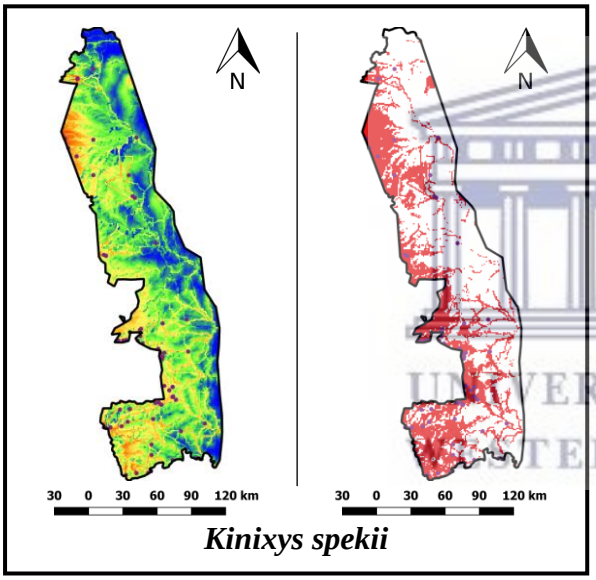
Hemirhagerrhis nototaenia



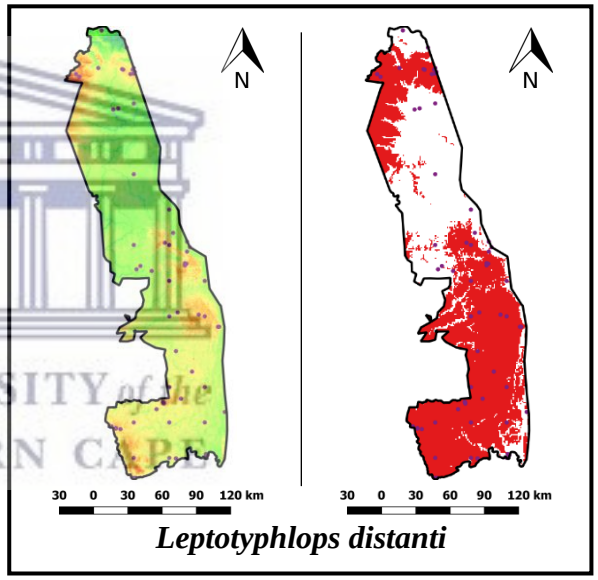
Homopholis wahlbergii



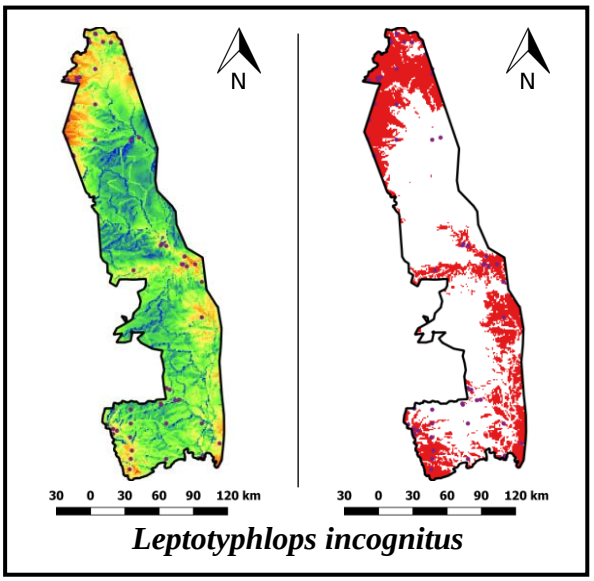
Kinixys natalensis



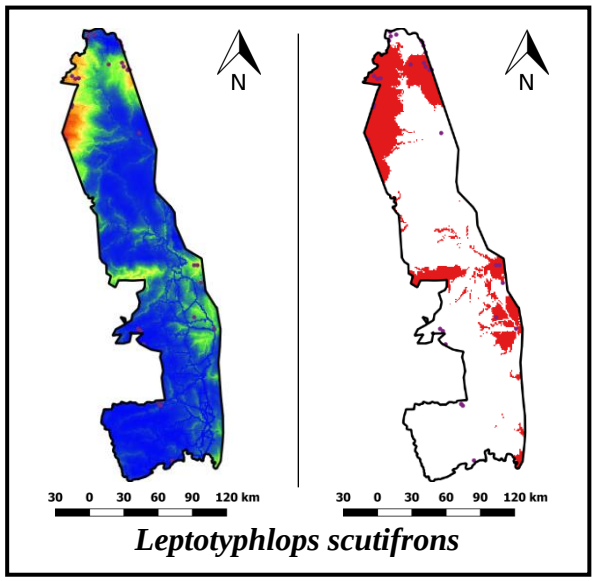
Kinixys spekii



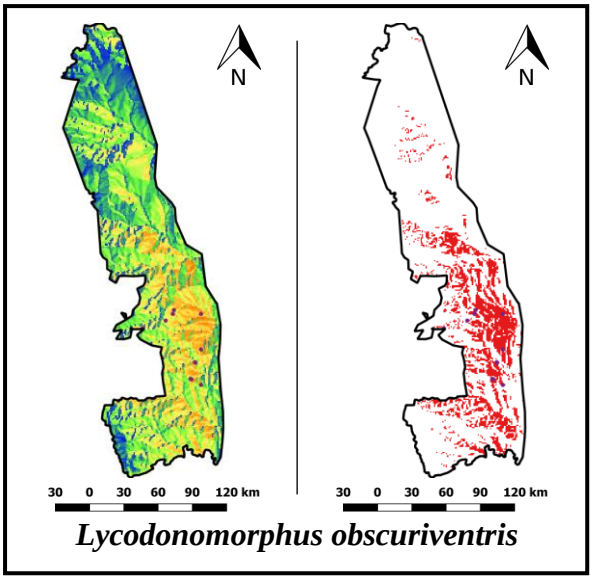
Leptotyphlops distanti



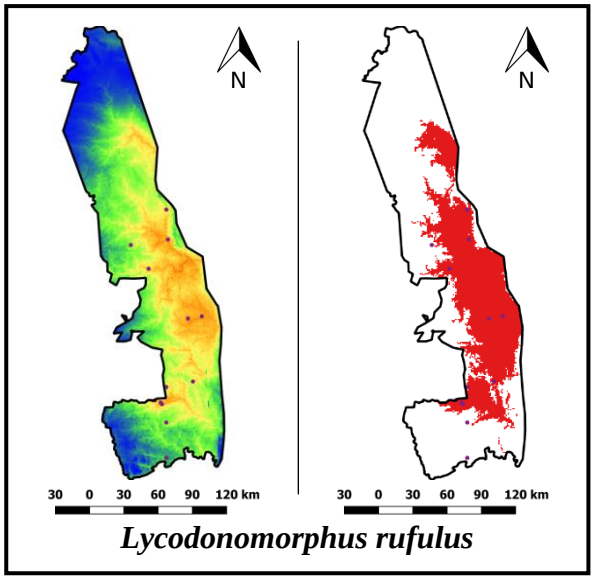
Leptotyphlops incognitus



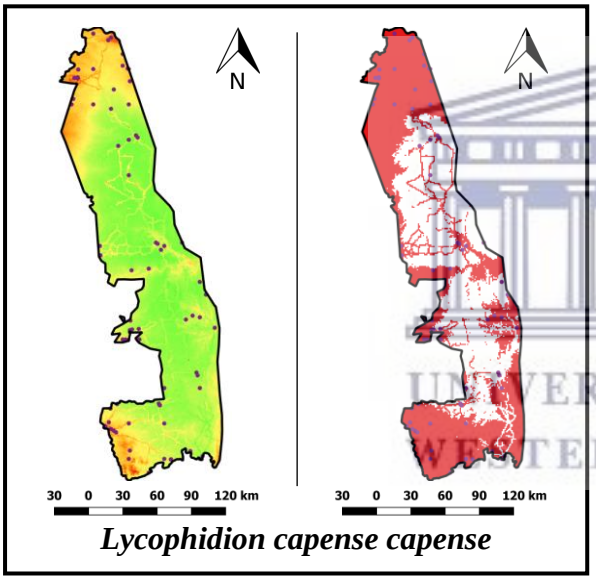
Leptotyphlops scutifrons



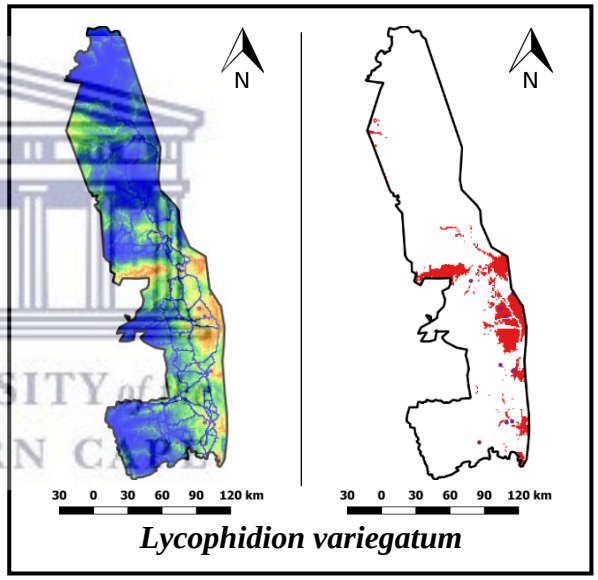
Lycodonomorphus obscuriventris



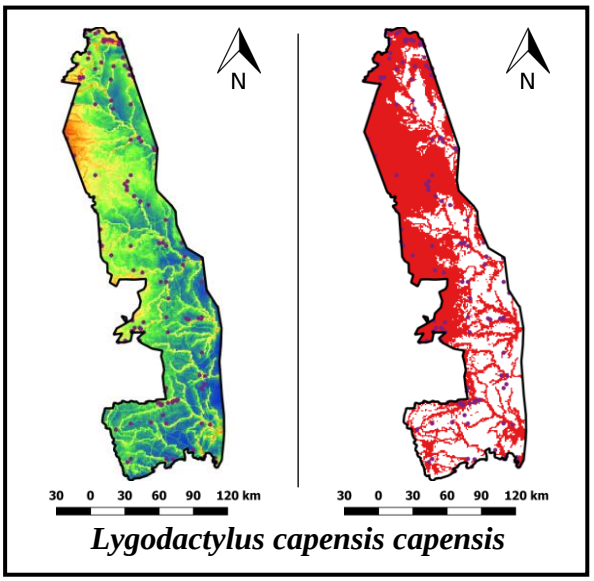
Lycodonomorphus rufulus



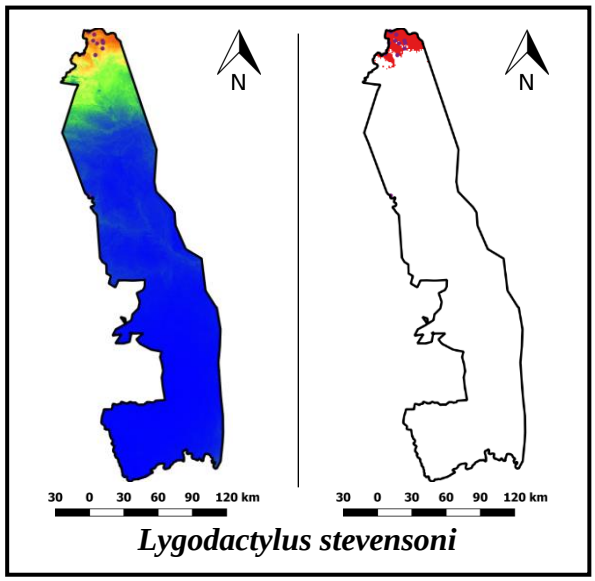
Lycophidion capense capense



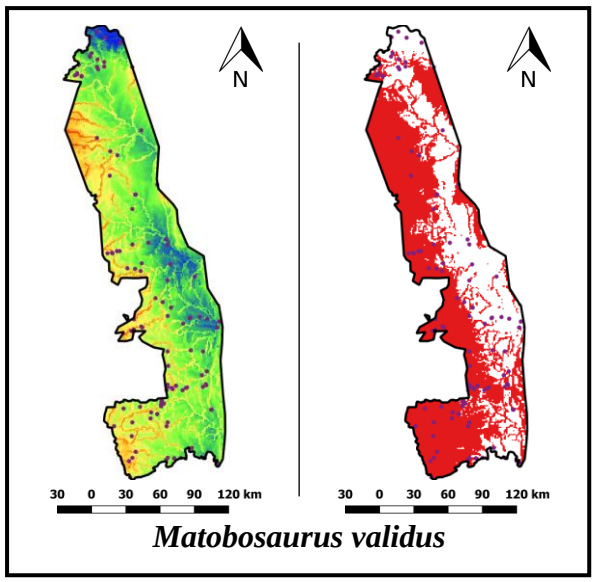
Lycophidion variegatum



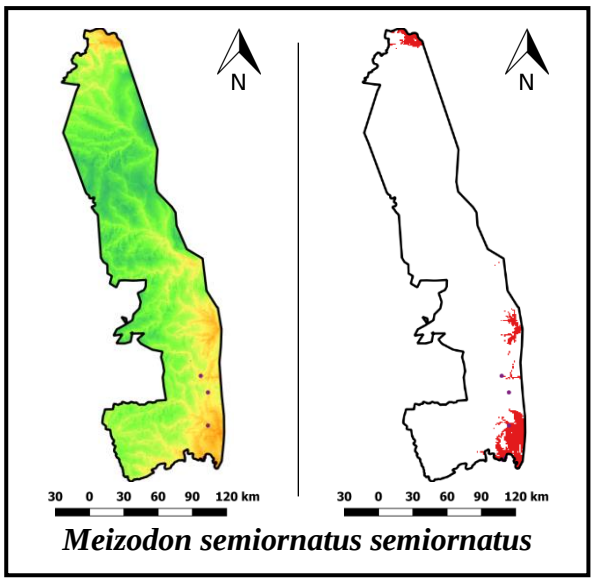
Lygodactylus capensis capensis



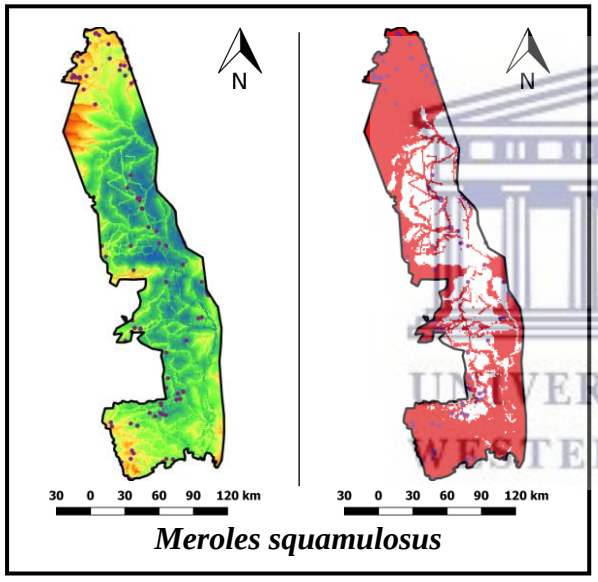
Lygodactylus stevensoni



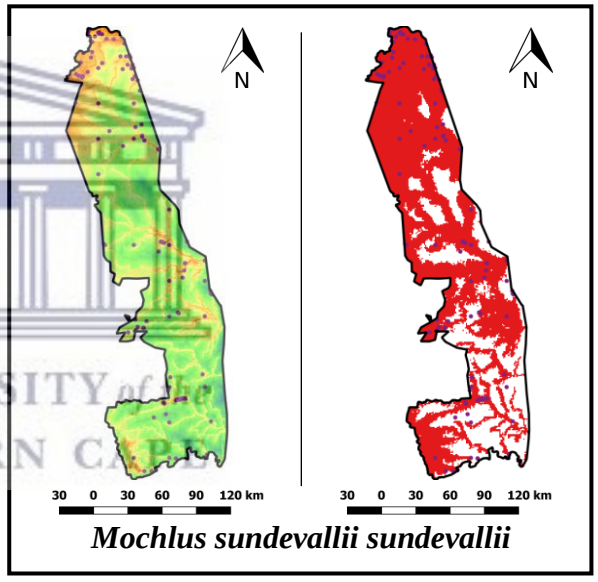
Matobosaurus validus



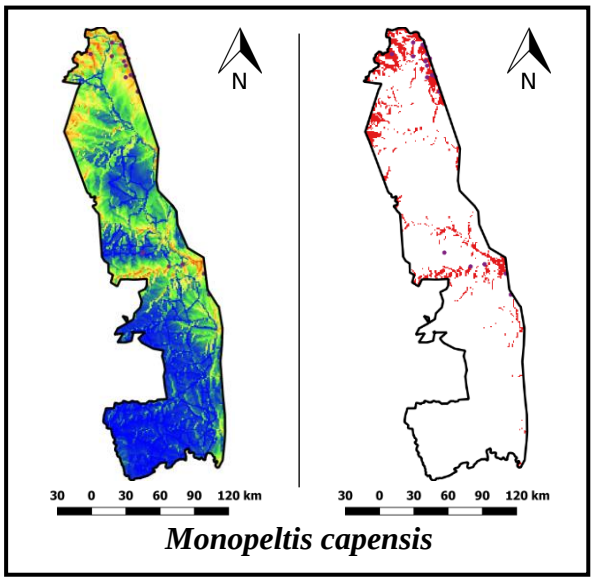
Meizodon semiornatus semiornatus



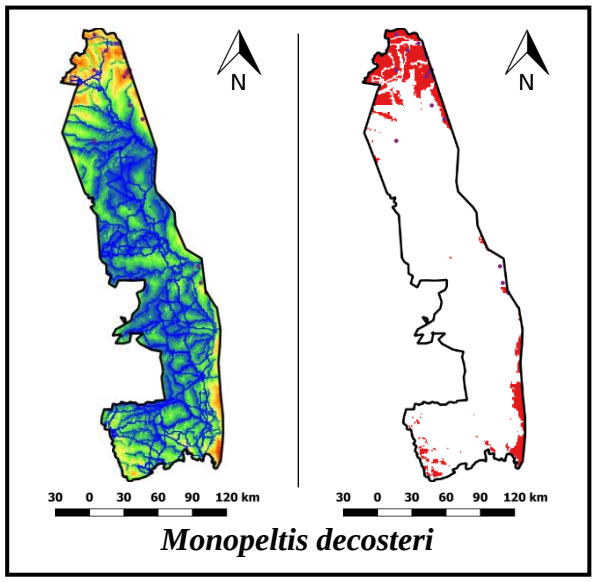
Meroles squamulosus



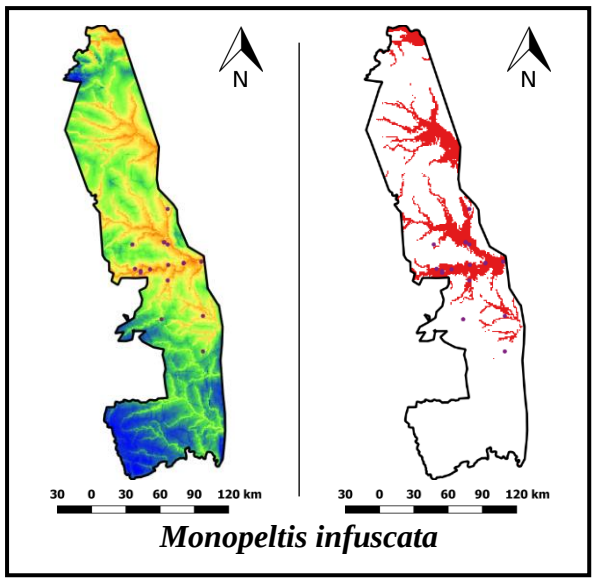
Mochlus sundevallii sundevallii



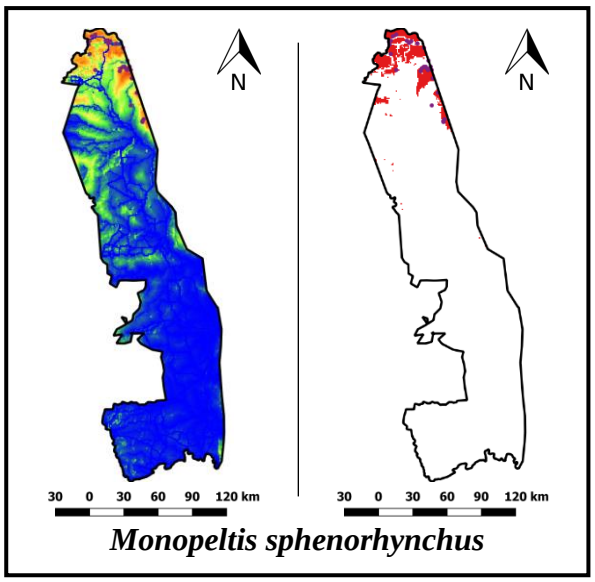
Monopeltis capensis



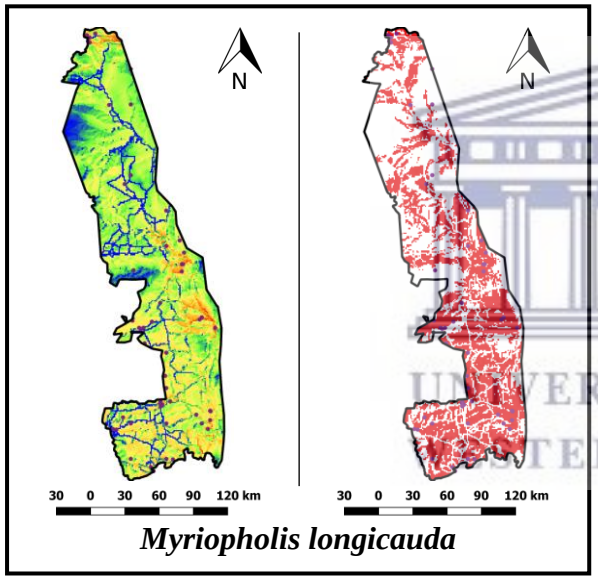
Monopeltis decosteri



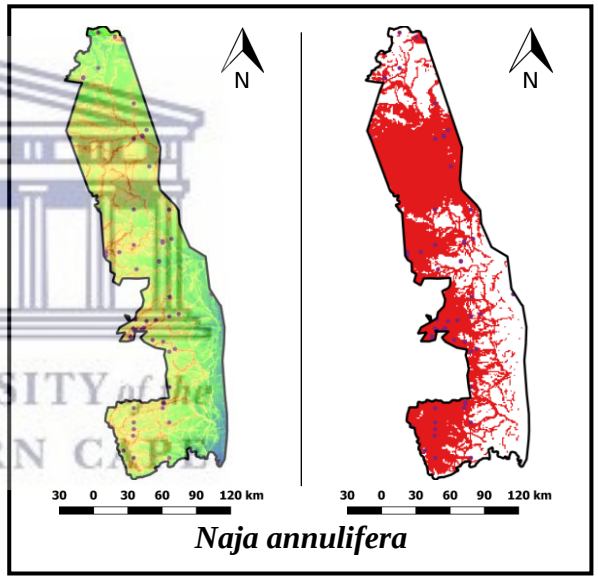
Monopeltis infuscata



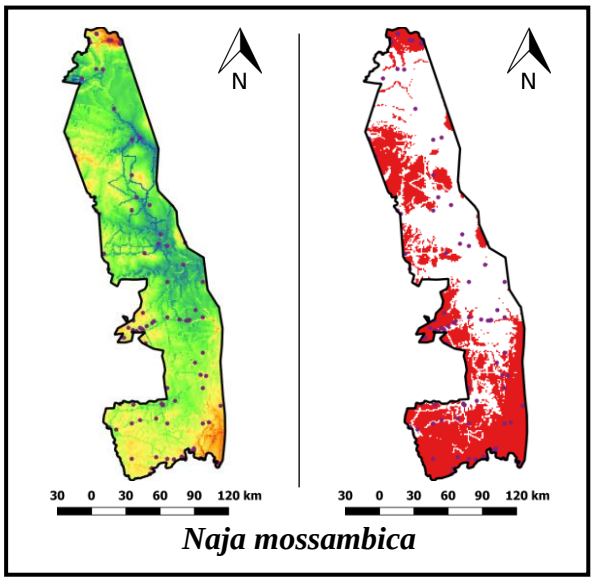
Monopeltis sphenorhynchus



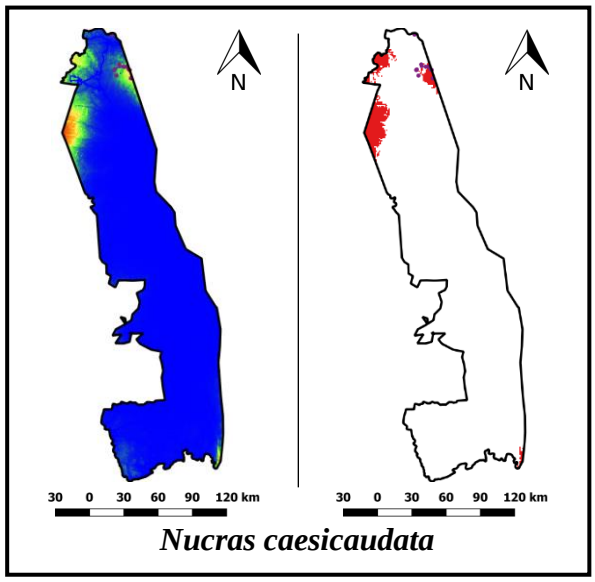
Myriopholis longicauda



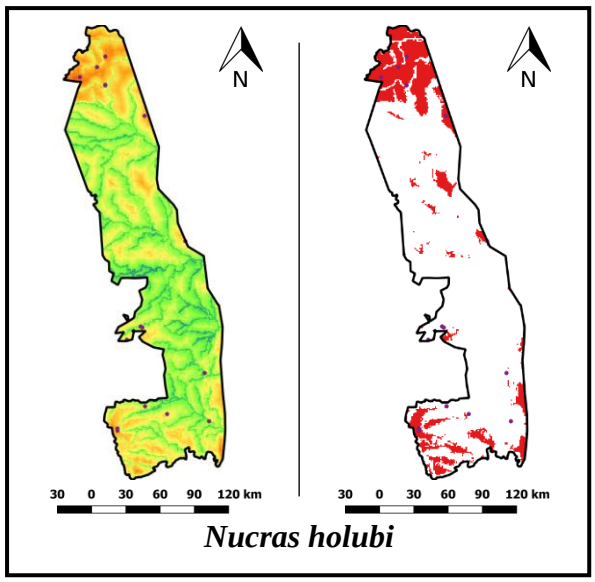
Naja annulifera



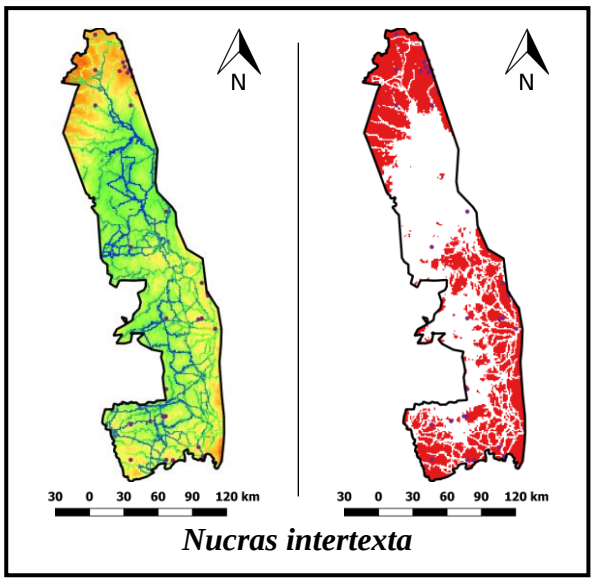
Naja mossambica



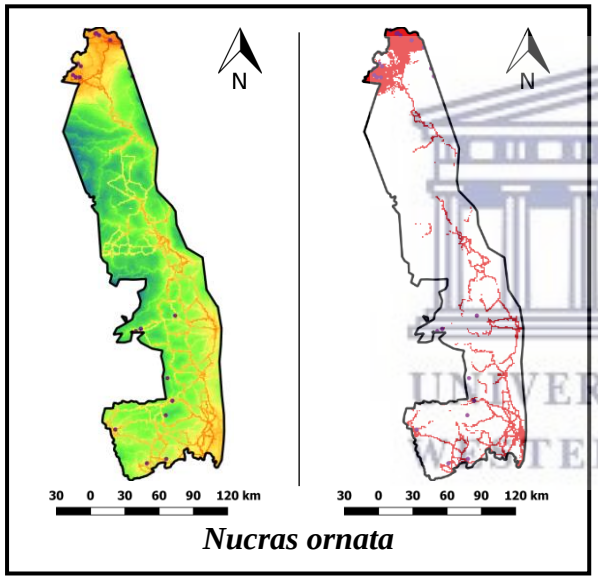
Nucras caesicaudata



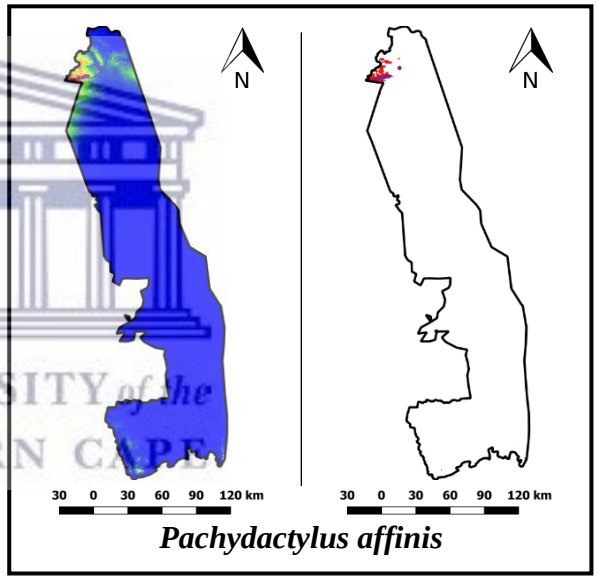
Nucras holubi



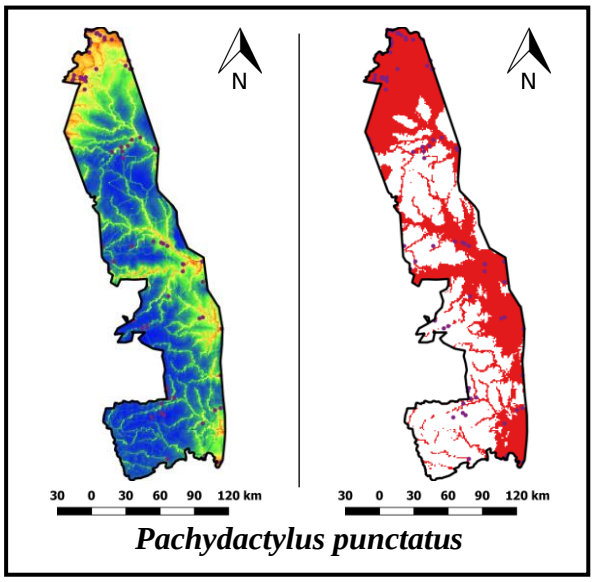
Nucras intertexta



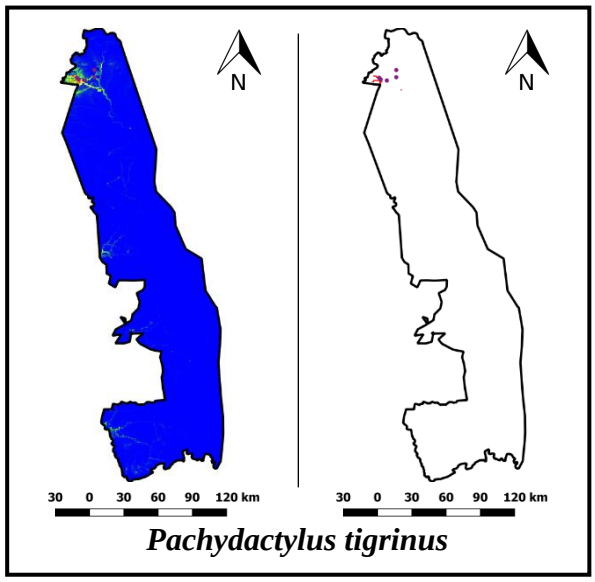
Nucras ornata



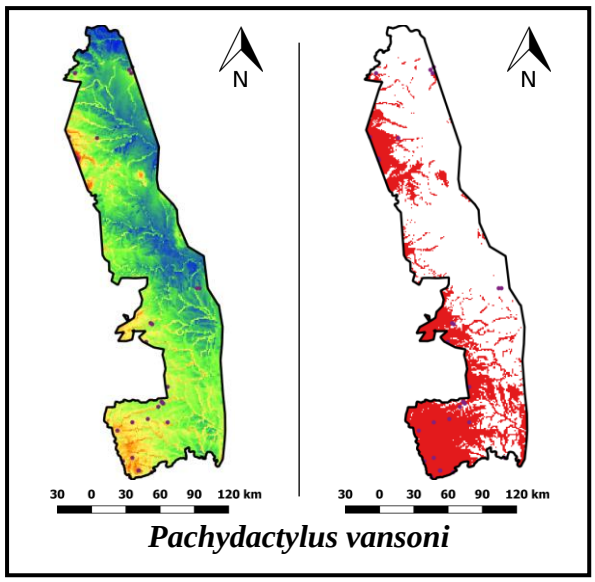
Pachydactylus affinis



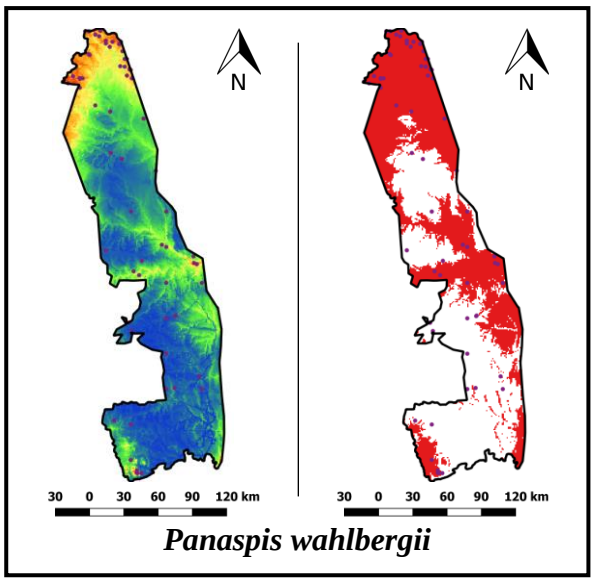
Pachydactylus punctatus



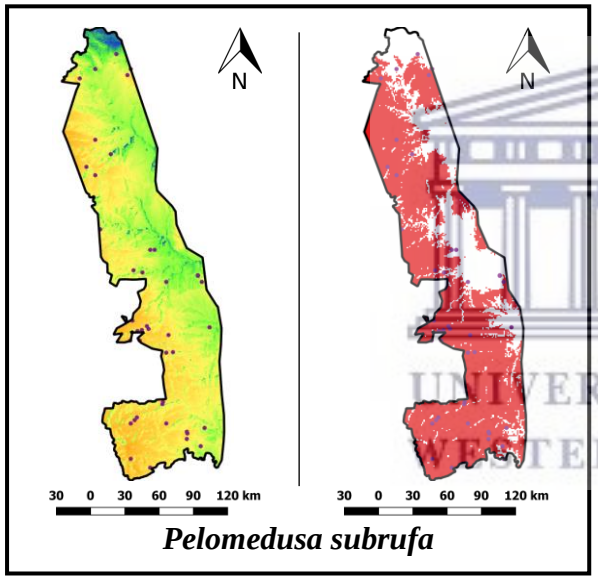
Pachydactylus tigrinus



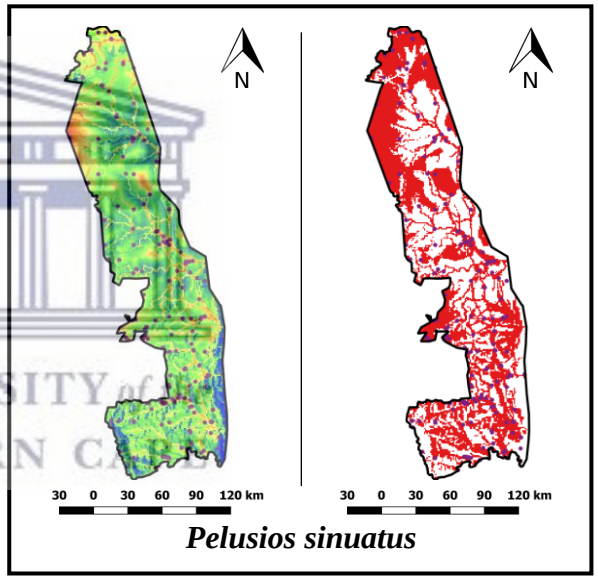
Pachydactylus vansoni



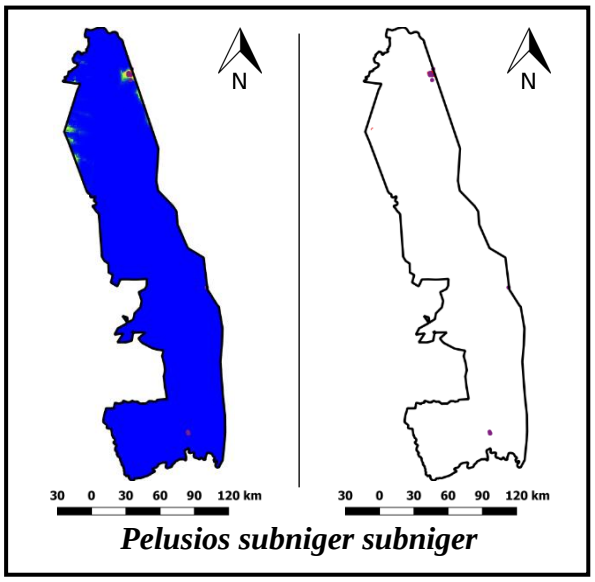
Panaspis wahlbergii



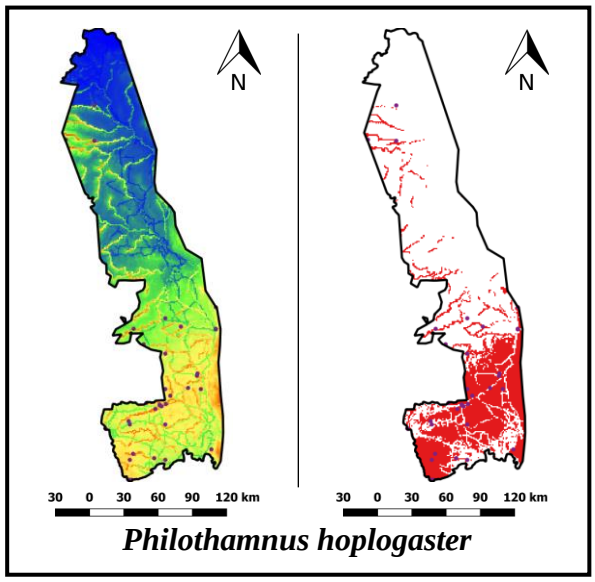
Pelomedusa subrufa



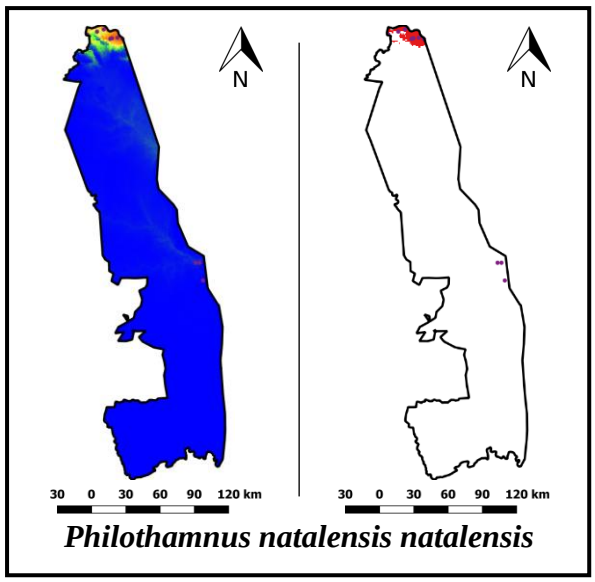
Pelusios sinuatus



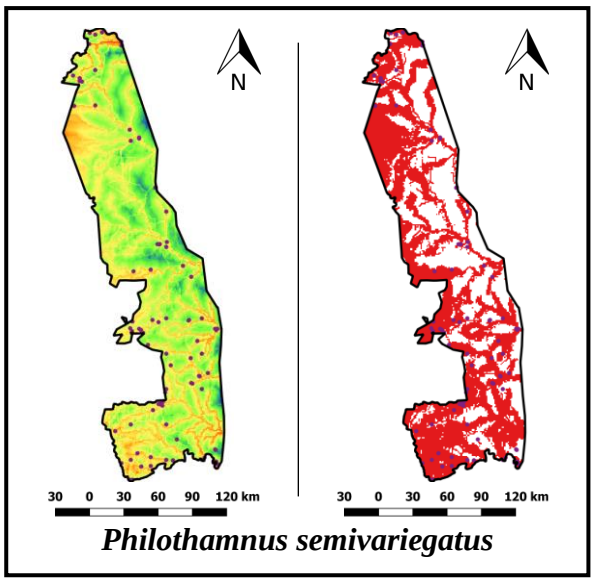
Pelusios subniger subniger



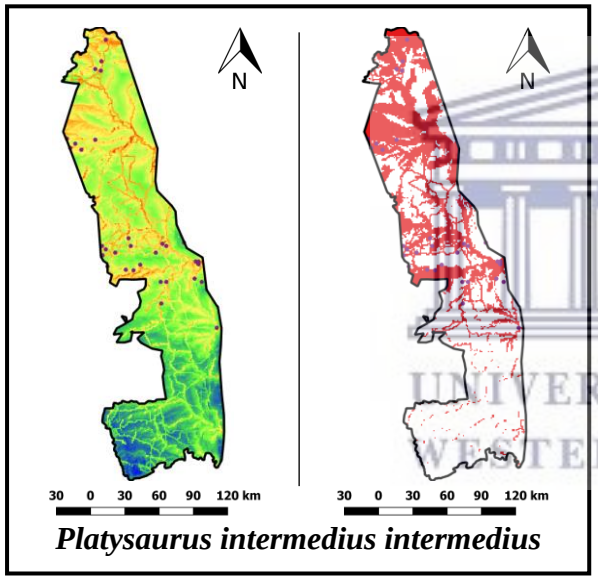
Philothamnus hoplogaster



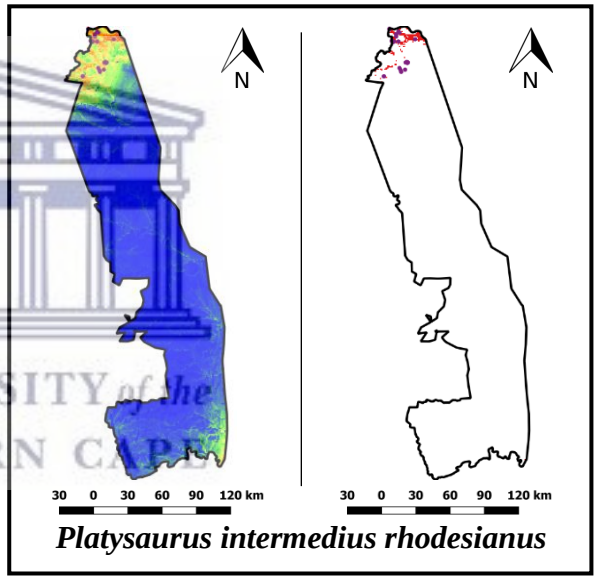
Philothamnus natalensis natalensis



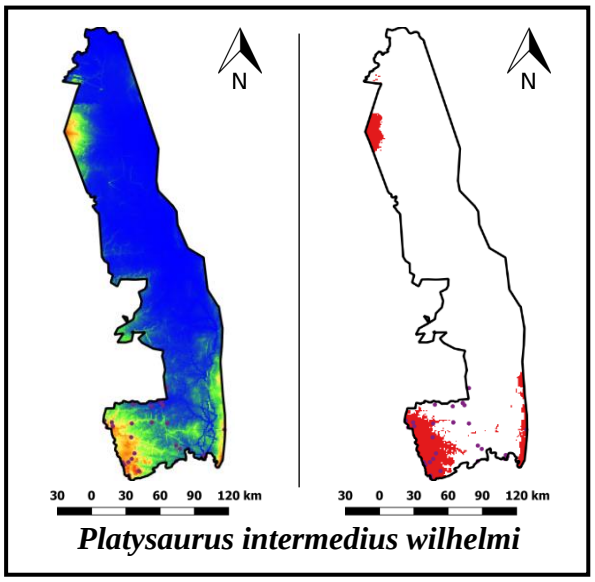
Philothamnus semivariegatus



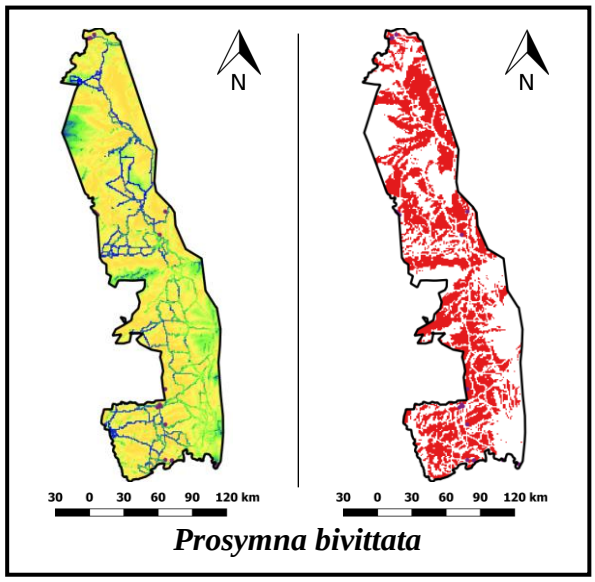
Platysaurus intermedius intermedius



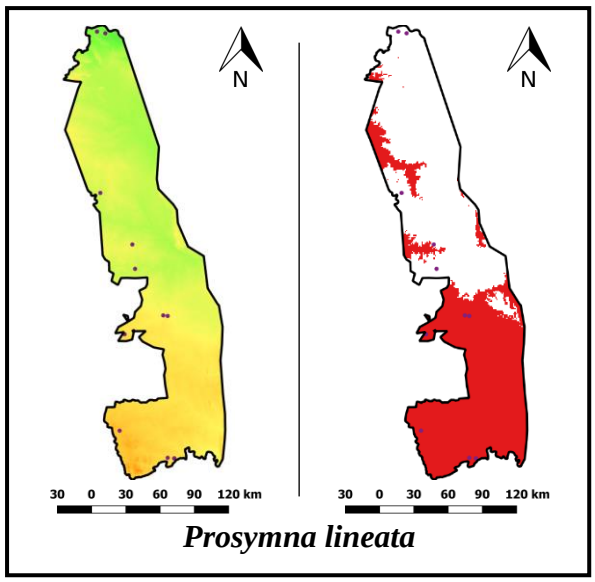
Platysaurus intermedius rhodesianus



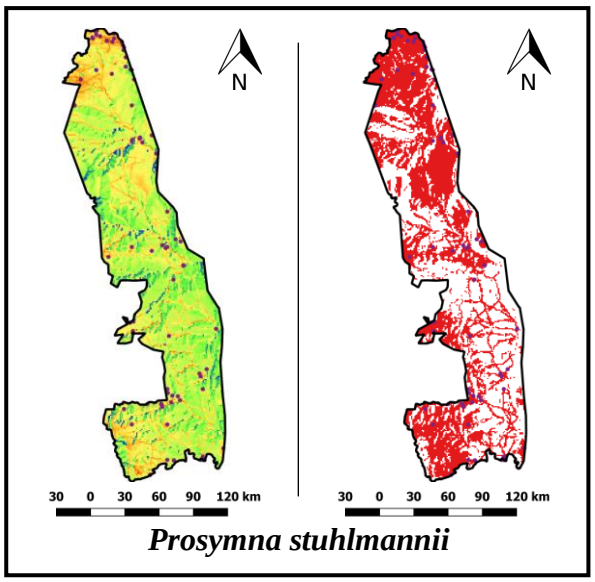
Platysaurus intermedius wilhelmi



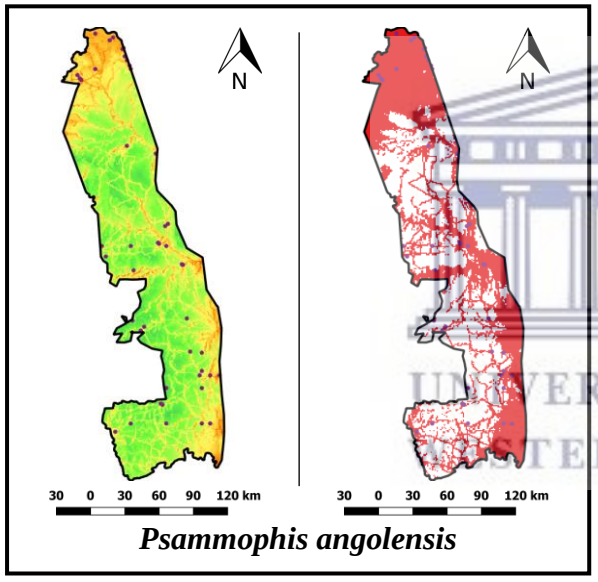
Prosymna bivittata



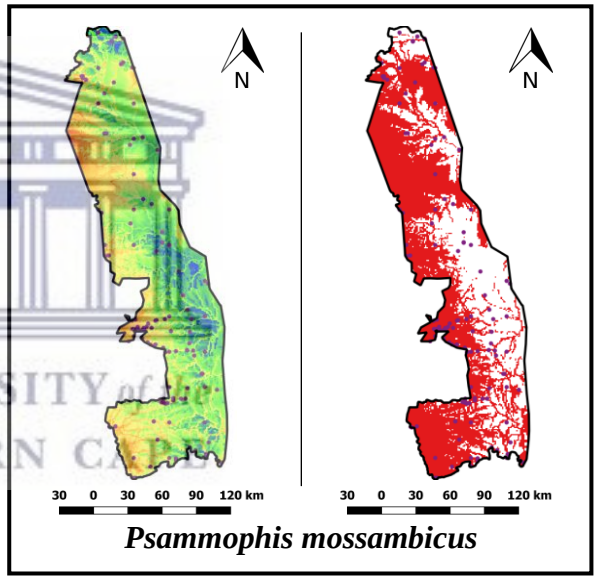
Prosymna lineata



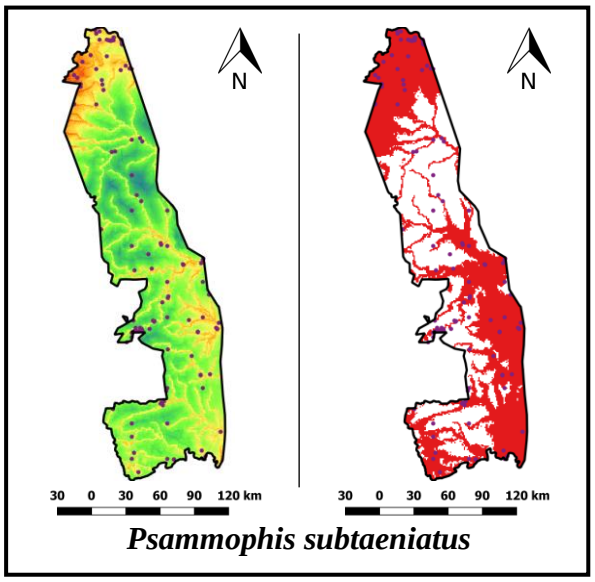
Prosymna stuhlmannii



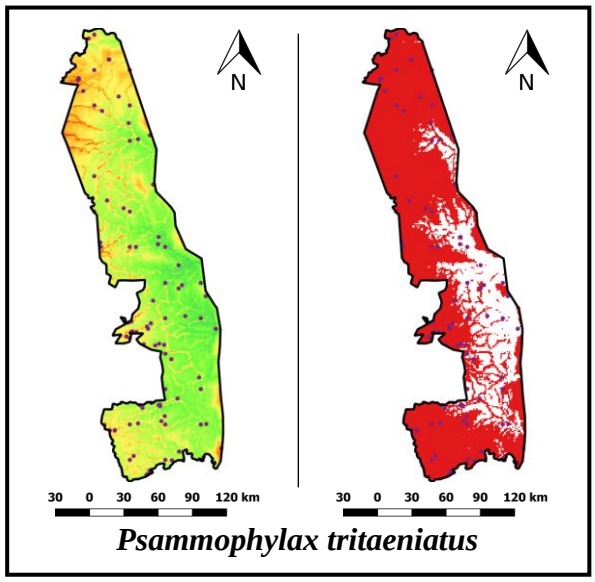
Psammophis angolensis



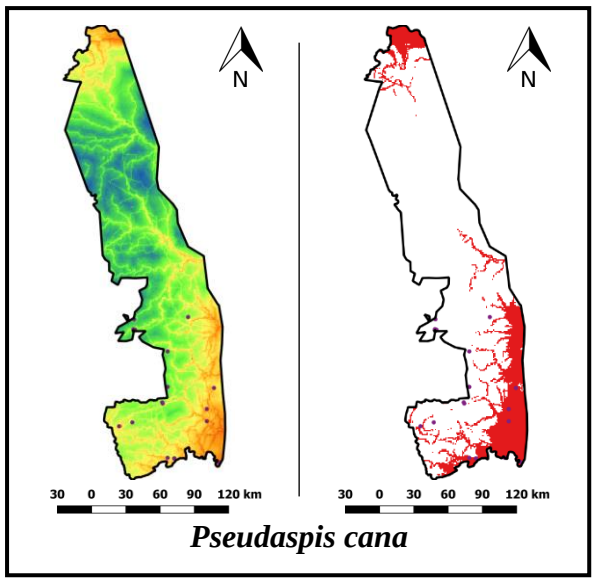
Psammophis mossambicus



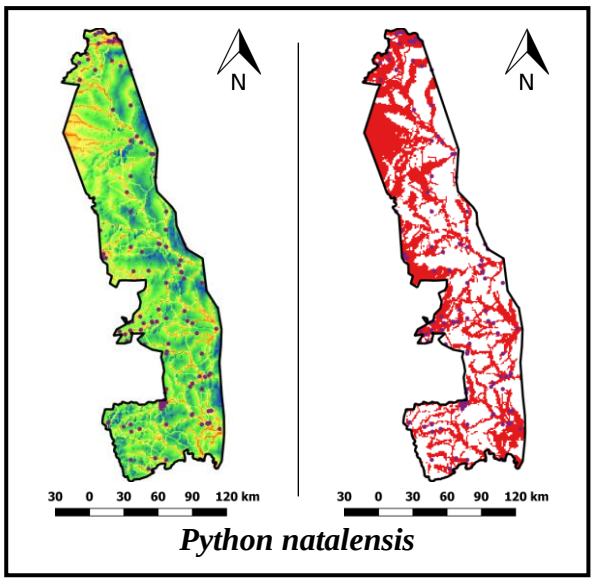
Psammophis subtaeniatus



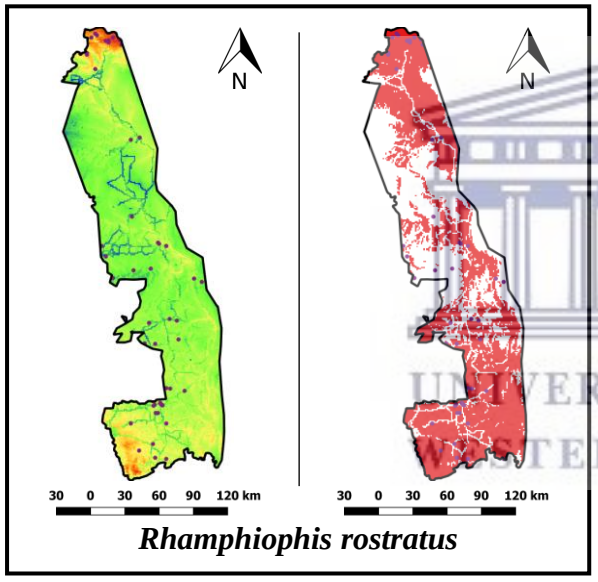
Psammophylax tritaeniatus



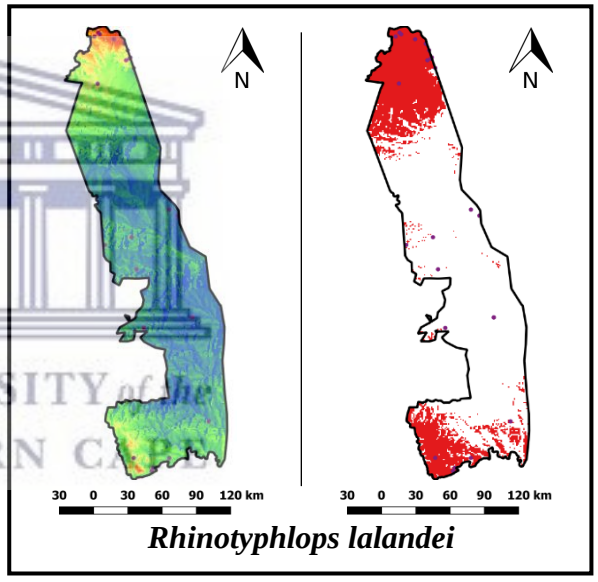
Pseudaspis cana



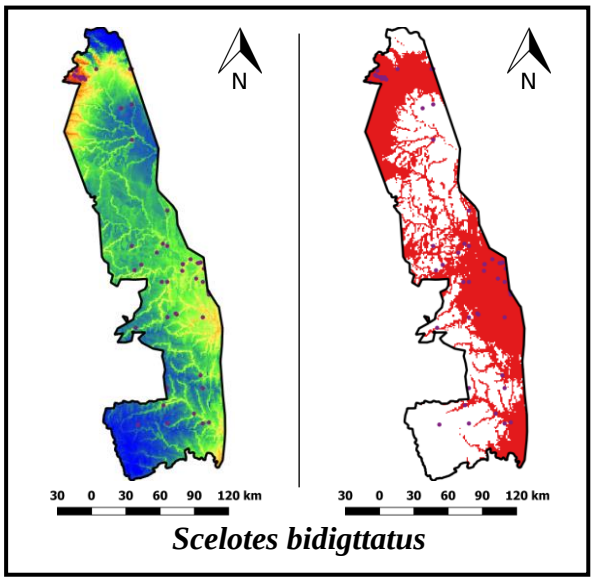
Python natalensis



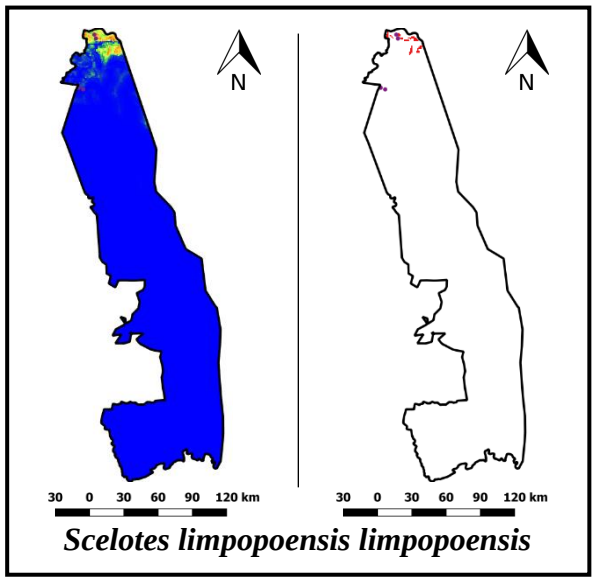
Rhamphiophis rostratus



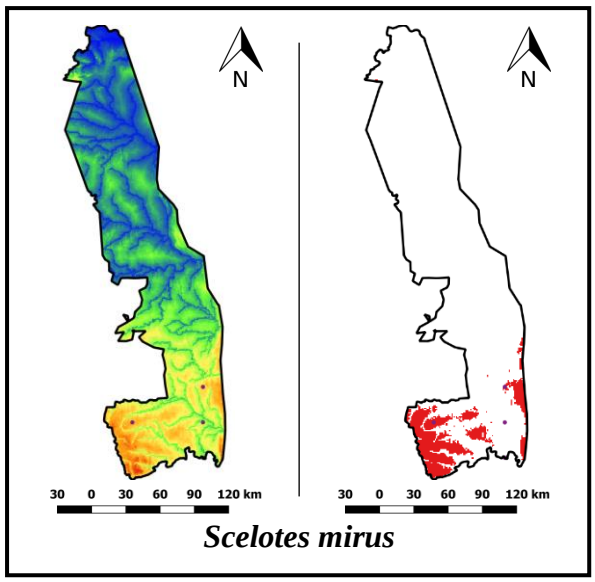
Rhinotyphlops lalandei



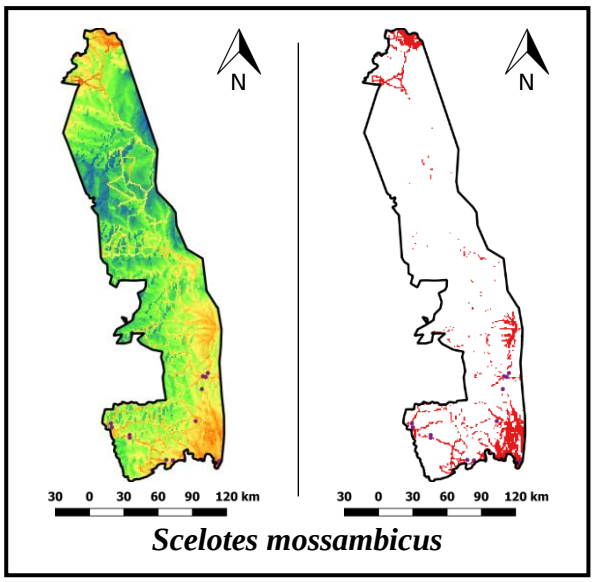
Scelotes bidigttatus



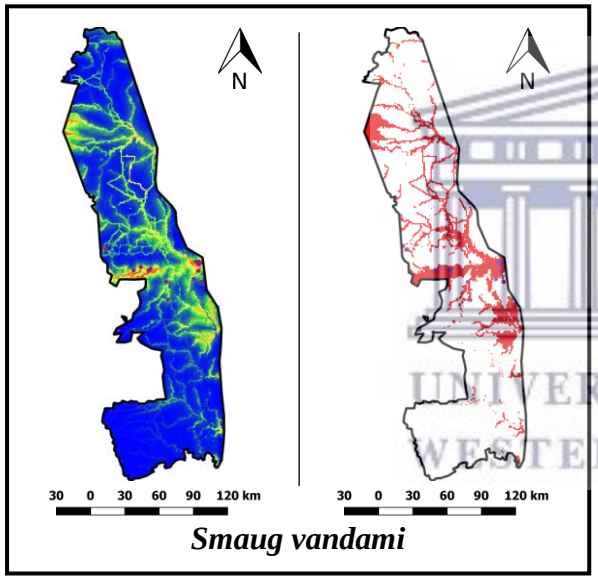
Scelotes limpopoensis limpopoensis



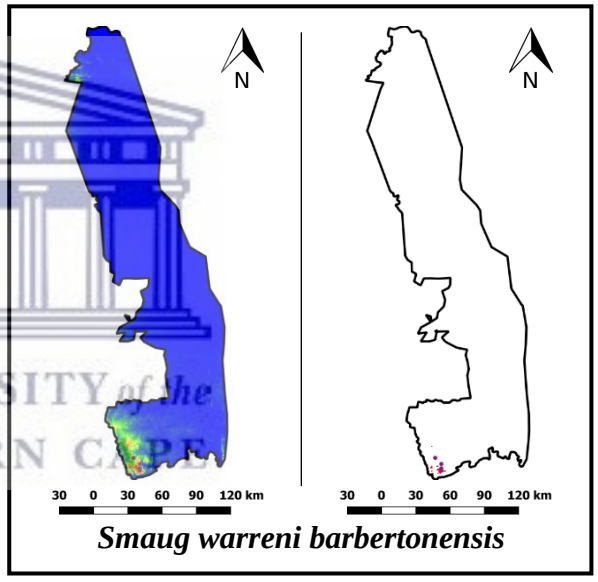
Scelotes mirus



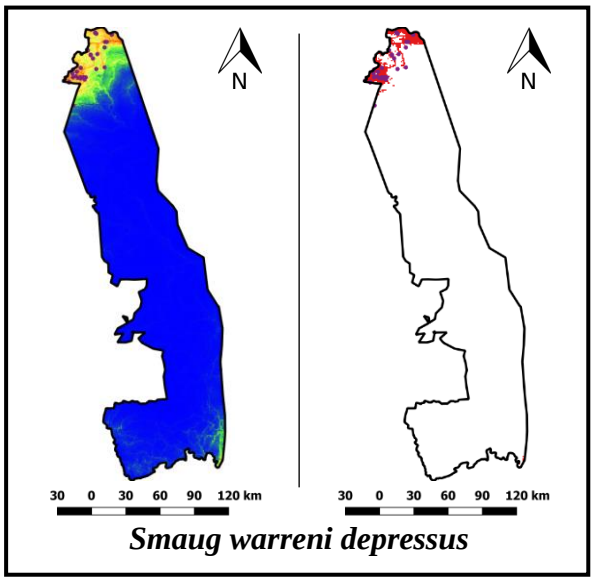
Scelotes mossambicus



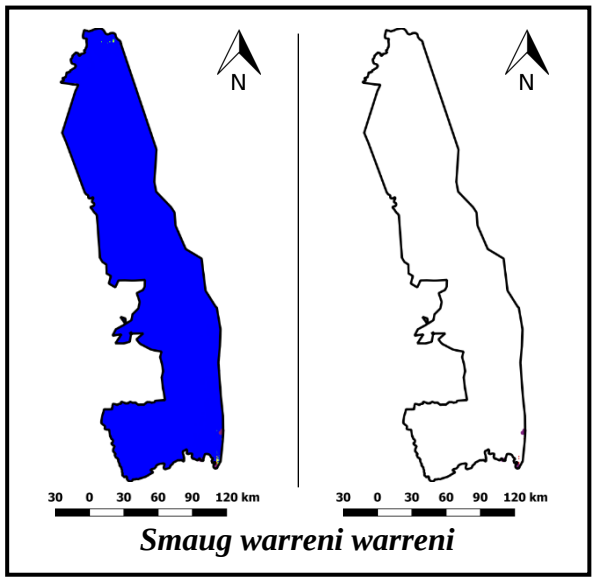
Smaug vandami



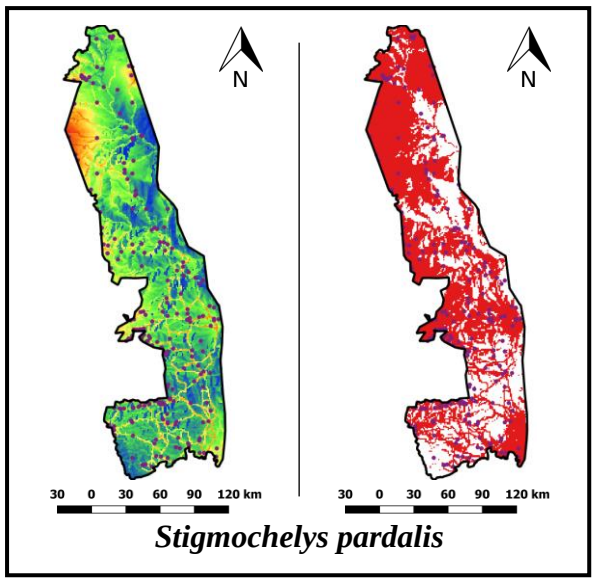
Smaug warreni barbertonensis



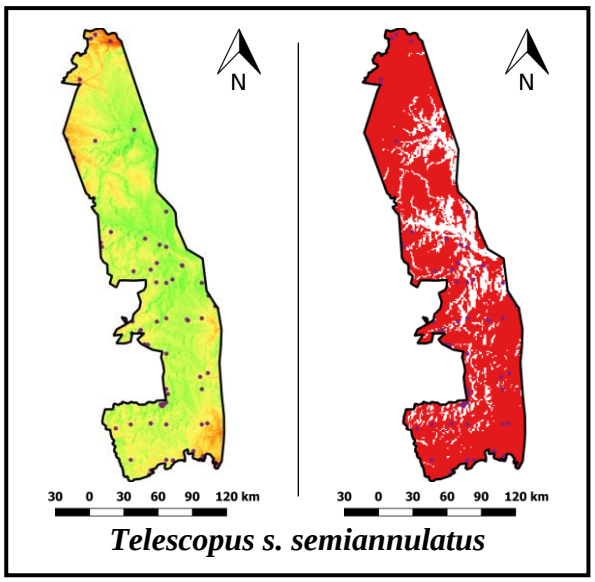
Smaug warreni depressus



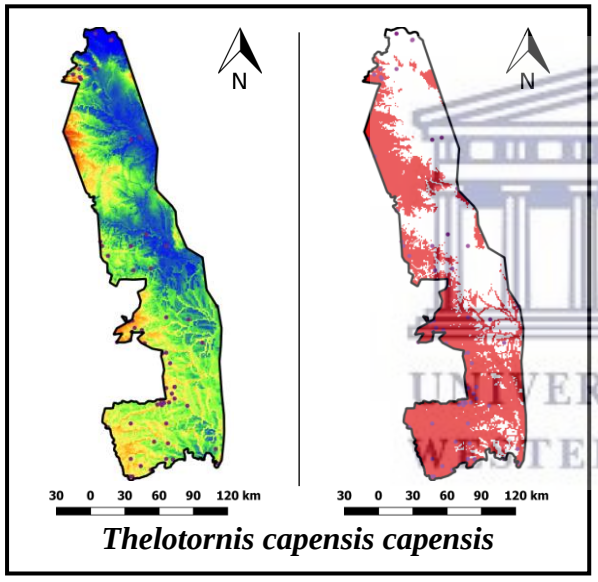
Smaug warreni warreni



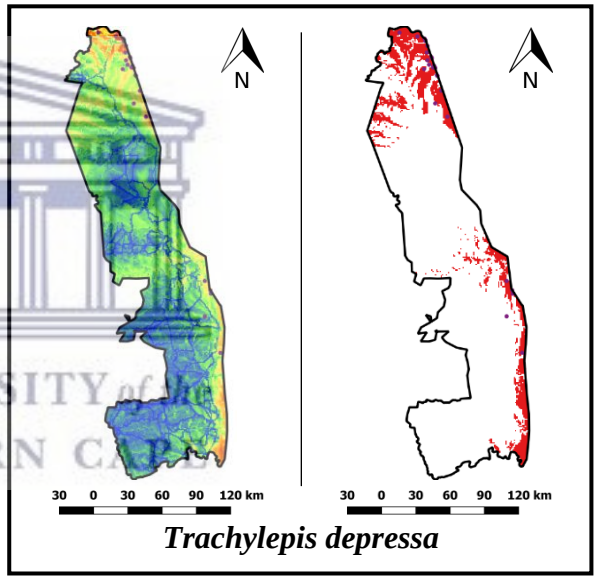
Stigmochelys pardalis



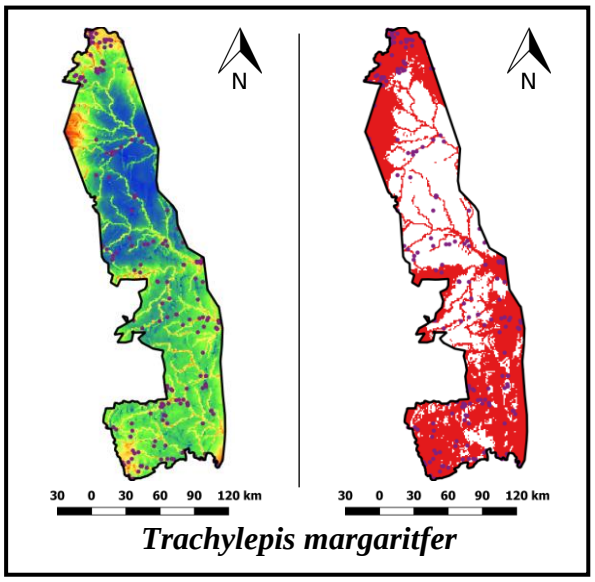
Telescopus s. semiannulatus



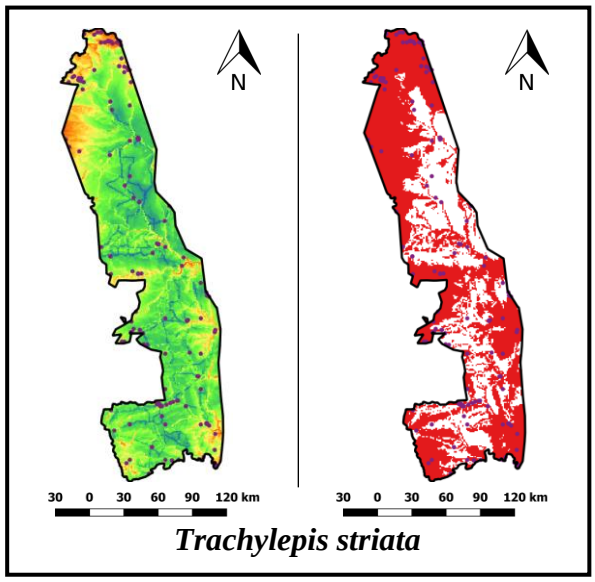
Thelotornis capensis capensis



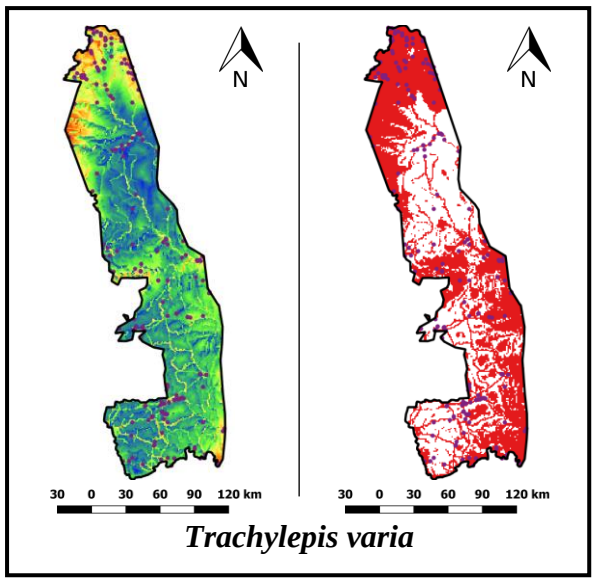
Trachylepis depressa



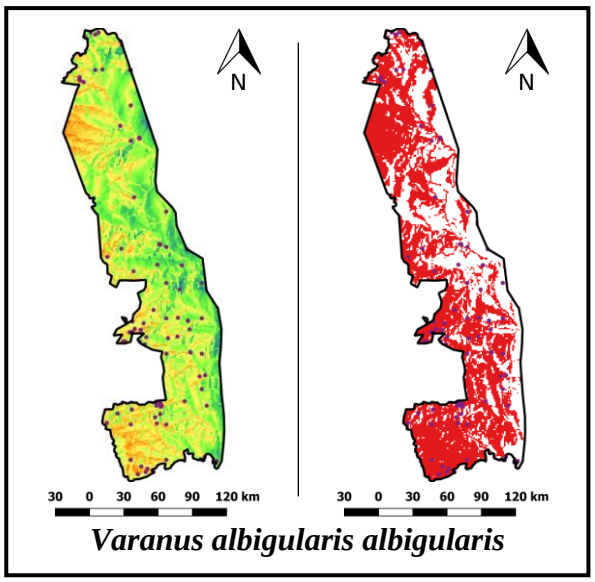
Trachylepis margaritfer



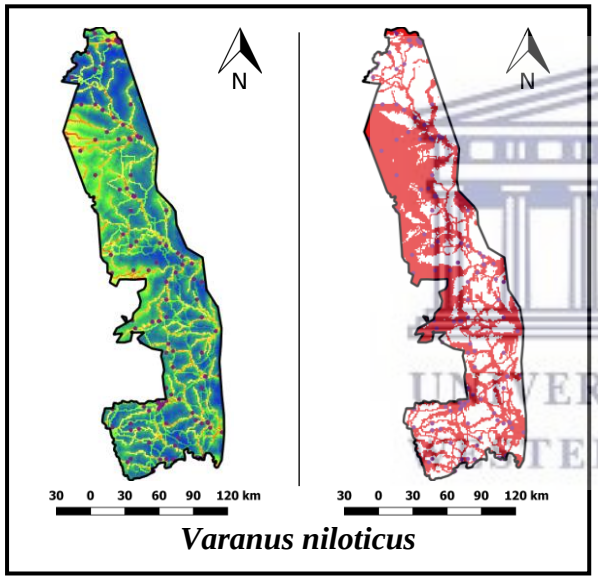
Trachylepis striata



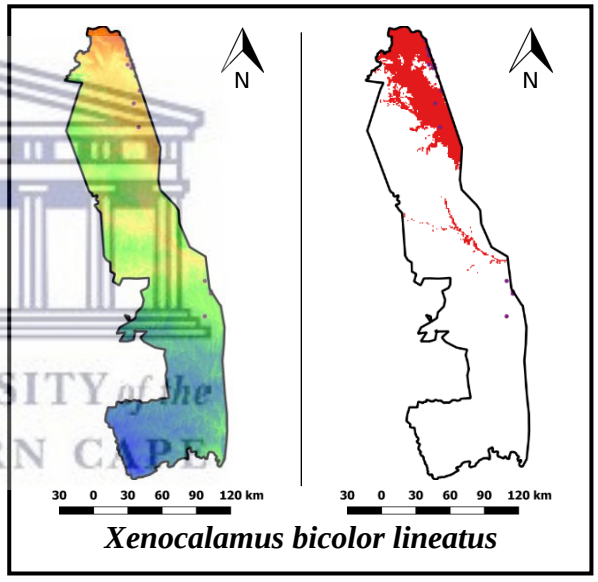
Trachylepis varia



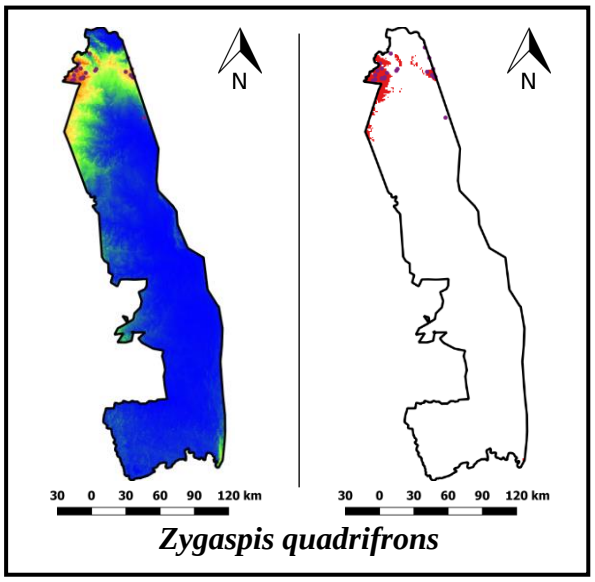
Varanus albigularis albigularis



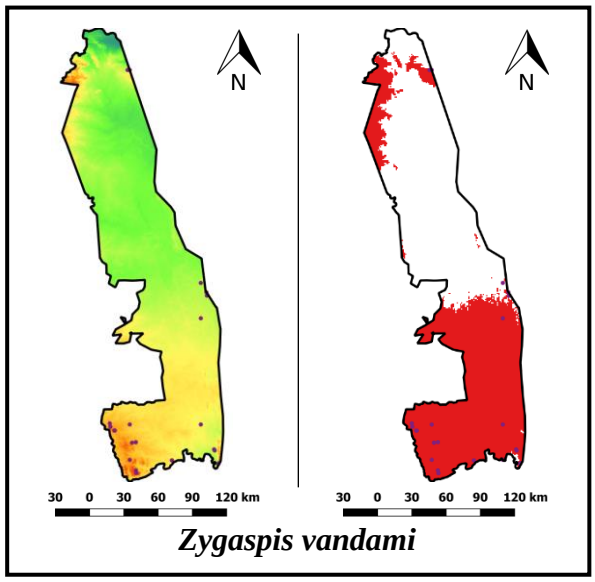
Varanus niloticus



Xenocalamus bicolor lineatus



Zygaspis quadrifrons



Zygaspis vandami

Appendix 4: Species distribution model information.

Species	Training AUC	Test AUC	10th percentile training presence threshold	Number of records
Acanthocercus atricollis atricollis	0.71	0.64	0.32	92
Acontias aurantiacus fitzsimonsi	0.94	0.90	0.37	29
Acontias cregoi	0.97	0.98	0.32	12
Acontias plumbeus	0.78	0.69	0.32	45
Afroablepharus maculicollis	0.92	0.87	0.27	26
Afroedura langi	0.95	0.90	0.33	12
Afroedura marleyi	0.99	0.99	0.46	5
Afroedura multiporis haackei	0.98	0.96	0.63	5
Afroedura pienaari	0.98	0.98	0.64	10
Afroedura transvaalica	0.97	0.96	0.54	12
Afrotyphlops bibronii	0.85	0.76	0.47	5
Afrotyphlops schlegelii	0.73	0.68	0.40	89
Agama aculeata distanti	0.71	0.62	0.34	48
Agama armata	0.96	0.95	0.54	12
Amblyodipsas concolor	0.95	0.93	0.55	4
Amblyodipsas m. microphthalma	0.96	0.94	0.50	12
Amblyodipsas microphthalma nigra	0.98	0.97	0.65	6
Amblyodipsas polylepis polylepis	0.80	0.71	0.41	34
Aparallactus capensis	0.71	0.64	0.35	100
Aparallactus lunulatus lunulatus	0.75	0.67	0.42	33

Aspidelaps scutatus intermedius	0.70	0.62	0.38	53
Atractaspis bibronii	0.71	0.62	0.43	52
Bitis arietans arietans	0.68	0.61	0.36	120
Boaedon capensis	0.66	0.60	0.38	97
Broadleysaurus major	0.71	0.65	0.41	55
Causus defilippii	0.77	0.68	0.47	50
Causus rhombeatus	0.68	0.54	0.44	4
Chamaeleo dilepis dilepis	0.69	0.64	0.36	125
Chirindia langi langi	0.98	0.98	0.32	43
Chondrodactylus turneri	0.69	0.65	0.37	104
Cordylus jonesii	0.64	0.54	0.39	42
Cordylus vittifer	0.85	0.75	0.51	5
Crocodylus niloticus	0.79	0.74	0.31	143
Crotaphopeltis hotamboeia	0.70	0.61	0.46	75
Dasypeltis scabra	0.69	0.63	0.41	82
Dendroaspis polylepis	0.70	0.65	0.44	100
Dipsadoboa aulica	0.78	0.66	0.45	32
Dispholidus typus typus	0.67	0.61	0.42	102
Elapsoidea boulengeri	0.82	0.76	0.41	27
Elapsoidea sundevallii	0.93	0.90	0.44	17
Gerrhosaurus flavigularis	0.68	0.61	0.43	70
Gerrhosaurus intermedius	0.76	0.69	0.44	68
Gonionotophis capensis capensis	0.84	0.76	0.41	20
Gonionotophis nyassae	0.76	0.62	0.49	17
Heliobolus lugubris	0.71	0.65	0.35	63

Hemidactylus mabouia	0.70	0.66	0.28	115
Hemirhagerrhis nototaenia	0.68	0.58	0.42	64
Homopholis wahlbergii	0.76	0.64	0.45	39
Kinixys natalensis	0.90	0.87	0.60	7
Kinixys spekii	0.84	0.78	0.41	56
Leptotyphlops distanti	0.71	0.61	0.42	58
Leptotyphlops incognitus	0.83	0.77	0.33	57
Leptotyphlops scutifrons	0.84	0.78	0.31	32
Lycodonomorphus obscuriventris	0.83	0.75	0.50	9
Lycodonomorphus rufulus	0.73	0.63	0.50	13
Lycophidion capense capense	0.71	0.63	0.40	67
Lycophidion variegatum	0.86	0.83	0.50	9
Lygodactylus capensis capensis	0.71	0.68	0.31	131
Lygodactylus stevensoni	0.97	0.97	0.65	11
Matobosaurus validus	0.67	0.61	0.40	98
Meizodon semiornatus semiornatus	0.93	0.88	0.60	5
Meroles squamulosus	0.70	0.65	0.35	90
Mochlus sundevallii sundevallii	0.71	0.68	0.36	97
Monopeltis capensis	0.92	0.85	0.44	18
Monopeltis decosteri	0.84	0.76	0.47	16
Monopeltis infuscata	0.89	0.84	0.48	14
Monopeltis sphenorhynchus	0.96	0.94	0.36	27
Myriopholis longicauda	0.77	0.66	0.47	50
Naja annulifera	0.70	0.60	0.42	54
Naja mossambica	0.72	0.64	0.45	88

Nucras caesicaudata	0.95	0.95	0.42	9
Nucras holubi	0.80	0.74	0.49	14
Nucras intertexta	0.73	0.64	0.39	36
Nucras ornata	0.82	0.77	0.37	22
Pachydactylus affinis	0.99	0.99	0.72	7
Pachydactylus punctatus	0.79	0.75	0.32	75
Pachydactylus tigrinus	0.98	0.97	0.66	5
Pachydactylus vansoni	0.83	0.75	0.34	27
Panaspis wahlbergii	0.75	0.71	0.30	60
Pelomedusa subrufa	0.70	0.56	0.46	42
Pelusios sinuatus	0.72	0.65	0.39	153
Pelusios subniger subniger	0.84	0.71	0.50	8
Philothamnus hoplogaster	0.83	0.75	0.40	32
Philothamnus natalensis natalensis	0.96	0.93	0.60	10
Philothamnus semivariegatus	0.72	0.65	0.40	78
Platysaurus intermedius intermedius	0.80	0.71	0.29	32
Platysaurus intermedius rhodesianus	0.98	0.95	0.62	14
Platysaurus intermedius wilhelmi	0.93	0.90	0.34	28
Prosymna bivittata	0.76	0.67	0.49	17
Prosymna lineata	0.64	0.53	0.50	11
Prosymna stuhlmannii	0.72	0.65	0.38	68
Psammophis angolensis	0.71	0.58	0.44	42
Psammophis mossambicus	0.67	0.61	0.44	112
Psammophis subtaeniatus	0.73	0.67	0.38	97
Psammophylax tritaeniatus	0.66	0.58	0.46	86

Pseudaspis cana	0.82	0.81	0.51	16
Python natalensis	0.73	0.68	0.40	112
Rhamphiophis rostratus	0.74	0.65	0.42	45
Rhinotyphlops lalandei	0.76	0.67	0.46	22
Scelotes bidigittatus	0.74	0.68	0.34	47
Scelotes limpopoensis limpopoensis	0.97	0.94	0.60	4
Scelotes mirus	0.88	0.94	0.51	4
Scelotes mossambicus	0.92	0.87	0.48	15
Smaug vandami	0.83	0.76	0.42	16
Smaug warreni barbertonensis	0.99	0.98	0.71	4
Smaug warreni depressus	0.98	0.98	0.35	25
Smaug warreni warreni	0.99	0.99	0.67	4
Stigmochelys pardalis	0.73	0.67	0.37	198
Telescopus s. semiannulatus	0.66	0.57	0.44	64
Thelotornis capensis capensis	0.73	0.65	0.38	65
Trachylepis depressa	0.89	0.83	0.35	21
Trachylepis margaritifer	0.74	0.70	0.34	176
Trachylepis striata	0.72	0.70	0.28	116
Trachylepis varia	0.75	0.69	0.29	166
Varanus albigularis albigularis	0.70	0.61	0.45	81
Varanus niloticus	0.73	0.70	0.33	116
Xenocalamus bicolor lineatus	0.86	0.80	0.43	17
Zygaspis quadrifrons	0.97	0.96	0.39	20
Zygaspis vandami	0.87	0.82	0.36	20