

ECOLOGICAL SUITABILITY MODELLING FOR ANTHRAX IN THE KRUGER NATIONAL PARK, SOUTH AFRICA

by

Pieter Johannes Steenkamp

Submitted in partial fulfilment of the requirements for the degree
Master of Veterinary Medicine, Specialization: Wildlife

Faculty of Veterinary Science
University of Pretoria
PRETORIA

(January 2013)

Keywords

Bacillus anthracis, Anthrax, Species distribution modelling, Maxent, Geographic Information System (GIS), Wildlife, Kruger National Park, South Africa

Acknowledgements

The supervisors for this project were Drs. Louis van Schalkwyk and Henriette van Heerden. Thank you for your time, effort and valuable inputs. Especially thanks to Dr. van Schalkwyk for all the guidance, motivation and the initiation of the project.

I wish to thank my wife and best friend, Danielle, who has inspired, encouraged, and fully supported me always.

Thank you to the staff at the State Veterinary Services in Skukuza, Mr. At Dekker, Drs. Linmarie de Klerk-Lorist and Roy Bengis. Thank you for all the data and willingness to help.

Thank you to Dr. Richard Burroughs for all of your advice and guidance throughout the course of this degree.

I wish to thank my siblings for their kind encouragement and love.

Last but not least, I wish to thank my parents. I am forever indebted to them for their understanding, endless patience, love and support.

And Asus.

Table of Contents

Acknowledgements	ii
List of Figures	v
List of Tables	vii
List of Appendices.....	viii
Summary	ix
1. Introduction	1
1.1. Research Background.....	1
1.1.1. Context	1
1.1.2. Study Area.....	2
1.1.3. Species Distribution Modelling (SDM).....	7
1.1.4. Maximum Entropy (Maxent) modelling.....	15
1.1.5. Anthrax	18
1.1.6. Anthrax in the Kruger National Park.....	19
1.2. Problem Statement and Justification	22
1.3. Research questions, study objectives and hypotheses	23
1.4. Assumption and limitations	24
1.5. Software.....	24
2. Materials and Methods	26
2.1. Species Observation Data.....	26
2.1.1. <i>Bacillus anthracis</i> occurrence data.....	26
2.1.2. Processing of Species Observation Data	26
2.2. Environmental Predictors	30
2.2.1. Selection Criteria	30
2.2.2. Spatiotemporal Framework	33
2.2.3. Topography and Soil Variables	33
2.2.4. Processing of the NDVI variable.....	35
2.2.5. Processing of Land Cover variables	38
2.2.6. Processing of Climate Variables.....	38
2.3. Modelling technique	40

2.3.1.	Maxent settings.....	40
2.3.2.	Modelling Procedure	41
2.3.3.	Maxent and sample bias	42
2.3.4.	Null Model.....	43
2.3.5.	Simple multiple regression	43
2.3.6.	Feature selection.....	43
2.3.7.	Probability classes	45
2.3.8.	Gap Analyses.....	46
3.	Results	47
4.	Discussion	64
5.	Synthesis.....	76
5.1.	Conclusion.....	76
5.2.	Recommendations and future work.....	76
6.	References	78
7.	Appendices	90
	Appendix A	90
	Appendix B.....	94
	Appendix C.....	95
	Appendix D	97

List of Figures

FIGURE 1: LANDSCAPES OF KRUGER NATIONAL PARK AS ORIGINALLY DEFINED BY GERTENBACH (1983) THAT DIVIDED THE PARK INTO 35 LAND TYPES.....	4
FIGURE 2: RECLASSIFICATION OF THE 35 LAND TYPES OF GERTENBACH (1983) INTO 56 LAND TYPES BY VENTER (1990).	5
FIGURE 3: POSITIVE <i>BACILLUS ANTHRACIS</i> CASES DIAGNOSED IN THE KRUGER NATIONAL PARK FROM 1988 TO 2011. NOTE THAT SOME DOTS CAN REPRESENT MULTIPLE CASES.	6
FIGURE 4: FRAMEWORK THAT ILLUSTRATES THE PROCESSES INVOLVED IN SDM. SOURCES OF UNCERTAINTY AND DECISION STEPS IN CHOOSING DATA AND METHODS TO MATCH MODELLING OBJECTIVES ARE SHOWN (FRANKLIN, 2009).....	9
FIGURE 5: PRINCIPLE OF PARSIMONY. THE BEST MODELS HAVE A NUMBER OF PARAMETERS THAT ARE CLOSE TO THE INTERSECTING LINES (BURNHAM, 2001).	11
FIGURE 6: EXAMPLE OF A ROC GRAPH INDICATING THE SENSITIVITY AND SPECIFICITY.	14
FIGURE 7: MAP OF THE PAFURI REGION IN KNP. NOTE THE VARIOUS DRAINAGE CHANNELS INTO THE LUVUVHU RIVER AND LOWER LYING NORTHERN DEPRESSION.	20
FIGURE 8: POSITIVE ANTHRAX CASES DURING 1950-1959 OUTBREAK IN KRUGER NATIONAL PARK AS INDICATED BY PIENAAR (1960).	28
FIGURE 9: POSITIVE ANTHRAX CASES DURING 1959 ANTHRAX OUTBREAK IN KRUGER NATIONAL PARK AS INDICATED BY PIENAAR (1961).	29
FIGURE 10: PRESENTATION OF THE DIFFERENT INDICES (THE SLOPES OF INCREASE (SPRING) AND DECREASE (AUTUMN), THE MAXIMUM NDVI VALUE, THE INTEGRATED NDVI (INDVI, I.E. THE SUM OF NDVI VALUES OVER A YEAR), THE DATE WHEN THE MAXIMUM NDVI VALUE OCCURS, THE RANGE OF ANNUAL NDVI VALUES, AND THE DATE OF GREEN-UP (I.E. THE BEGINNING OF THE GROWING SEASON)) THAT COULD BE DERIVED FROM NDVI TIME SERIES OVER A YEAR. IMAGE ADAPTED FROM PETORELLI <i>ET AL.</i> (2005).....	36
FIGURE 11: NDVI VALUES (RANGING FROM 0 TO 1) FOR KRUGER NATIONAL PARK IN (A) NOVEMBER – PERIOD OF INCREASE, (B) JANUARY – MAXIMUM NDVI AND (C) MAY 2008 – PERIOD OF DECREASE.	37
FIGURE 12: BACKGROUND BIAS LAYER CREATION IN MAXENT. THE PURPLE MAP ON THE RIGHT INDICATES THE BIAS LAYER WHICH MAXENT USED TO CREATE BACKGROUND POINTS FROM.	43
FIGURE 13: VARIABLE SELECTION PROCEDURE INDICATING THE LOSS OF GAIN AND ASSOCIATED DECREASE IN AUC VALUE WHEN VARIABLES ARE REMOVED.	50
FIGURE 14: SUITABLE ANTHRAX AREAS WITHIN KRUGER NATIONAL PARK. BLACK RECTANGLES INDICATE CORE AREAS (GREATER THAN 80% PROBABILITY).....	52
FIGURE 15: JACKKNIFE RESULTS OF THE TWELVE VARIABLES USED TO CONSTRUCT THE FINAL MODEL.	54
FIGURE 16: PROBABILITY OF PRESENCE OF THE DIFFERENT CLASSES THAT MAKE UP THE SOTERSOILID VARIABLE. RED BARS INDICATE THE AVERAGE VALUE OVER THE 10 MODEL RUNS, DARK BLUE BARS INDICATE THE MAXIMUM VALUES AND THE TURQUOISE BARS INDICATE THE MINIMUM VALUES.	55
FIGURE 17: INDVI VALUES FOR THE 10 MODEL RUNS. RED LINE INDICATES AVERAGE AND BLUE AREAS ARE DEVIATIONS.....	56
FIGURE 18: 1959 OUTBREAK DATA PLOTTED AGAINST MODELLED SUITABILITY MAP. A. FIRST OUTBREAK OF 1959 (PIENAAR, 1960) B. SECOND OUTBREAK OF 1959 (PIENAAR, 1961) C. PRESENCE POINTS USED TO TRAIN THE MODEL (1988 - 2011)	58
FIGURE 19: TWO MODELS WERE DEVELOPED IN THIS STUDY – A PRELIMINARY 40 VARIABLE MODEL AND A FINAL 12 VARIABLE MODEL, 5000 ITERATIONS, WITH ALL PRESENCE POINTS.	59
FIGURE 20: GAP ANALYSES ON MAXENT MODEL OUTPUT. RED AREAS INDICATE AREAS WHERE NO PRESENCE POINTS WERE USED AS TRAINING DATA, BUT THAT HAVE A VERY HIGH LIKELIHOOD OF PRODUCING POSITIVE CASES. THUS RED INDICATES HIGHEST SAMPLING PRIORITY, FOLLOWED BY ORANGE AND LASTLY YELLOW.....	63
FIGURE 21: SUITABLE ANTHRAX AREAS WITHIN KRUGER NATIONAL PARK. BLACK RECTANGLES INDICATE CORE AREAS (GREATER THAN 80% PROBABILITY). GREY RECTANGLES INDICATE NOTABLE AREAS (60% - 80% PROBABILITY).	66
FIGURE 22: PAFURI REGION SUITABILITY MAP FOR <i>BACILLUS ANTHRACIS</i> . THIS MAP DEPICTS THE PREDICTED AREA IN MORE DETAIL.	67
FIGURE 23: SHINGWEDZI REGION SUITABILITY MAP FOR <i>BACILLUS ANTHRACIS</i> . THIS MAP DEPICTS THE PREDICTED AREA IN MORE DETAIL.	69

FIGURE 24: LETABA-OLIFANTS REGION SUITABILITY MAP FOR <i>BACILLUS ANTHRACIS</i> . THIS MAP DEPICTS THE PREDICTED AREA IN MORE DETAIL.	70
FIGURE 25: KINGFISHERSPRUIT REGION SUITABILITY MAP FOR <i>BACILLUS ANTHRACIS</i> . THIS MAP DEPICTS THE PREDICTED AREA IN MORE DETAIL.	72
FIGURE 26: CROCODILE BRIDGE REGION SUITABILITY MAP FOR <i>BACILLUS ANTHRACIS</i> . THIS MAP DEPICTS THE PREDICTED AREA IN MORE DETAIL.	73
FIGURE 27: CREATION OF A PERSONAL GEODATABASE IN ARCCATALOG TO ENABLE ADDITION OF INDIVIDUAL RECORDS OF ANTHRAX CASES.	97
FIGURE 28: MAXENT GUI, MAIN SCREEN INDICATING FEATURES SELECTED.	97
FIGURE 29: BASIC OPTIONS SELECTED IN MAXENT. NOTE THE 25 RANDOM TEST PERCENTAGE.....	98
FIGURE 30: ADVANCED OPTIONS IN MAXENT. NOTE THE NUMBER OF MAXIMUM ITERATIONS TO ENSURE CONVERGENCE.	98
FIGURE 31: NULL MODEL CREATION WITH ENMTOOLS.....	99
FIGURE 32: MAXENT MODEL SURVEYOR OUTPUT. TWO VARIABLE SETS WITH OPTIMAL AUC ARE DISPLAYED CONTAINING 10 AND 11 VARIABLES RESPECTIVELY.	99

List of Tables

TABLE 1: CURRENT MODELLING TECHNIQUES USED IN SDM, INCLUDING TYPE OF DATA REQUIRED AND THE TECHNIQUE REFERENCE.	7
TABLE 2: SOME PUBLISHED METHODS FOR OCCURRENCE THRESHOLD SELECTION (PEARSON, 2007). FOR THE DIFFERENT THRESHOLD VALUES APPLIED TO THE FINAL MODEL OUTPUT, SEE APPENDIX B.	12
TABLE 3: OVERVIEW OF ENVIRONMENTAL DATA USED IN MAXENT INDICATING THE VARIABLES, TYPE OF DATA, SOURCE, SPATIAL RESOLUTION AND REFERENCES.	30
TABLE 4: NUMBER OF CLASSES PER SELECTED CATEGORICAL VARIABLE.	35
TABLE 5: HYDROLOGICAL INDICES AND RIVER CLASSES USED IN THIS STUDY (HANNART AND HUGHES, 2003). A HYDROLOGICAL INDEX LESS THAN 16.110 INDICATES A PERMANENT RIVER. A HYDROLOGICAL INDEX BETWEEN 16.110 AND 37.81 INDICATES A SEASONAL RIVER AND A HYDROLOGICAL INDEX GREATER THAN 37.81 INDICATES AN EPHEMERAL RIVER. ...	39
TABLE 6: BEST SUBSET MMS PROCEDURE. VARIABLES ARE ADDED TO THE MAXENT VARIABLE SET UNTIL THE AUC VALUE FOR ALL POSSIBLE COMBINATIONS HAVE BEEN CALCULATED.	45
TABLE 7: SIMPLE MULTIPLE REGRESSION ON CONTINUOUS VARIABLES.	48
TABLE 8: SIMPLE MULTIPLE REGRESSION ON CATEGORICAL VARIABLES.	48
TABLE 9: VARIABLE SELECTION PROCEDURE. THE VARIABLE REMOVED FROM THE MODEL BASED ON THE JACKKNIFE METHOD IS LISTED IN THE RIGHT COLUMN. MODEL GAIN AND AUC VALUES ARE DISPLAYED IN COLUMNS 2 AND 3 TO ILLUSTRATE THE CHANGE (DECREASE) IN VALUE AS MORE VARIABLES ARE REMOVED.	49
TABLE 10: VARIABLE CONTRIBUTION TABLE INCLUDING PERCENT CONTRIBUTION AND PERMUTATION IMPORTANCE OF THE VARIABLE LISTED.	53
TABLE 11: SOTERSOILID CLASS AND NAME AS LISTED IN THE SOTER DATABASE.	55
TABLE 12: THRESHOLD POINT VALUES FOR THE TOP 12 VARIABLES AND THE THREE AREAS IDENTIFIED BY THE MODEL (FIGURE 14) AS HAVING THE HIGHEST SUITABILITY FOR ANTHRAX OCCURRENCE.	60
TABLE 13: SELECTED PARAMETERS FROM SOIL SAMPLE POINTS IN KRUGER NATIONAL PARK INDICATING ANTHRAX PRESENCE OR ABSENCE (DIJKSHOORN <i>ET AL.</i> , 2008).	65
TABLE 14: SELECTED SOIL PROPERTIES OF PAFURI LAND TYPE (PA05) WHERE ANTHRAX IS ENDEMIC COMPARED WITH ANTHRAX UNFAVOURABLE PRETORIUSKOP LAND TYPE (SK01) (VENTER, 1990)	68
TABLE 15: SELECTED SOIL PROPERTIES OF SHINGWEDZI LAND TYPE (LE05) FAVOURABLE FOR ANTHRAX COMPARED TO UNFAVOURABLE ANTHRAX PRETORIUSKOP LAND TYPE (SK01) (VENTER, 1990)	69
TABLE 16: SELECTED SOIL PROPERTIES OF LETABA LAND TYPE (LE02) FAVOURABLE FOR ANTHRAX COMPARED TO UNFAVOURABLE ANTHRAX PRETORIUSKOP LAND TYPE (SK01) (VENTER, 1990)	71
TABLE 17: SELECTED SOIL PROPERTIES OF ORPEN LAND TYPE (SA05) SUITABLE ECOLOGICAL CONDITIONS FOR ANTHRAX COMPARED TO UNFAVOURABLE ANTHRAX PRETORIUSKOP LAND TYPE (SK01) (VENTER, 1990)	72
TABLE 18: SELECTED SOIL PROPERTIES OF SATARA LAND TYPE (SA01) SUITABLE ECOLOGICAL CONDITIONS FOR ANTHRAX COMPARED TO UNFAVOURABLE ANTHRAX PRETORIUSKOP LAND TYPE (SK01) (VENTER, 1990)	74

List of Appendices

Appendix A.....	90
Appendix B.....	94
Appendix C.....	95
Appendix D.....	97

Summary

Ecological suitability modelling for anthrax in the Kruger National Park, South Africa

by

Pieter Johannes Steenkamp

Supervisor: Dr. O.L. van Schalkwyk
Co-Supervisor: Dr. H. van Heerden
Department: Production Animal Studies
Degree: MMedVet(Fer)

Bacillus anthracis is the causal agent of anthrax which primarily affects ungulates, occasionally carnivores and less frequently humans. The endospores of this soil-borne bacterium are highly resistant to extreme conditions, and under ideal conditions, anthrax spores can survive for many years in the soil. The bacterium is generally found in soil at sites where infected animals have died. When these spores are exposed, they have the potential to be ingested by a mammalian species which could lead to an anthrax outbreak. Anthrax is almost never transmitted directly from host to host, but is rather ingested by herbivores while drinking, grazing or browsing in a contaminated environment, with the exception of scavengers and carnivores consuming infected prey. Anthrax is known to be endemic in the northern part of Kruger National Park (KNP) in South Africa (SA), with occasional epidemics spreading southward into the non-endemic areas.

The aim of this study is to identify and map areas that are ecologically suitable for the harbouring of *B. anthracis* spores within the KNP. Anthrax surveillance data and selected environmental variables were used as inputs to the maximum entropy (Maxent) species distribution modelling method.

Five-hundred and ninety-seven anthrax occurrence records, dating from the year 1988 to 2011, were extracted from the Skukuza State Veterinary Office's database. A total of 40 environmental variables were used and their relative contribution to predicting suitability for anthrax occurrence was evaluated using Maxent software (version 3.3.3k). Variables showing the highest gain were then used for subsequent, refined model iterations until the final model parameters were established.

The environmental variables that contributed the most to the occurrence of anthrax were soil type, normalized difference vegetation index (NDVI), land type and precipitation. A map was created using a geographic information system (GIS) that illustrates the sites where anthrax spores are most likely to occur throughout the Park. This included the known endemic Pafuri region as well as the low lying soils along the Shingwedzi-Phugwane-Bubube rivers and the Letaba-Olifants river drainage area.

The outputs of this study could guide future targeted surveillance efforts to focus on areas predicted to be highly suitable for anthrax, especially since the KNP uses passive surveillance to detect anthrax outbreaks. Knowing where to look can improve sampling efficiency and lead to increased understanding of the ecology of anthrax within the KNP.

1. Introduction

1.1. Research Background

1.1.1. Context

Anthrax has been present in the KNP, South Africa (SA) for centuries and is considered a natural part of the ecosystem. One area within the Park, the northern Pafuri region, has been described as an anthrax endemic area, with outbreaks originating and spreading from this location (De Vos, 1990). The disease cycle of anthrax involves host species, predators, scavengers, insects, water, soil and various environmental factors. A multitude of the above requirements must be met for an anthrax outbreak to occur. An even more stringent set of requirements is needed for the spores to be able to survive in the soil. The ecology of anthrax within the Park will be discussed in more detail to enable a better understanding of the environmental factors necessary to sustain the bacterium and how this pertains to the data selection for modelling.

Species distribution modelling (SDM) is a process by which the potential distribution of a species is mapped as a function of its ecological niche. A species' fundamental niche is the set of all conditions that allow for the species' long term survival while the realized niche is that subset of the fundamental niche that the species actually occupies (Hutchinson, 1957). Environmental conditions at the occurrence locations constitute samples from the realized niche. Thus, a certain species is very likely to occur in a given area if all the environmental conditions favour its sustained survival.

The data used in this study were collected throughout the KNP over an extended period and consist of confirmed positive anthrax cases. These data were considered presence-only data, since only anthrax positive cases were documented. The available environment, referred to as the background, can be determined by a random sample, random stratified sample, all non-presence locations, or all locations within the study area (Franklin, 2009).

Theoretically, environments with a high anthrax incidence should have a high potential to harbour *B. anthracis* spores. The objective of this study was to identify all possible variables that play a role in the distribution and survival of anthrax spores and to use these variables to develop a suitability map for spore occurrence.

1.1.2. Study Area

The study area includes the KNP in the Mpumalanga and Limpopo provinces of SA. The KNP was formally established in 1926 and is an elongated conservation area in the north-eastern corner of SA, encompassing almost two million hectares of subtropical savannah woodland (Braack and Teske, 1997). The Park stretches from 20°19'S to 25°32'S and 31°01'E to 32°02' E. This Park is SA's largest conservation area and includes the following number of species: 336 trees, 49 fish, 34 amphibians, 114 reptiles, 507 birds and 147 mammals with an approximate 3700 kg biomass of free-ranging animals per 100 ha (Braack and Teske, 1997).

The topography of the Park reflects differences in weathering and dissection intensity of the underlying rock, especially in areas that flank major rivers. The KNP lies on average 300 m above sea level and is divided into two climate zones as defined by the South African weather service. The south and central areas of the Park fall within the lowveld bushveld zone and have an average rainfall of 500-700 mm per year. The northern part of KNP falls within the northern arid bushveld zone with an average rainfall of 300-500 mm per year. The temperature also varies between the two zones with the southern zone being cooler than the north. Soil profiles generally become shallower as rainfall decreases toward the north of the Park. Wet and dry cycles occur that significantly influence animal population dynamics (De Vos, 1990). During dry periods, plains game such as zebra and wildebeest increase in number and long grass feeders such as buffalo and roan antelope decrease in number (Venter *et al.*, 2003).

The soil diversity and topographical features such as hills, valleys and plains in KNP harbour distinct vegetation types that in turn determine the distribution and abundance of animals. Gertenbach (1983) classified the diverse landscape of KNP into 35 types (Figure

1). Venter (1990) reclassified Gertenbach's system into 11 land systems and 56 land types (Figure 2). Each land system consists of between one and 12 different land types (Venter, 1990).

Anthrax mortalities occur most frequently in the northern part of KNP and spread southwards (Figure 3) with the Pafuri area (most northern region of KNP) deemed endemic to the disease (De Vos, 1990).

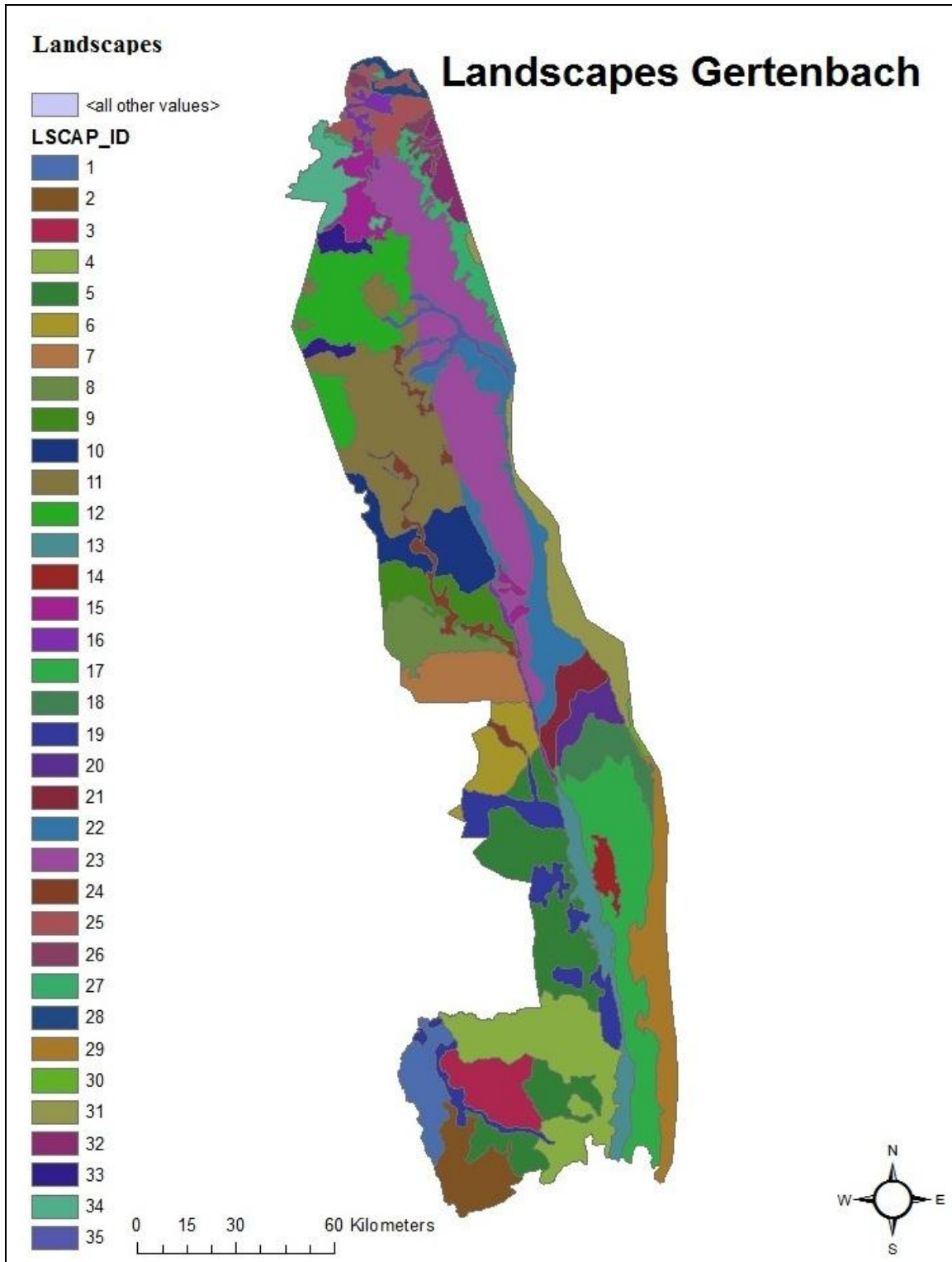


Figure 1: Landscapes of Kruger National Park as originally defined by Gertenbach (1983) that divided the Park into 35 land types.

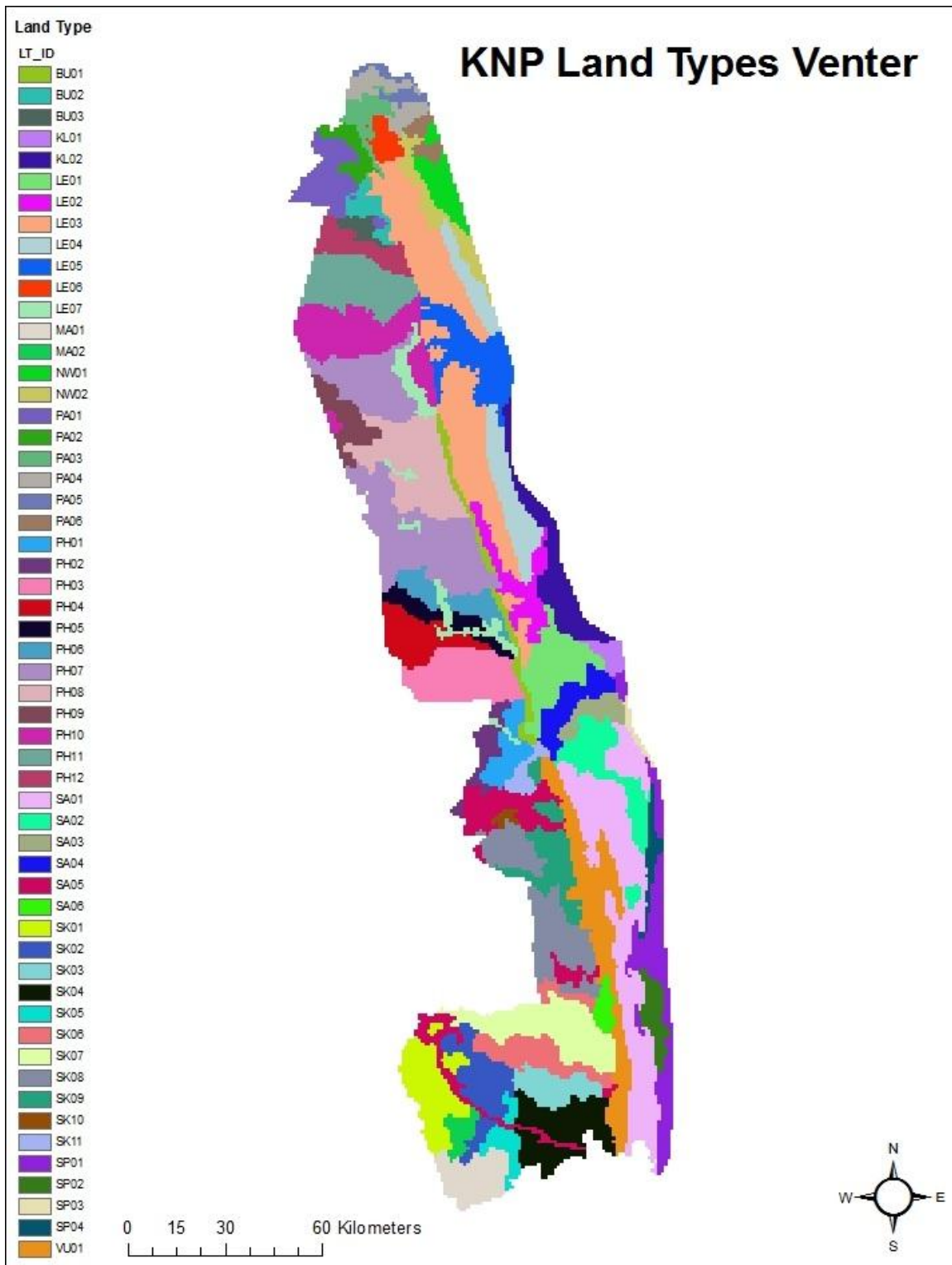


Figure 2: Reclassification of the 35 land types of Gertenbach (1983) into 56 land types by Venter (1990).

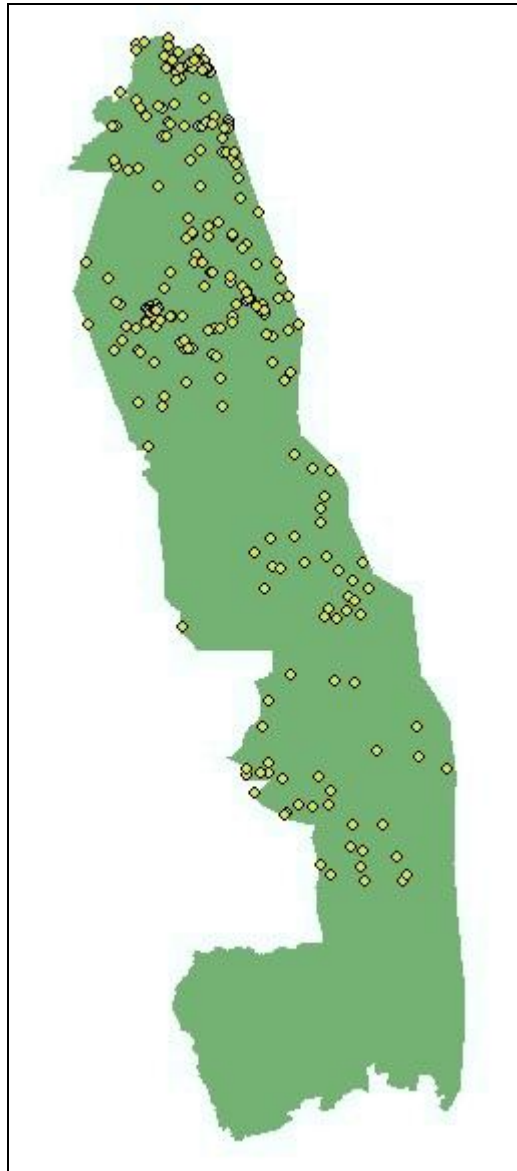


Figure 3: Positive *Bacillus anthracis* cases diagnosed in the Kruger National Park from 1988 to 2011. Note that some dots can represent multiple cases.

1.1.3. Species Distribution Modelling (SDM)

SDM is the subset of mathematical ecology techniques attempting to quantify the spatial and or temporal relationship of a species with the environment in which it occurs. The earliest attempts to quantify this relationship were during the late 1970's by Nix *et al.* (1977). Today, various tools and methods exist for the predictive modelling of species' environmental requirements and geographic distributions. SDM has been used in conservation planning, ecology, evolution, epidemiology, invasive species management and other fields (Phillips *et al.*, 2006).

The type and quality of data available generally determines which modelling method is going to produce the best result. Data that were collected without any specific sampling strategy are often associated with errors and biases, reflecting the haphazardness of the collection method (Hijmans *et al.*, 2000; Reese *et al.*, 2005; Elith *et al.*, 2006).

Furthermore, data can be divided into presence-only or presence-absence. Presence-only data are a set of records of observed or classified presences of the research species. Absence data are a set of records of where the research species was not found within the extent of the study area. Absence is difficult to determine, because the species involved might be present, but just not observed.

The need for having a modelling approach that handles presence-only data stems from incomplete historical species records in museums leading to the development of the following modelling techniques during the last number of years (Table 1):

Table 1: Current modelling techniques used in SDM, including type of data required and the technique reference.

Technique	Description	Type of data	Reference
BIOCLIM	Predicts suitable conditions in a 'bioclimatic envelope', consisting of the range of observed presence values in each environmental dimension. This envelope specifies the model in terms of percentiles or upper and lower tolerances.	Presence-only	Busby, 1986; Nix, 1986
DOMAIN	Gives a predicted suitability index by computing the minimum distance in environmental space to any	Presence-only	Carpenter <i>et al.</i> , 1993

	presence record. Resulting predictions range between 0 and 100.		
LIVES	Limiting factor method that postulates that the occurrence of a species is determined only by the environmental factor that most limits its distribution. The limiting factor of the species is defined as the environmental factor that has the minimum similarity among the environmental factors considered in the model.	Presence-only	Carpenter <i>et al.</i> , 1993
Generalised Linear models (GLMs)	Regression based. Uses occurrence and background data as dependent variables and environmental data as independent variables.	Presence, background	Graham <i>et al.</i> , 2008
Generalised Additive models (GAMs)	Regression based. Similar to GLM, but uses non-parametric, data-defined smoothers to fit non-linear functions. Considered more capable (than GLMs) to model complex ecological response shapes.	Presence, background	Yee and Mitchell, 1991; Guisan <i>et al.</i> , 2002; Wintle <i>et al.</i> , 2005; Graham <i>et al.</i> , 2008
Multivariate adaptive regression splines (MARS)	Regression based. Uses piecewise linear fits rather than smooth functions and a fitting procedure that makes them much faster to implement than GAMs.	Presence, background	Leathwick <i>et al.</i> , 2005; Elith <i>et al.</i> , 2006; Graham <i>et al.</i> , 2008
Genetic Algorithm for Rule Set Prediction (GARP)	Uses a set of rules that together gives a binary prediction – rules are prioritized according to their significance based on a sample of background and presence data.	Presence, background	Stockwell and Noble, 1992; Stockwell and Peters, 1999
Maximum Entropy (Maxent)	Maximum entropy approach to determine the distribution that is closest to uniform using a set of presence-only data, background data and environmental variables.	Presence, background	Phillips <i>et al.</i> , 2006
Boosted regression trees (BRT)	Combines two algorithms: the boosting algorithm iteratively calls the regression-tree algorithm to construct a combination or ‘ensemble’ of trees. The regression trees are fitted sequentially, and use a gradient descent algorithm to model iteratively the residuals that reflect the lack of fit from the previous set of trees.	Presence, background	Schapire, 2003; Elith <i>et al.</i> , 2006
Environmental Niche Factor Analysis (ENFA)	Analyses presence data and environmental data for the entire study area and transforms it into different factors. Environmental suitability is then modelled in the transformed space	Presence, background	Hirzel <i>et al.</i> , 2002

The above table is by no means a complete list of all the techniques implemented today. The reader is referred to Elith *et al.* (2006) and Austin (2007) who provides a very comprehensive comparison of current modelling techniques.

Figure 4 illustrates the processes involved in SDM. Maxent displayed promising results compared to other methods when dealing with presence-only data and small sample sizes

(Phillips *et al.*, 2004; Elith *et al.*, 2006; Herkt, 2007; Baldwin, 2009). It can also handle categorical data and sample bias, while over-fitting can be avoided (Phillips *et al.*, 2004).

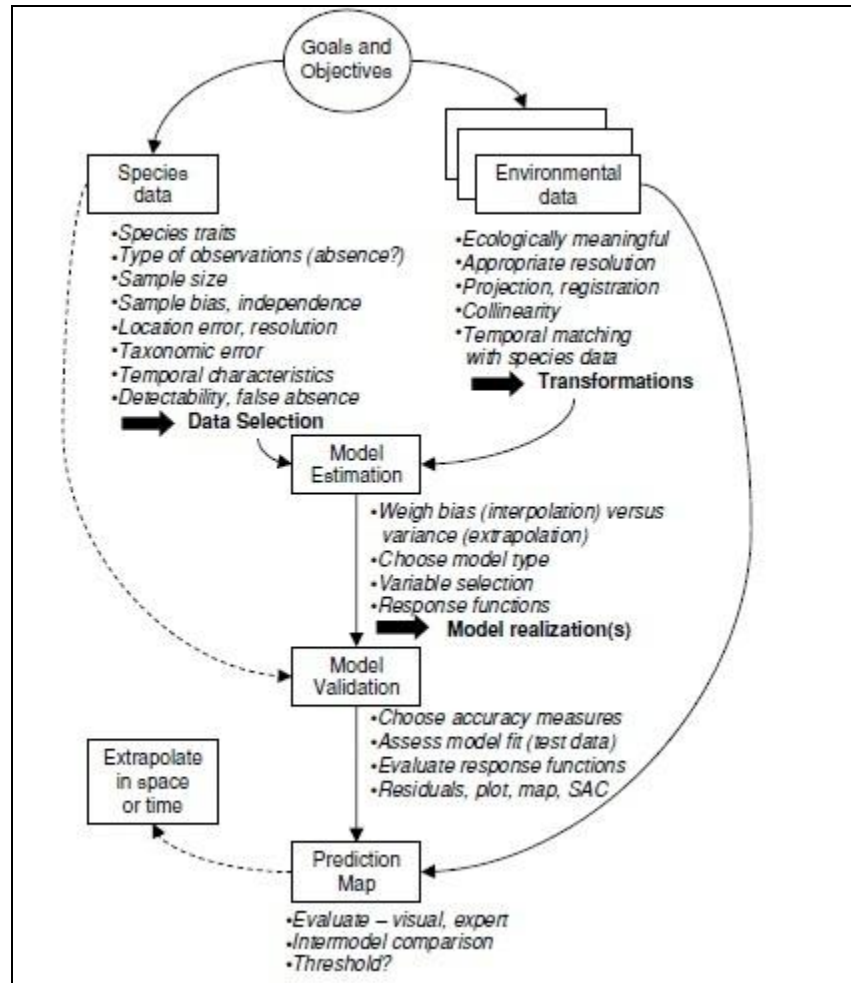


Figure 4: Framework that illustrates the processes involved in SDM. Sources of uncertainty and decision steps in choosing data and methods to match modelling objectives are shown (Franklin, 2009).

Spatial autocorrelation (SAC) is a statistical property of most ecological variables and represents the relationship between values of the given variable at different geographical separations (Legendre, 1993; Naimi *et al.*, 2011). Spatial data is said to exhibit SAC when values measured nearby in space are more similar than values measured farther away from each other. SAC can occur with the collection of specimens from several nearby localities in a certain area (Phillips *et al.*, 2006). The robustness of a SDM to species positional uncertainty is affected positively by SAC in environmental variables

(Naimi *et al.*, 2011). SAC statistics measure and analyse the degree of dependency among observations in a geographic space. On average, the closer together two locations are, the more similar their measures of species abundances or occurrences.

If the SAC pattern remains present in the residuals (resulting probability distribution) of a statistical model based on such data, one of the key assumptions of standard statistical analyses, that residuals are independent and identically distributed, is violated. The violation of the assumption of independent and identically distributed residuals may bias parameter estimates (Dormann *et al.*, 2007). The environmental layers in Maxent are converted into ‘features’ and they are used to constrain the resulting model residuals. Therefore, Maxent inherently deals with, and is not sensitive to SAC (Cheng, 2007).

The benefit of including SAC in a model is that the values of neighbours are incorporated, which ultimately improves the predictive power of the model (Costanza and Ruth, 2001). In addition, spatial models that include SAC may improve variable selection (Eller and Seifu, 2002; Keitt *et al.*, 2002).

Deciding on how many parameters a model should have depends on the ecology of the organism involved and the research objectives (Figure 5). Models with too few parameters have biased predictions, whereas models with too many variables have poor precision (Franklin, 2009).

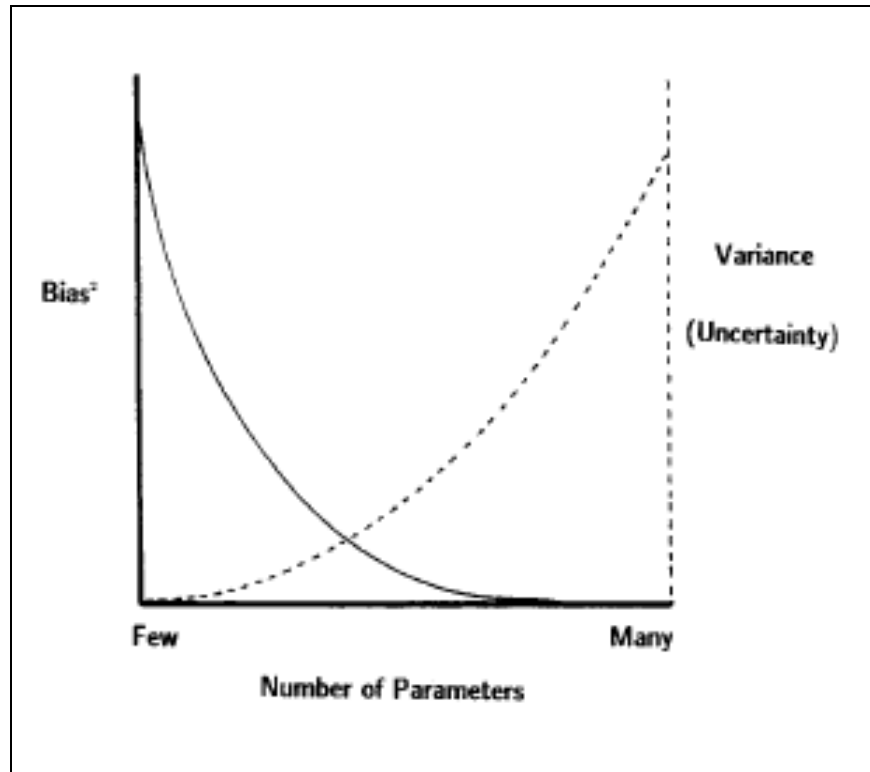


Figure 5: Principle of parsimony. The best models have a number of parameters that are close to the intersecting lines (Burnham, 2001).

To provide informative predictions, it is necessary for a model to successfully predict a high proportion of test localities (i.e. have a low omission rate) whilst not predicting as suitable such a large proportion of the study area as to make the model statistically indistinguishable from a random prediction (Anderson *et al.*, 2002). The model needs to be validated to determine if the results are in fact better than random. In SDM, quantifying prediction accuracy is used as a measure of model performance or validity (Franklin, 2009). The first step is to verify that the model performed better than random (Phillips *et al.*, 2006). Two methods have been used to accomplish this, namely receiver operating characteristic (ROC) plots and defined thresholds.

For this study, verification that the model performs significantly better than random was done by performing:

- binomial test based on omission and predicted area;
- AUC (ROC) analyses;

- null model statistics.

Threshold-dependent measures

A threshold is a value that is determined by the model creator and signifies the probability value above which species presence is assumed. These thresholds are established by maximizing sensitivity while minimizing specificity. To aid model validation and interpretation, it is usually desirable to distinguish ‘suitable’ from ‘unsuitable’ areas by setting a decision threshold. Above this decision threshold, model output is considered to be a prediction of presence (Pearson *et al.*, 2004). A number of different methods have been employed for selecting thresholds (Table 2) (Pearson, 2007). The choice of an appropriate threshold is dependent on the type of data that are available and the question that is being addressed. This study dealt with presence-only data and the suitability of the environment for the long-term survival of *B. anthracis* spores. The simplest approach for threshold selection is to use an arbitrary value, even though this method is subjective (Liu *et al.*, 2005).

Table 2: Some published methods for occurrence threshold selection (Pearson, 2007). For the different threshold values applied to the final model output, see Appendix B.

Method	Definition	Species data type	Reference(s)
Fixed value	An arbitrary fixed value (e.g. probability = 0.8)	presence-only	Manel <i>et al.</i> , 1999; Robertson <i>et al.</i> , 2001
Lowest predicted value	The lowest predicted value corresponding with an observed occurrence record	presence-only	Pearson <i>et al.</i> , 2006; Phillips <i>et al.</i> , 2006
Fixed sensitivity	The threshold at which an arbitrary fixed sensitivity is reached (e.g. 0.90, meaning that 90% of observed localities will be included in the prediction)	presence-only	Pearson <i>et al.</i> , 2004
Sensitivity-specificity equality	The threshold at which sensitivity and specificity are equal	presence and absence	Pearson <i>et al.</i> , 2004
Sensitivity-specificity sum	The sum of sensitivity and specificity is maximized	presence and absence	Manel <i>et al.</i> , 2001

maximization			
Maximize Kappa	The threshold at which Cohen's Kappa statistic is maximized	presence and absence	Huntley <i>et al.</i> 1995; Elith <i>et al.</i> , 2006
Average probability/suitability	The mean value across model output	presence-only	Cramer, 2003
Equal prevalence	Species' prevalence (the proportion of presences relative to the number of sites) is maintained the same in the prediction as in the calibration data.	presence and absence	Cramer, 2003

After applying a threshold, model performance can be evaluated using the extrinsic omission rate and proportional predicted area. The extrinsic omission rate is the fraction of test locations that was predicted as unsuitable for the species and the proportional predicted area is the fraction of all locations predicted as suitable for the species. A one-tailed binomial test is used to determine whether a model predicts the test locations significantly better than random ($p < 0.05$) (Phillips *et al.*, 2006).

Threshold-independent measures

Area under the curve (AUC)

The most widely used evaluation method in SDM is the AUC of the receiver operating characteristic (ROC) curve (Lobo *et al.*, 2007; Hijmans *et al.*, 2011). It measures the probability that the model will assign a higher probability of occurrence to the observed presences (Merckx *et al.*, 2011). Maxent develops a ROC plot for AUC evaluation automatically. A good model is defined by a curve that maximizes sensitivity for low values of the false-positive fraction (Baldwin, 2009). AUC provides a ranked approach for prediction accuracy compared to a random distribution. This method assigns a single number to the performance of a model (Hanley and McNeil, 1982; Phillips *et al.*, 2006; Baldwin, 2009).

The ROC plot is a plot of sensitivity and one minus specificity, with sensitivity representing how well the data correctly predicts presence, whereas specificity provides a

measure of correctly predicted absences (Fielding and Bell, 1997; Baldwin, 2009) (Figure 6). The AUC is calculated by summing the area under the ROC graph. The more the value of the ROC graph tends toward a specificity and sensitivity of 1, the better the model is. If the ROC graph follows or is close to the diagonal line, the model predictions are no better than random.

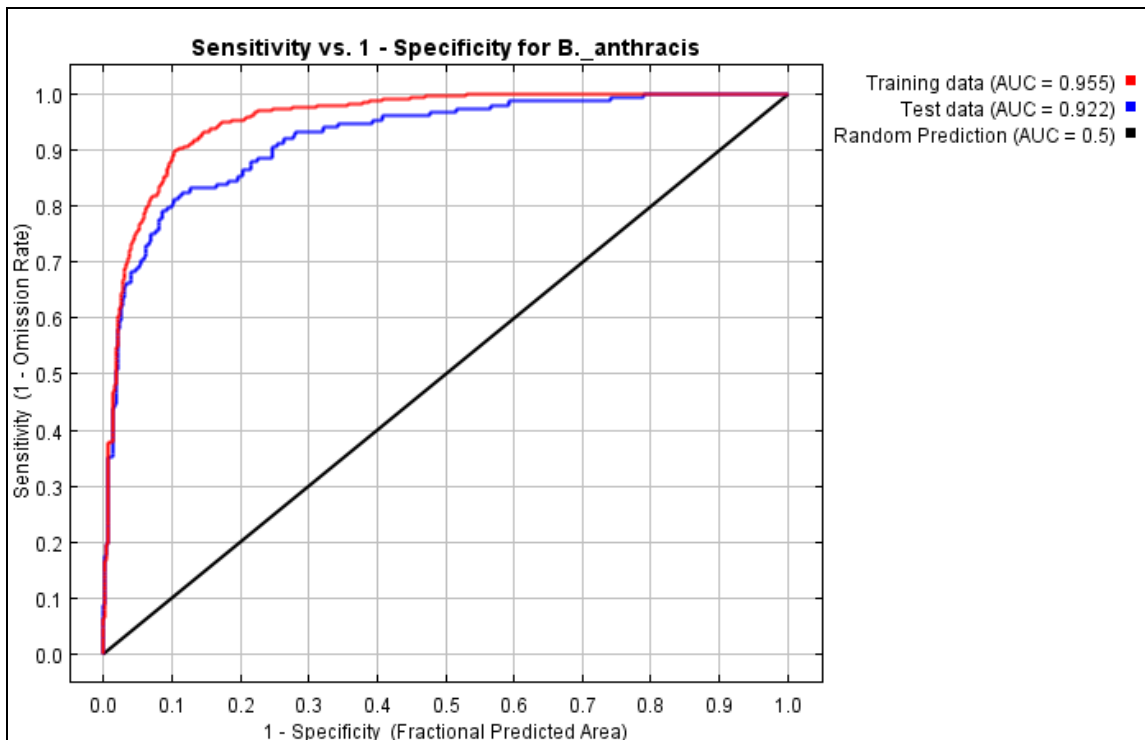


Figure 6: Example of a ROC graph indicating the sensitivity and specificity.

Successful models have AUC scores approaching 1.0 and models predicting no better than random will have an AUC approaching 0.5 (Blackburn *et al.*, 2007; Guo and Lui, 2010). Araújo and Guisan (2006) defined a rough guide for classifying model accuracy: 0.6–0.7 poor, 0.7–0.8, average, 0.8–0.9 good, 0.9–1 excellent. The AUC of a classifier is thus the probability that it will rank a randomly chosen positive instance higher than a randomly chosen negative one, making it equivalent to the Wilcoxon test of ranks (Fawcett, 2006). AUC (ROC) analysis is independent of both threshold setting and prevalence, making it a very effective method for model evaluation when working with presence-only data (Allouche *et al.*, 2006). However, to use ROC curves with presence-only data, one must interpret all grid cells with no occurrence localities as “negative examples”, even if they support good environmental conditions for the species. The

maximum AUC is therefore less than one, and is smaller for wider-ranging species (Phillips *et al.*, 2004).

Null Model

With this testing method, various random pseudo-presence sets are generated throughout the extent of the model area. The same number of presence records as the original model is generated for each set. Each of these datasets is used as the presence data for independent Maxent models and the resulting AUCs are compared to the AUC of the model to be tested. One-sided 95% confidence intervals (CI's) are used to test for significance. If the AUC of the model falls within the top five percent of all AUCs, it is considered statistically significant and the model is predicting better than random (Raes and Ter Steege, 2007; Merckx *et al.*, 2011).

1.1.4. Maximum Entropy (Maxent) modelling

Maxent modelling is a general purpose machine learning modelling method with a simple and precise mathematical formulation, designed to make decisions from incomplete data (Baldwin, 2009; Phillips *et al.*, 2006). Machine learning involves a number of advanced statistical methods that handle regression and classification tasks with multiple dependent and independent variables (Hill and Lewicki, 2007). The idea of Maxent is to estimate a target probability distribution of sampling points compared to background locations by finding the probability distribution that is closest to uniform, subject to a set of constraints, that represent the incomplete information about the target distribution (Grendár and Grendár, 2001; Phillips *et al.*, 2004; Phillips *et al.*, 2006; Baldwin, 2009). The algorithm will converge to the maximum entropy probability distribution.

Phillips *et al.* (2006) listed the following as some of the advantages Maxent have over other modelling methods:

- it only requires presence data and environmental information;
- it can utilize both categorical and continuous data;
- it has very efficient algorithms that are guaranteed to converge to the optimal probability distribution;

- it contains a precise mathematical formulation, amenable to analysis;
- over-fitting can be avoided;
- output is continuous and;
- it can also utilize absence data by implementing a conditional model.

Maxent is prone to over-fitting, meaning that the resulting distribution will congregate around provided presence points. Over fitting can be minimized by regularization – a relaxation parameter that allows the average value of each variable to approximate its empirical average but not equal it (Baldwin, 2009). A higher regularization value leads to a wider predicted distribution. This parameter can and should be adjusted according to the sample size and strategy (Phillips *et al.*, 2004; Elith *et al.*, 2011).

Relative variable predictive importance can be measured using the jackknife procedure. The jackknife procedure considers a set of n variables; the gain of each variable on its own is determined and compared to the gain of all the other variables combined, should the former be omitted from the model. This procedure is repeated $n - 1$ times, providing a list of relative variable importance. The gain of a variable can be defined as the sum of the likelihood of the data plus a penalty function (the regularization part). Calculating the gain exponent gives the average ratio of the likelihood assigned to an observed presence location to the likelihood assigned to a background location (Phillips *et al.*, 2006).

Maxent provides continuous outputs in raw, cumulative and logistic formats. The raw output is an exponential function that assigns a probability value to each site and the sum of these values must equal one, making interpretation more complex (Phillips *et al.*, 2006; Baldwin, 2009; Franklin, 2009). Cumulative outputs represent a range from 0-100 in probabilities predicted by the model, again complicating direct interpretation. The logistic format indicates the probability of presence at each site and thus easier and potentially more accurate interpretation. Presences for which not all predictors have a value are assigned NoData values, thereby omitting the presence points (Phillips *et al.*, 2004). A graphical user interface allows the user to customize the model parameters.

Maxent writes the outputs of the model to a specified folder. This folder will contain the following files after the model was successfully executed:

- a hypertext markup language (html) file describing model results. This html file contains an analysis of omission/commission, pictures of the modelled suitability, variable response curves, analysis of variable contributions, raw data outputs, control parameters and links to other files;
- an Ascii file containing the probabilities in raster format;
- an explain tool, provided that the use of product features were disabled since this tool can only be used for additive models. This is a graphical user interface, providing maps and statistics of all contributing variables;
- a text file called maxentResults.csv - listing the number of training samples used for learning, values of training gain and test gain and AUC. Test gain and AUC are given only when a test sample file is provided or when a specified percentage of the samples are set aside for testing. If a jackknife is performed, the regularized training gain and (optionally) test gain and AUC for each part of the jackknife is included;
- a text file called maxent.log - records the parameters and options chosen for the model run, and some details of the model run that are useful for troubleshooting;
- x.lambdas - containing the computed values of the constants c_1, c_2, \dots ;
- x.png - is a picture of the mapped prediction;
- a text file called x_omission.csv - describing the predicted area and training and (optionally) test omission for various raw and cumulative thresholds and;
- various plots for jackknifing and response curves, in the plots subdirectory.

Apart from the above-mentioned program for maximum entropy modelling, various other software packages such as ModEco, R and ENMTools can implement the maximum entropy algorithm for suitability prediction. Due to the high parameter customization ability of Maxent's default GUI, none of the other methods were used.

1.1.5. Anthrax

Anthrax is a rapidly fatal disease caused by the spore-forming bacterium *B. anthracis*. The disease can affect most species, but ruminants are particularly susceptible. Multiple host and environmental factors are thought to play a role in the transmission of anthrax. In domestic species, the primary disease control measures are prophylactic, and consist of breaking the disease cycle through means such as vaccination, treatment and quarantine. Treatment of anthrax is not always a curative measure, but is also used as a way to lessen the bacterial load of infected animals.

Bacillus anthracis is endemic in many parts of Africa, but outbreaks are becoming less frequent in managed species. Free-ranging wildlife experiences more outbreaks than domestic species, simply due to the fact that bacterial transmission and spore survival are harder to regulate. Many of Africa's wildlife reserves experience cyclic anthrax outbreaks, one such area being the KNP in SA.

Despite anthrax being a disease of antiquity, little is known about its spatial ecology or epidemiology (Blackburn *et al.*, 2007). The general thought is that *B. anthracis* is an obligate *in vivo* pathogen and that little propagation occurs in soil (De Vos and Turnbull, 1994). If the environmental conditions are suitable, the bacterium will rapidly form spores once outside the host. Soil pH and soil calcium levels are considered the most important properties for spore survival and, therefore, endemicity of *B. anthracis* is associated with elevated calcium and neutral-to-alkaline soils (Van Ness and Stein, 1956; Van Ness, 1959a; Van Ness, 1959b; Van Ness, 1971; Dragon and Rennie, 1995; Smith *et al.*, 2000). Although these are considered the most important factors for the long-term survival of *B. anthracis* in the soil, other variables such as environmental temperature, rainfall, vegetation, presence of scavengers and mechanical vectors also play a vital role in the spread of the disease.

Septicaemic infection with anthrax causes impaired clotting function (De Vos and Turnbull, 1994). When an animal succumbs to anthrax, the host's impaired clotting ability results in blood draining from any orifice or draining into the soil when carcass is

opened by scavengers. Vegetative anthrax cells are thus exposed to environmental oxygen and begin to sporulate. Depending on the environmental conditions, these spores can survive for decades in the soil until infection of a suitable host takes place.

1.1.6. Anthrax in the Kruger National Park

Anthrax is considered an indigenous and integral part of the KNP ecosystem (De Vos and Turnbull, 1994; Hugh-Jones and Blackburn, 2009). The first confirmed case of anthrax in the KNP was in 1954, but it has been in the northern region of the Park for at least 200 years, as was proven by isolation of spores from archaeological bones dating back to 1700 ± 50 BC (De Vos and Bryden, 1996).

Anthrax outbreaks in KNP appear to have a cyclical pattern of roughly 10 years and are most often associated with a dry climatological spell after a couple years of above average rainfall (De Vos, 1990). The Pafuri region in the far north of the Park is considered endemic for anthrax with periodic outbreaks, typically occurring during late winter and spreading southward. This endemic area is low-lying (Figure 7) and consists of many small pans, which during late winter usually dry up, thus creating an ideal situation for the start of an anthrax cycle. A large number of drainage channels occur into this area from the higher lying southern landscape.

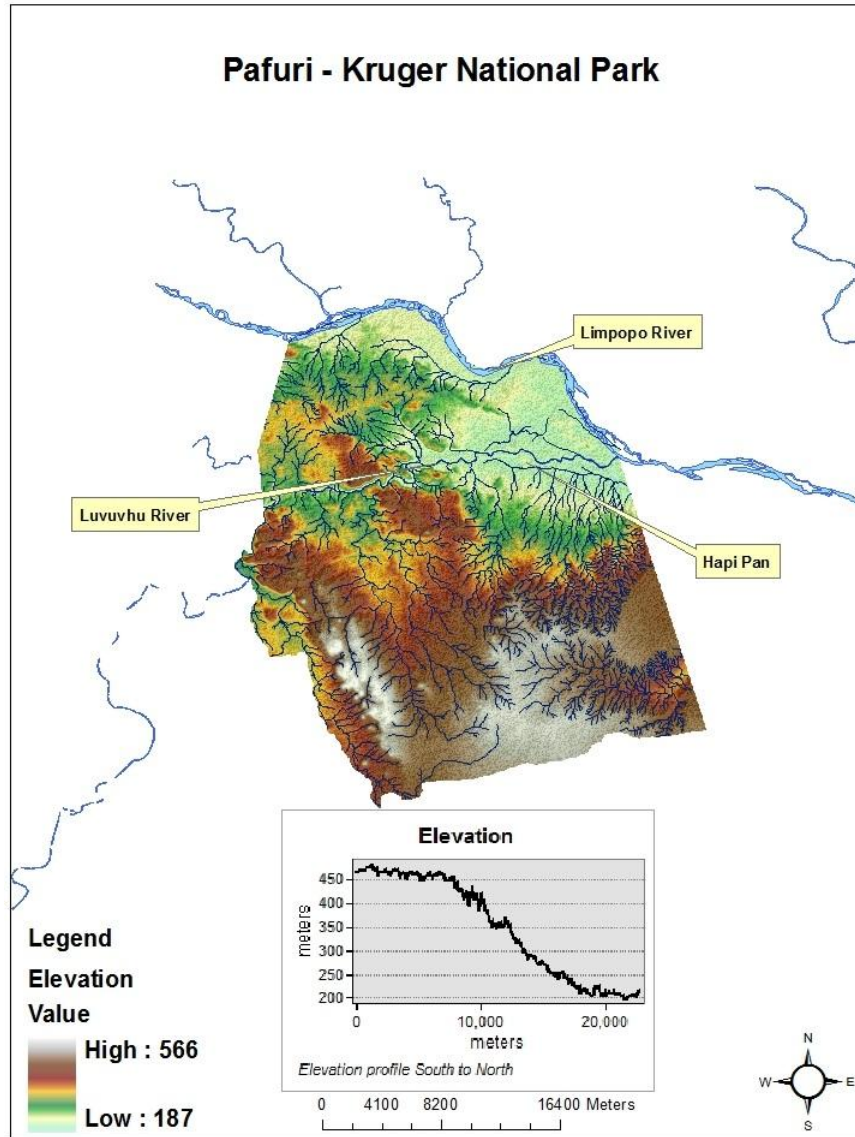


Figure 7: Map of the Pafuri region in KNP. Note the various drainage channels into the Luvuvhu river and lower lying northern depression.

Soil sample analyses from the Pafuri depression clearly indicated that it acts as a catchment and accumulation area for *B. anthracis* spores (De Vos, 1990). The soil depth that anthrax spores are found at, is as shallow as 3 cm during dry, high risk outbreak conditions and as deep as 15 cm during wetter, deposition periods. Deposition periods can be defined as the washing away of spores during heavy rains and deposition in low lying silt beds (De Vos, 1990).

KNP soil profiles becomes shallower and soil type diversity decreases towards the north (Venter *et al.*, 2003). Animals will stir up and ingest spores during the dry season when congregations around water points occur. Once an epidemic starts, the maintenance and spread are determined by biotic factors. Death from anthrax is per-acute to acute and invariably occurs close to the infection site, although buffalo and eland can travel great distances (~30 km) before a new infection locus will be established (De Vos, 1990).

Kudu are especially important in the spread of the disease since their numbers are in direct positive correlation to the amount of rainfall in the Park. If the kudu numbers are high, then an anthrax outbreak can be considered more likely. They feed at a level where infected blow-fly droplets are deposited (1-3 m) on leaves and are gregarious, which means that once an animal in the herd becomes infected, most of the others will invariably become infected too (De Vos, 1990).

Blow-flies will deposit infected droplets onto leaves and can disperse up to a distance of 65 km (Braack and Retief, 1986). Predators also play a role in the dissemination of spores by the opening up and dispersal of carcasses. Spores are passed in their faeces to new sites. Vultures will bathe in nearby water sources after feeding on infected carcasses, contaminating it with spores and will also pass spores in their droppings (De Vos, 1990).

Detecting an anthrax outbreak can be difficult, especially in a vast wilderness region like KNP. Currently this disease is monitored through passive surveillance. Control measures in KNP will only be implemented when biodiversity is negatively impacted (e.g. threatening the survival of low density or vulnerable species), and/or where man-made features (e.g. permanent watering holes) can propagate/sustain an outbreak (De Vos and Turnbull, 1994). However, management policies in the Park can influence the disease. The western boundary fence of the Park was constructed in 1960, thus preventing animals to migrate to their preferred dry season grazing areas to the west. The provision of artificial water sources was then instituted and between 1930 and 1980, more than 300 boreholes were drilled and 50 dams constructed to counteract the impacts of the fence. These artificial water sources had a negative impact on the biodiversity of species

(especially the rare antelope) - due to an increase in number of plains game that now permanently stayed close to wherever a water source was. It was subsequently decided to close down some of these boreholes, a step which proved to be successful, since the plains game moved out of these areas and allowed the rare antelope species population time to recover. To date, a total of 196 boreholes have been closed. According to Park ecologist, Dr. Freek Venter, more of the remaining 141 boreholes will be closed until there are only about 50 left (Travers, 2005). It is difficult to quantify the effect that the closing of boreholes had on the distribution of anthrax. It can be argued that an open borehole can act as a concentration point for spores and that some of the closed down boreholes were contaminated. Once a borehole has been closed, soil disturbance and animal activity in the area significantly decreases, subsequently decreasing the risk of exposure and new infections.

In an extensive wildlife reserve, such as the KNP, it is very difficult, if not impossible, to ensure immediate and proper disposal of anthrax infected carcasses. Carcasses are opened by scavengers whereupon sporulation takes place. Contaminated areas with anthrax spores in the soil are thus constantly created by animals dying from the disease. Spores are disseminated by insects (blowflies in KNP) that contaminate browse in the vicinity, vultures and mammalian scavengers which contaminate water sources. Water run-off contaminates the grazing that is ingested by herbivores (Hugh-Jones and De Vos, 2002).

1.2. Problem Statement and Justification

The identification of potential sites suitable as environmental reservoirs for anthrax spores is critical for the surveillance and management of the disease in wildlife, as wide scale immunization in wildlife remains untenable. Passive surveillance is currently used to locate potentially infected carcasses and monitor the extent of an outbreak. Modelling of ecologically suitable areas for anthrax in the KNP can lead to a better understanding of anthrax ecology and epidemiology. Site identification can be achieved via modelling that can support and improve surveillance and control strategies of anthrax in the KNP. Additionally, it also has practical applications for anthrax control in the smaller game

parks surrounding KNP (Hugh-Jones and De Vos, 2002). Commercial and subsistence farming communities adjacent to the Park can benefit from increased surveillance at targeted locations by allowing implementation of increased prophylactic measures for their livestock.

A limited amount of modelling studies has been done on anthrax. An anthrax distribution modelling study was done by Blackburn *et al.* (2007), in which they utilized GARP as a modelling system. These authors modelled the ecological niche for *B. anthracis* in the contiguous United States of America (USA), using wildlife and livestock outbreaks as well as several environmental variables. The study found that the modelled niche was able to be defined as a narrow index of NDVI, precipitation and elevation. Because of the limited studies available, the environmental variables deemed most important in the Blackburn study may not be applicable in all environments or at all scales. Other modelling techniques and variables should also be investigated.

1.3. Research questions, study objectives and hypotheses

Null hypothesis: Ecologically suitable areas for the occurrence of *B. anthracis* cases that differ significantly from random cannot be modelled using Maxent (with a regularized training gain of more than 1.5 and predictor data with a resolution of 1km or finer).

Alternative hypothesis: Ecologically suitable areas for the occurrence of *B. anthracis* cases, that differ significantly from random, can be modelled using Maxent (with a regularized training gain of more than 1.5 and predictor data with a resolution of 1km or finer).

Objectives

- Create an electronic database of confirmed anthrax positive (carcass/soil) locations in the KNP from historic records;

- Identify the environmental conditions most suitable for the occurrence of anthrax and subsequent spore dissemination;
- Develop a SDM to evaluate areas suitable for anthrax occurrence and long term spore survival;
 - Evaluate the model output using different sets of environmental variables;
 - Identify the top environmental predictors for anthrax occurrence in KNP;
 - Evaluate the model output against known propagating epidemic occurrences;
 - Create quantitative distribution maps representing the relative likelihood of anthrax occurrence within the environment in the KNP;

1.4. Assumption and limitations

In this dissertation anthrax positive case locations without specific coordinates were given standard coordinates based on the locality described in the record. It is assumed that these coordinates will still accurately reflect real presence since Maxent is not sensitive to small changes in coordinate values (Graham *et al.*, 2008; Baldwin, 2009). Place names in KNP have been very well described with reliable spatial information (Kloppers and Bornman, 2005)

1.5. Software

The following software packages were used in this project:

- ESRI ArcGIS Desktop 9.3.1 (ESRI, 2012)
- Maxent version 3.3.3k (Phillips *et al.*, 2004)
- ENMTools (Warren *et al.*, 2008)
- R statistical software (R Core Development Team, 2008)
- Diva-GIS (Hijmans *et al.*, 2012)
- StataSE12 (Statacorp, 2001)

- ModEco (Guo and Lui, 2010)
- Microsoft Excel 2010 (Microsoft Corporation, 2010)
- Apache OpenOffice 3.4.1. (Apache Software Foundation, 2012)
- Maxent Model Surveyor (Verbruggen, 2012)

2. Materials and Methods

In this study two models were developed, namely a preliminary model to eliminate variables with lesser importance and the final model with the most important environmental variables.

Anthrax positive case locations were collated from different datasets for the period 1950-2011. This included 597 records from passive surveillance data from 1988-2011, data of 1950-1959 outbreaks obtained from Pienaar (1960) and data of the 1959 outbreak consisting of 1151 anthrax cases obtained from Pienaar (1961). The private reserves adjacent to KNP were not explicitly included in this study, since no anthrax presence data was available.

2.1. Species Observation Data

2.1.1. *Bacillus anthracis* occurrence data

The State Veterinary office in Skukuza, KNP, SA provided anthrax surveillance data. These data were in electronic format for the period 2010-2011, and hard copy for the period 1988- 2009. These data were collated in an electronic database. A total of 597 confirmed anthrax cases from 1988-2011 were used in this study. Figure 3 indicates the positive anthrax cases in KNP from 1988-2011. Furthermore, two separate sets of data for the propagating epidemics were sourced from the literature as described by Pienaar (1960, 1961).

2.1.2. Processing of Species Observation Data

Each record contained: date sampled, processing date, species involved, gender of the species, presumed cause of death, location, ranger section, result of anthrax culture or blood smear and in some cases the Global Positioning System (GPS) coordinates of where the carcass was found. The sample points that did not contain GPS coordinates were given coordinates based on the recorded location within each ranger section (Kloppers and Bornman, 2005).

The GPS coordinates were not all in the same format and was either in degrees-minutes-seconds, degrees-decimal minutes or decimal degrees. The point data were standardized and converted to Universal Transverse Mercator (UTM) coordinates for use in ArcMap using Microsoft Excel. Only the species involved and UTM coordinates were used as input for the model. From the Excel database, the data was exported as a .dbf file using Apache OpenOffice, and imported into ArcCatalog for indexing of the individual records (Figure 27, Appendix D). From here it was exported for use in ArcMap. In ArcMap, the individual XY-coordinates were added as a point data layer and a comma separated values (.csv) file was created for use in Maxent.

Pienaar (1960) listed the data for the initial propagating epidemics in table format and map form. A map indicating the occurrence of positive anthrax cases during the outbreak is included in Figure 8. Both the table and the map from Pienaar (1960) were used for constructing an occurrence database. As only the location was given in the table, without coordinates, the exact location had to be tested against the map (Figure 8). Data were heads-up digitized (clicking on screen at the correct location on a digital map) using ArcGIS 9.3.

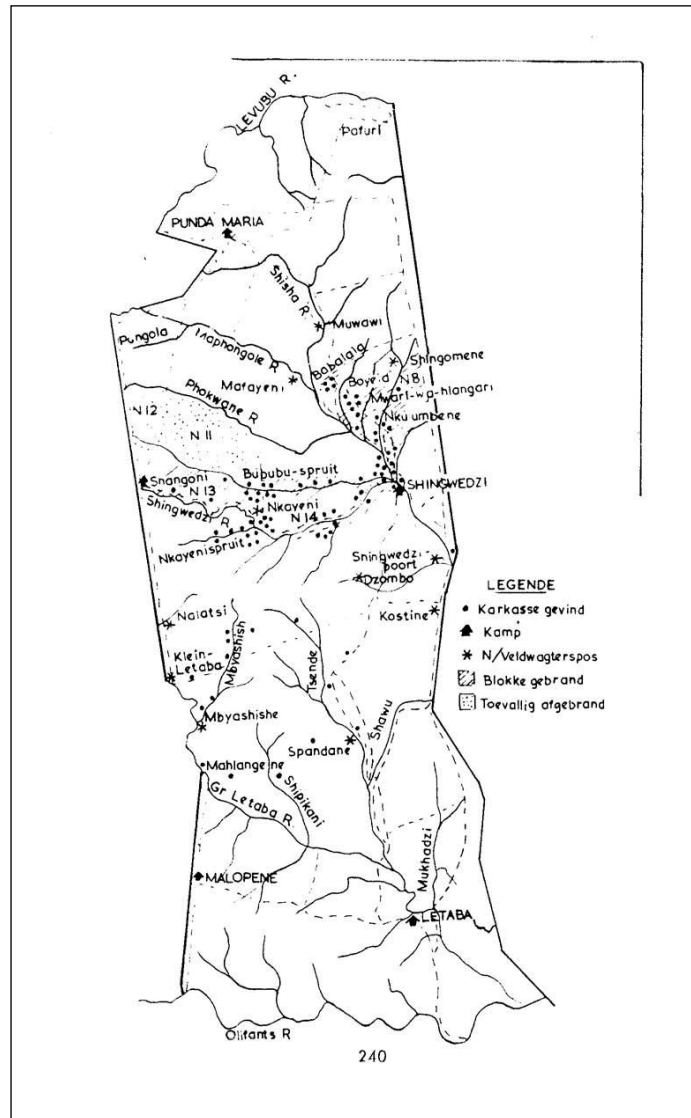


Figure 8: Positive anthrax cases during 1950-1959 outbreak in Kruger National Park as indicated by Pienaar (1960).

Data for the second outbreak in 1959 were obtained from a map (Figure 9: Positive anthrax cases during 1959 anthrax outbreak in Kruger National Park as indicated by Pienaar (1961)). Each point occurrence was heads-up digitized in ArcGIS to create a XY occurrence layer. The above mentioned datasets were constructed for qualitative testing of the model outcomes.

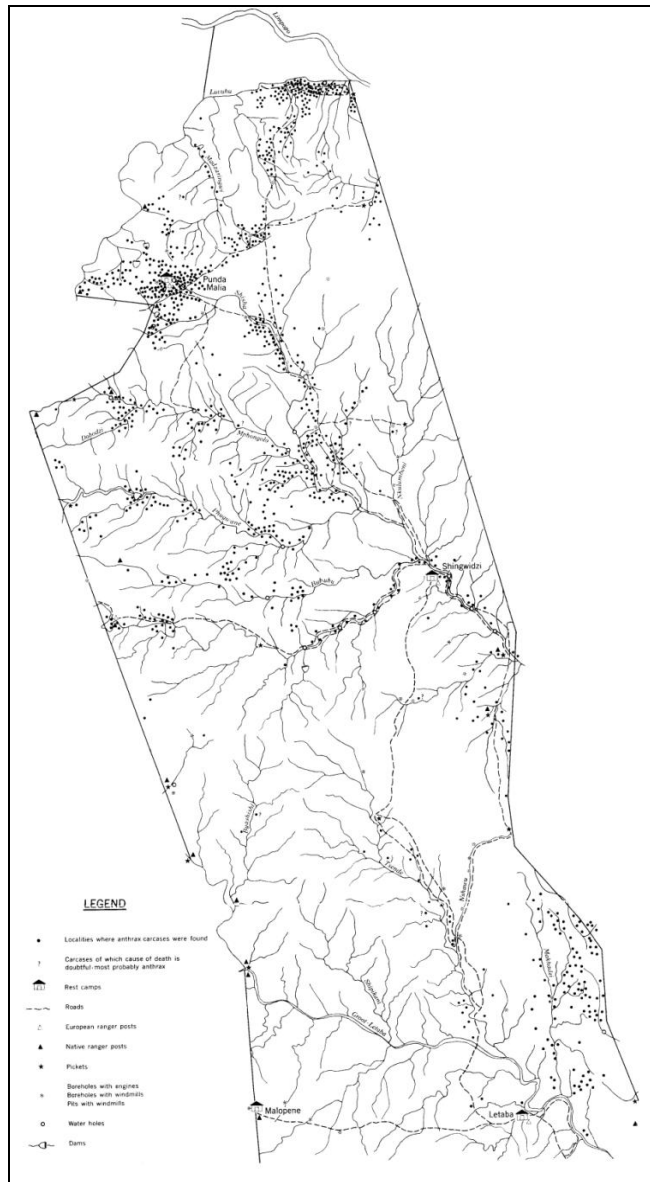


Figure 9: Positive anthrax cases during 1959 anthrax outbreak in Kruger National Park as indicated by Pienaar (1961).

2.2. Environmental Predictors

A total of forty environmental variables were used for initial model construction.

2.2.1. Selection Criteria

Care was taken to select variables that were ecologically meaningful. There is a large amount of environmental data available from various sources. Environmental variables were selected based on the possible impact that the variable involved can have on spore survival. All available variables were included in the first Maxent runs for exploratory analysis. The list of environmental variables contained 16 bioclimatic variables, two NDVI values (based on aggregated data from 2000 to 2009), an Aster Digital Elevation Model (DEM), distance of positive locations from dams, pans, rivers, springs, troughs, bore holes and water holes, soil data and vegetation data (Table 3).

Table 3: Overview of environmental data used in Maxent indicating the variables, type of data, source, spatial resolution and references.

#	Variable (*)	Type of data	Source	Original spatial resolution	E - Epidemiological importance references T – Technical references
1	Integrated NDVI (indvi)	Reflectance derived	MODIS-TERRA	1 km	E – (Jönsson and Eklundh, 2004; Petorelli <i>et al.</i> , 2005; Blackburn <i>et al.</i> , 2007) T – (Epistis, 2012)
2	Maximum NDVI (maxndvi)	Reflectance derived	MODIS-TERRA	1 km	E – (Jönsson and Eklundh, 2004; Petorelli <i>et al.</i> , 2005; Blackburn <i>et al.</i> , 2007) T – (Epistis, 2012)
3	Elevation (altitude)	Elevation derived	Aster-DEM	1 arc second (~ 30 m)	E – (; Blackburn <i>et al.</i> , 2007; Hugh-Jones and Blackburn, 2009; Joyner <i>et al.</i> , 2010) T – (USGS, 2012)
4	Slope (slope)	Elevation derived	DEM-derived	1 arc second (~ 30 m)	E – (De Vos, 1994) T – (USGS, 2012)
5	Aspect (aspect)	Elevation derived	DEM-derived	1 arc second (~ 30 m)	E – (De Vos, 1994) T – (USGS, 2012)
6	Distance to permanent water (permdist)	Distance metrics	ArcGIS Spatial Analyst	1 km	E – (De Vos, 1990; De Vos, 1994; Hugh-Jones and Blackburn, 2009)

			extension		T – (Hannart and Hughes, 2003)
7	Distance to seasonal water (seasdist)	Distance metrics	ArcGIS Spatial Analyst extension	1 km	E – (De Vos, 1990; De Vos, 1994; Hugh-Jones and Blackburn, 2009) T – (Hannart and Hughes, 2003)
8	Distance to ephemeral water (ephdist)	Distance metrics	ArcGIS Spatial Analyst extension	1 km	E – (De Vos, 1990; De Vos, 1994; Hugh-Jones and Blackburn, 2009) T – (Hannart and Hughes, 2003)
9	Distance to boreholes (boreholedist)	Distance metrics	ArcGIS Spatial Analyst extension	1 km	E – (De Vos, 1990; De Vos, 1994; Hugh-Jones and Blackburn, 2009)
10	SOTER Soil ID (sotersoilid)	Soils	SOTER database	1 km	E – (De Vos, 1990; De Vos, 1994; Hugh-Jones and Blackburn, 2009) T – Batjes, 2004; Dijkshoorn <i>et al.</i> , 2008
11	Land Type (ltypeventer)	Soils	KNP Scientific Services Skukuza	1 km	E – (De Vos, 1990; Venter, 1990; De Vos, 1994; Hugh-Jones and Blackburn, 2009)
12	Landscape (landscapegert)	Soils	KNP Scientific Services Skukuza	1 km	E – (Gertenbach, 1983; De Vos, 1990; De Vos, 1994; Hugh-Jones and Blackburn, 2009)
13	Basalt or Granite (basaltgranite)	Soils	KNP Scientific Services Skukuza	1 km	(Gertenbach, 1983)
14	Land Cover (landcover)	Soils		1 km	E – (Hugh-Jones and Blackburn, 2009) T – (Peace Parks Foundation, 2008)
15	Geology (geologyventer)	Soils	KNP Scientific Services Skukuza	1 km	(Gertenbach, 1983; Venter, 1990)
16	Calcium (caventer)	Soils	Interpolated from Venter database	1 km	E – (De Vos, 1990; Venter, 1990; De Vos, 1994; Dragon and Rennie, 1995) T – (Batjes, 2004; Dijkshoorn <i>et al.</i> , 2008; Spectrum Analytic Inc, 2012)
17	Lithology SOTER (lithosoter)	Soils	SOTER database	1 km	T – (Batjes, 2004; Dijkshoorn <i>et al.</i> , 2008)
18	pH Venter (phventer)	Soils	Interpolated from Venter database	1 km	E – (Venter, 1990; Dragon and Rennie, 2005; Blackburn <i>et al.</i> , 2007) T – (Batjes, 2004; Dijkshoorn <i>et al.</i> , 2008; Spectrum Analytic Inc, 2012)
19	pH SOTER (soterph)	Soils	SOTER database	1 km	E – (Dragon and Rennie, 2005; Blackburn <i>et al.</i> , 2007) T – (Batjes, 2004; Dijkshoorn <i>et al.</i> , 2008; Spectrum Analytic Inc, 2012)
20	Cation Exchange Capacity (cec)	Soils	Interpolated from Venter database	1 km	E – (Venter, 1990) T – (Spectrum Analytic Inc, 2012)

21	Total Available Water Capacity (tawc)	Soils	SOTER database	1 km	E – (Blackburn <i>et al.</i> , 2007) T – (Batjes, 2004; Dijkshoorn <i>et al.</i> , 2008)
22	Soil Clay % (clay)	Soils	SOTER database	1 km	T – (Batjes, 2004; Dijkshoorn <i>et al.</i> , 2008)
23	Soil Silt % (silt)	Soils	SOTER database	1 km	T – (Batjes, 2004; Dijkshoorn <i>et al.</i> , 2008)
24	Soil Sand % (sand)	Soils	SOTER database	1 km	T – (Batjes, 2004; Dijkshoorn <i>et al.</i> , 2008)
25	Annual mean temperature (annualtemp)	Climate	Worldclim	30 arc seconds (~ 1 km)	E – (Blackburn <i>et al.</i> , 2007; Joyner <i>et al.</i> , 2010) T – (Hijmans <i>et al.</i> , 2005)
26	Annual precipitation (annualprecipitation)	Climate	Worldclim	30 arc seconds (~ 1 km)	E – (Blackburn <i>et al.</i> , 2007; Joyner <i>et al.</i> , 2010) T – (Hijmans <i>et al.</i> , 2005)
27	Isothermality (isothermality)	Climate	Worldclim	30 arc seconds (~ 1 km)	E – (Blackburn <i>et al.</i> , 2007) T – (Hijmans <i>et al.</i> , 2005)
28	Maximum temperature warmest month (maxwarmmonth)	Climate	Worldclim	30 arc seconds (~ 1 km)	E – (Blackburn <i>et al.</i> , 2007) T – (Hijmans <i>et al.</i> , 2005)
29	Mean diurnal range (meandirange)	Climate	Worldclim	30 arc seconds (~ 1 km)	E – (Blackburn <i>et al.</i> , 2007) T – (Hijmans <i>et al.</i> , 2005)
30	Mean temperature warm quarter (meantwarmq)	Climate	Worldclim	30 arc seconds (~ 1 km)	E – (Blackburn <i>et al.</i> , 2007) T – (Hijmans <i>et al.</i> , 2005)
31	Mean temperature wet quarter (meantwetq)	Climate	Worldclim	30 arc seconds (~ 1 km)	E – (Blackburn <i>et al.</i> , 2007) T – (Hijmans <i>et al.</i> , 2005)
32	Mean temperature dry quarter (meantdryq)	Climate	Worldclim	30 arc seconds (~ 1 km)	E – (Blackburn <i>et al.</i> , 2007) T – (Hijmans <i>et al.</i> , 2005)
33	Minimum temperature of coldest month (mintcoldmonth)	Climate	Worldclim	30 arc seconds (~ 1 km)	E – (Blackburn <i>et al.</i> , 2007) T – (Hijmans <i>et al.</i> , 2005)
34	Precipitation of driest month (precdrymonth)	Climate	Worldclim	30 arc seconds (~ 1 km)	E – (Blackburn <i>et al.</i> , 2007; Joyner <i>et al.</i> , 2010) T – (Hijmans <i>et al.</i> , 2005)
35	Precipitation of driest quarter (precdryq)	Climate	Worldclim	30 arc seconds (~ 1 km)	E – (Blackburn <i>et al.</i> , 2007) T – (Hijmans <i>et al.</i> , 2005)
36	Precipitation seasonality (precseasonality)	Climate	Worldclim	30 arc seconds (~ 1 km)	E – (Blackburn <i>et al.</i> , 2007) T – (Hijmans <i>et al.</i> , 2005)
37	Precipitation of wettest month (precwetmonth)	Climate	Worldclim	30 arc seconds (~ 1 km)	E – (Blackburn <i>et al.</i> , 2007; Joyner <i>et al.</i> , 2010) T – (Hijmans <i>et al.</i> , 2005)
38	Precipitation of wettest quarter (precwetq)	Climate	Worldclim	30 arc seconds (~ 1 km)	E – (Blackburn <i>et al.</i> , 2007) T – (Hijmans <i>et al.</i> , 2005)
39	Temperature annual range (tempnrange)	Climate	Worldclim	30 arc seconds (~ 1 km)	E – (Blackburn <i>et al.</i> , 2007; Joyner <i>et al.</i> , 2010) T – (Hijmans <i>et al.</i> , 2005)
40	Temperature seasonality (tempseasonality)	Climate	Worldclim	30 arc seconds (~ 1 km)	E – (Blackburn <i>et al.</i> , 2007) T – (Hijmans <i>et al.</i> , 2005)

* Variable name as used in model included in parenthesis

2.2.2. Spatiotemporal Framework

The projected coordinate system, geographic extent and spatial resolution of all environmental layers were set to be the same before conversion from raster data to Ascii format. This was done to ensure positional accuracy and attribute integrity.

The projected coordinate system chosen was WGS 1984 UTM Zone 36S, suitable for use between 30°-36°E and between 0°-80°S onshore and offshore. This projection is based on the World Geodetic System of 1984 (WGS-84) ellipsoid. The central meridian was set to 33° E; False Easting was 500000; False Northing was 10000000; Linear Unit meters; Angular Unit degrees; Latitude of Origin 0; Datum WGS 84.

Spatial resolution was set to 1000m. The data set with the lowest resolution determined the resolution to be used by the model – in this case it was WorldClim (Hijmans *et al.*, 2005). Resampling using nearest neighbour (for discrete data) and bilinear interpolation (for continuous data) was performed on all layers. The analysis extent was set to: Top 7529415.63621m, Left 283091.00223m, Right 402091.00223m and Bottom 7175415.63621m (WGS 1984 UTM Zone 36S). This resulted in 119 columns and 354 rows in each clipped raster with a cell size of 1 km² and a total pixel count of 18983. In the KNP outline model, the environmental layers were clipped to fit the study area. As this was a purely spatial model, temporal variables were not considered, since positive cases were used as a proxy for the environment's potential suitability to store *B. anthracis* spores and not the time of death. The rather coarse spatial resolution also minimized the significance of the difference in the exact point of infection and the point of death.

2.2.3. Topography and Soil Variables

The following two paragraphs were taken from the ArcGIS Desktop Help file to explain resampling:

“Nearest neighbour assignment is the resampling technique of choice for discrete (categorical) data since it does not alter the value of the input cells. Once the location of

the cell's center on the output raster dataset is located on the input raster, nearest neighbour assignment will determine the location of the closest cell center on the input raster and assign the value of that cell to the cell on the output raster.” (ESRI, 2012).

“Bilinear interpolation uses the value of the four nearest input cell centers to determine the value on the output raster. The new value for the output cell is a weighted average of these four values, adjusted to account for their distance from the center of the output cell. This interpolation method results in a smoother looking surface than can be obtained using nearest neighbour. Since the values for the output cells are calculated according to the relative position and the value of the input cells, bilinear interpolation is preferred for data where the location from a known point or phenomenon determines the value assigned to the cell (that is, continuous surfaces).” (ESRI, 2012).

The environmental variable altitude was acquired as an Aster Digital Elevation Model (DEM) from the United States Geological Survey (USGS) at <http://earthexplorer.usgs.gov/>, with a resolution of 1 arc second (~30 m). The layer was reprojected, resampled using bilinear interpolation and clipped to extent.

The Aspect variable was created from the clipped DEM layer using the ArcGIS Spatial Analyst extension version 9.3.1. Since aspect was represented as degrees, it needed to be converted to linear decimal. For this study $\cos(\text{aspect})$ was used and the conversion was done in the raster calculator of ArcGIS.

Soil data were obtained from two different sources – KNP Scientific Services and the SOTER (SOil and TERRain) digital database (Batjes, 2004; Dijkshoorn *et al.*, 2008). The SOTER data was downloaded at <http://www.isric.org/data/data-download> with a 1 km resolution. SOTER variables used included lithology, soil type, total available water capacity (TAWC), pH, soil clay content, soil silt content and soil sand content. Table 4 illustrates the number of different classes that each categorical variable was divided into.

Table 4: Number of classes per selected categorical variable

Variable	Number of classes
lithosoter	8
geologyventer	15
sotersoilid	46
ltypeventer	56
landscapegert	35
basaltgranite	2

The polygon layers obtained from KNP Scientific Services included geology, land type, lithology and soil data. All the categorical layers were converted to rasters, reprojected and resampled using nearest neighbour technique.

A dataset (Venter, 1990) with 370 soil sampling sites was provided by the KNP Scientific services. Two layers - caventer and phventer were created from this data. Inverse Distance Weighted (IDW) interpolation was used in ArcGIS to derive Ca and pH values for the rest of the KNP. Ideally, specific measured values around the Park should be used, but this was not possible from the data provided. Since interpolation creates new values for missing ones by using known values, a possible decrease in the precision of the model can result and the Ca and pH layers were interpreted with caution. Finally, all rasters were converted to Ascii format in ArcGIS for use in Maxent.

2.2.4. Processing of the NDVI variable

NDVI is derived from satellite data and is commonly used as a proxy for vegetation productivity (Pettorelli *et al.*, 2005). The NDVI provides information about the spatial and temporal distribution of vegetation communities, vegetation biomass, CO₂ fluxes, vegetation quality for herbivores (because the rate of greening can be correlated with food quality) and the extent of land degradation in various ecosystems (Pettorelli *et al.*, 2005). Pre-processed data from the Moderate Resolution Imaging Spectroradiometer (MODIS–TERRA) were used in this study, spanning 2000-2009 at 250 m resolution (Epistis, 2012).

Two derived NDVI values were used in this study (Figure 10):

1. Integrated NDVI: INDVI is the sum of all the positive NDVI values over a given period and provides a measure of overall productivity and biomass. This is equivalent to the large integral of an NDVI curve (Jönsson and Eklundh, 2004; Petorelli *et al.*, 2005). A mosaic raster was created from the annual large integral rasters to produce the index of overall productivity.
2. Maximum NDVI: MaxNDVI is the annual maximum NDVI value. This is another measure of overall productivity and biomass (Jönsson and Eklundh, 2004; Petorelli *et al.*, 2005). The yearly maximum values were determined in ArcGIS and a mosaic raster was created to produce the index of overall productivity.

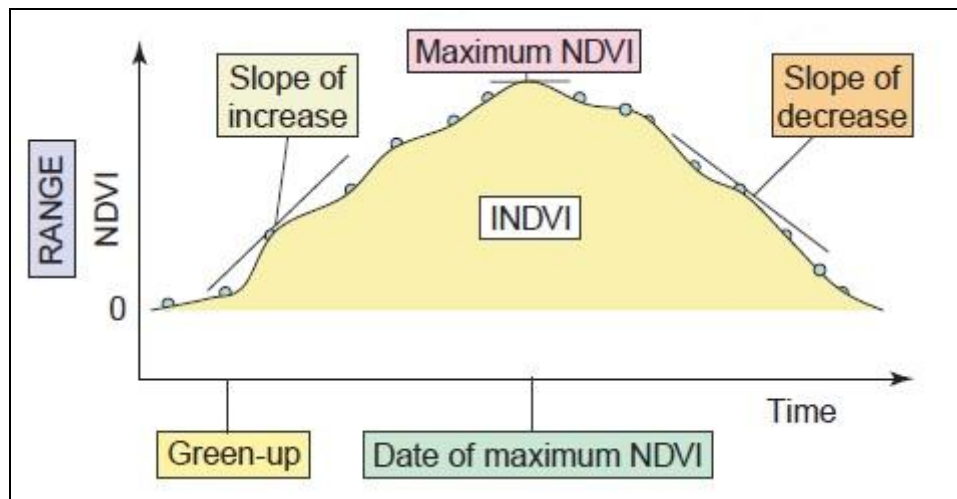


Figure 10: Presentation of the different indices (the slopes of increase (spring) and decrease (autumn), the maximum NDVI value, the integrated NDVI (INDVI, i.e. the sum of NDVI values over a year), the date when the maximum NDVI value occurs, the range of annual NDVI values, and the date of green-up (i.e. the beginning of the growing season)) that could be derived from NDVI time series over a year. Image adapted from Petorelli *et al.* (2005).

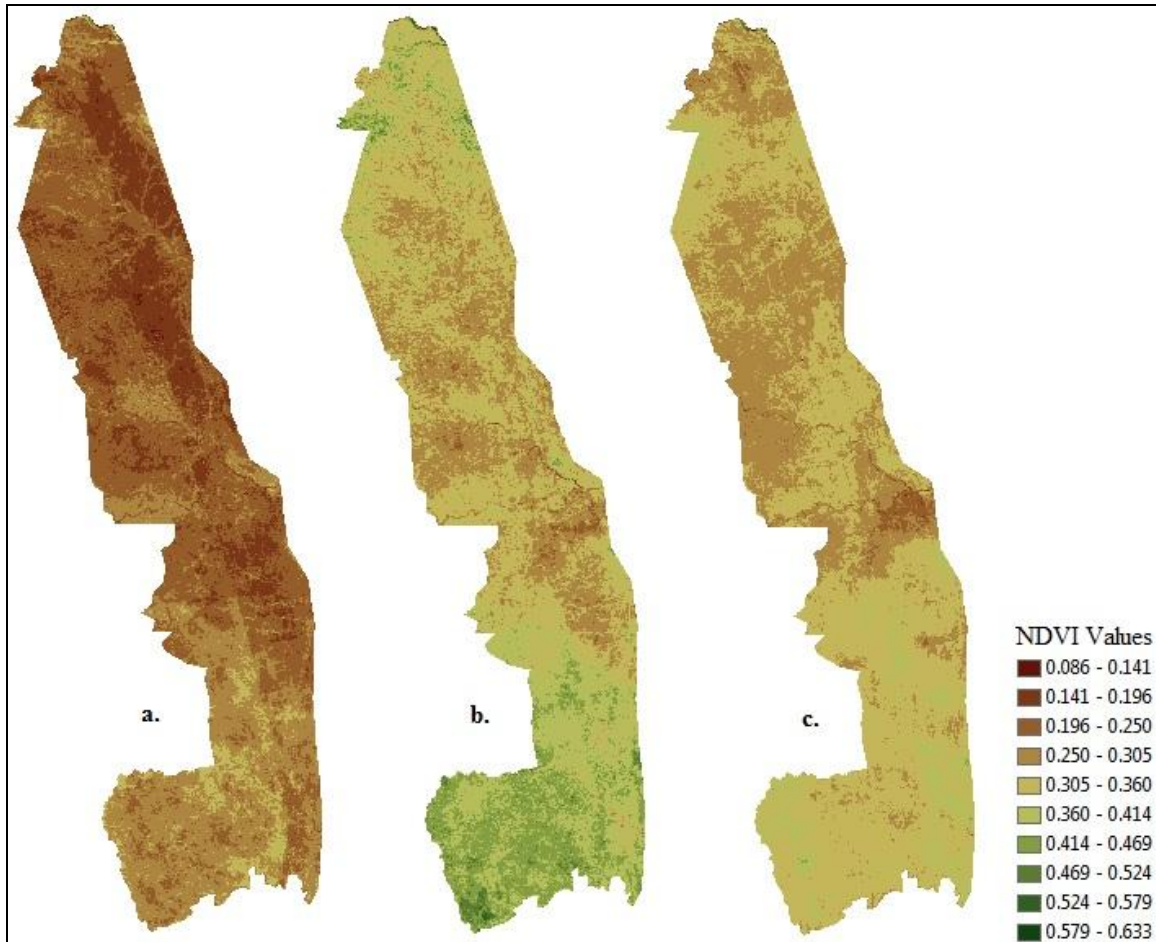


Figure 11: NDVI values (ranging from 0 to 1) for Kruger National Park in (a) November – period of increase, (b) January – maximum NDVI and (c) May 2008 – period of decrease.

Figure 11 displays the NDVI values corresponding to seasonal change. Figure 11a represents the transition into the growing season (spring). Figure 11b represents the peak of the growing season (summer) and Figure 11c represents the transition into the dormant season (autumn).

Other derivations of the NDVI variable include (1) beginning of season, (2) end of season, (3) left 90% level, (4) right 90% level, (5) peak, (6) amplitude, (7) length of season, (8) rate of increase / decrease and (9) relative annual range (Jönsson and Eklundh, 2004; Petorelli *et al.*, 2005). Only the INDVI and MaxNDVI values are used as indicators of overall productivity and biomass (Petorelli *et al.*, 2005).

The rasters were resampled (bilinear interpolation) to 1000 m and clipped to extent before conversion to Ascii format.

2.2.5. Processing of Land Cover variables

Land cover layers used in this study included land cover, land type, vegetation type, geology and landscapes. The land cover layer (Peace Parks Foundation, 2008) was resampled and clipped to extent. Since all other land cover layers were in polygon format, each one was converted to a raster for export to Ascii format. All of these variables were used as categorical data in Maxent. Vegetation type was used as defined by Gertenbach (1983) (Figure 1).

2.2.6. Processing of Climate Variables

Nineteen bioclimatic variables were downloaded from the WorldClim website under the number 37 tile - <http://www.worldclim.org/tiles.php?Zone=37>. Bioclimatic variables are derived from the monthly temperature and rainfall values in order to generate more biologically meaningful variables. These are often used in ecological niche modelling (e.g., BIOCLIM, GARP). The bioclimatic variables represent annual trends (e.g., mean annual temperature, annual precipitation) seasonality (e.g., annual range in temperature and precipitation) and extreme or limiting environmental factors (e.g., temperature of the coldest and warmest month, and precipitation of the wet and dry quarters; a quarter is a period of three months) (Hijmans *et al.*, 2005).

The rasters were all in 30 arc second resolution (~1 km) and were reprojected to the required projection. An extract by mask was performed on all rasters and each one resampled using bilinear interpolation using ArcGIS. Finally the layers were converted to Ascii format for use in Maxent.

2.2.7. Processing of Hydrological Variables

2.2.7.1. Rivers

The following hydrological layers were created and used: A river polygon layer was downloaded from http://www.dwa.gov.za/iwqs/gis_data/river/rivs500k.html (Resource Quality Services (formerly IWQS) and Chief Directorate of National Geo-Spatial Information (formerly Surveys and Mapping)). The layer was clipped to the KNP boundary and combined with a layer from Hannart and Hughes (2003), which provided a hydrological index and class, which in turn allowed the river to be classified as permanent, seasonal or ephemeral. The rivers within the KNP were divided into three categories as indicated in Table 5. For each of the river categories an euclidean distance raster was created using the spatial analyst extension in ArcGIS.

Table 5: Hydrological indices and river classes used in this study (Hannart and Hughes, 2003). A hydrological index less than 16.110 indicates a permanent river. A hydrological index between 16.110 and 37.81 indicates a seasonal river and a hydrological index greater than 37.81 indicates an ephemeral river.

Class	Hydrological index (HI) thresholds	Flow variability descriptors used in this study
1	HI 4.394	Permanent
2	4.394 < HI 7.535	
3	7.535 < HI 13.74	
4	13.745 < HI 16.110	
5	16.110 < HI 37.81	Seasonal
6	37.819 < HI 64.16	Ephemeral
7	64.169 < HI 92.70	
8	92.705 < HI 98.12	
9	98.124 < HI	

2.2.7.2. Boreholes

A shapefile containing all the boreholes within the KNP, listed as points, was provided by SANParks Scientific Services. Former boreholes (those currently closed) were not considered in this study. The remaining operational boreholes were mapped in ArcGIS and an euclidean distance raster layer was created. See Appendix C for information regarding each borehole such as drilling data and close down date.

2.2.7.3. Pans, troughs, dams, waterholes, springs

The challenge with classifying and mapping water sources was to determine which structures were ephemeral, permanent or seasonal. The dams layer provided information

on the working status of the dam. If the dam was operational, it was considered a permanent water source, otherwise it was considered ephemeral. Pans were all considered ephemeral except those linked to permanent river systems and some pans specifically listed as being permanent. Springs and waterholes were all considered ephemeral. Euclidean distance from feature rasters were created and incorporated into the above mentioned river rasters according to classification as ephemeral, permanent or seasonal. The number of seasonal and ephemeral rivers resulted in no change in the importance of the respective variables when combined with the above mentioned features. It should, however, be noted that very different dynamics and processes occur at these water sources. If there was a larger amount of rivers or the rivers had a negative impact on the importance of the variables, then the layers would have been separated into rivers and standalone features respectively. All the above euclidean distance rasters were converted to Ascii format for use in Maxent using ArcGIS.

2.3. Modelling technique

2.3.1. Maxent settings

Using standard settings, and thus auto feature selection, implicates that Maxent will automatically add modelling features with increasing number of samples in the training set: below 10 samples only linear functions are used; between 10 and 14 samples quadratic features are added; between 15 and 79 samples hinge features are added and above 79 samples product and threshold features are allowed (Merckx *et al.*, 2011). Since the number of samples in the presence data set was bigger than 100, the auto-features option was selected (Phillips *et al.*, 2004; Elith *et al.*, 2011) which automatically adjusts the beta regularization parameter for each feature type in the model (Figure 28, Appendix D).

The random seed option was selected to create a different training and test dataset for each model run. A different background dataset was also created for each model run. The option to remove duplicate presence records was deselected (default select). This was done because more presence locations in the same area signify more ecologically suitable conditions for *B. anthracis*. Random test percentage was set to 25% meaning that Maxent

randomly sets aside 25% of the provided presence data and uses these data to test the model. The amount of replicate runs was set to 10, and replicate run type to subsample – this removed the sampling set used for testing and selected a new set (excluding all previous set points) on each run (Figure 29, Appendix D). The advanced options in Maxent that were selected included the maximum iteration set to 5000 to allow the models enough time to reach convergence (Young *et al.*, 2011) (Figure 30, Appendix D).

2.3.2. Modelling Procedure

Two separate models were developed in this study namely (i) a preliminary model used to eliminate variables with lesser importance and (ii) a model with the most important environmental variables selected. Model (ii) was used to determine the final ecological suitability distribution of *B. anthracis*. The datasets used in this study consisted of (1) 597 anthrax cases, (2) a set of environmental variables specific to the clipped extent of the KNP, (3) 1950-1959 anthrax outbreak as described by Pienaar (1960) and (4) the 1959 outbreak documented by Pienaar (1961) with a total of 1151 positives. Datasets 1 and 2 were used to build the model and datasets 3 and 4 were used to visually evaluate the outcomes of the model.

Test and training data

There are two ways in which Maxent evaluates model performance – “test of fit” and “test to predict”. This is accomplished by splitting the set of occurrence data into test and training datasets. The training dataset is used to build the model and the test dataset is used to measure how accurately the model can predict the points within this test set. A random test percentage option is included in Maxent to indicate the percentage of data to use for testing and training. The dataset was divided into training (75%) and test (25%) presence points. Without this measure, the model would use the training data in its test thus inflating model performance (Young *et al.*, 2011).

2.3.3. Maxent and sample bias

By default, when using Maxent, the assumption is made that species occurrence data are unbiased, independent samples from the distribution of the species. The assumption of lack of bias is easily violated, for example if sample collection effort is biased towards more easily accessible areas such as areas close to roads or populated centers. A simple strategy to remove sample selection bias is to replace the uniform background data by a random sample of background data drawn from the sampling distribution (Dudík *et al.*, 2005; Phillips *et al.*, 2009). Since Maxent selects the distribution of maximum entropy relative to the provided background, the sample selection bias was effectively factored out.

To create a target background layer, the area surrounding the presence points was extracted from the KNP extent (Figure 12). This was done by selecting all the ranger sections within KNP that contained presence points. Since there are numerous ranger sections within the Park, this method provided an easy way to select the most appropriate background layer. The selection was converted to a raster and the spatial analyst raster calculator used to convert all the cells to the value of '1' or 'NoData' using a conditional function: `con("BGLayer" >= 0, 1, "BGLayer")` (Young *et al.*, 2011). This was an important step since Maxent only accepts values that are bigger than zero for this type of file. Finally the raster was converted to Ascii format for use in Maxent.

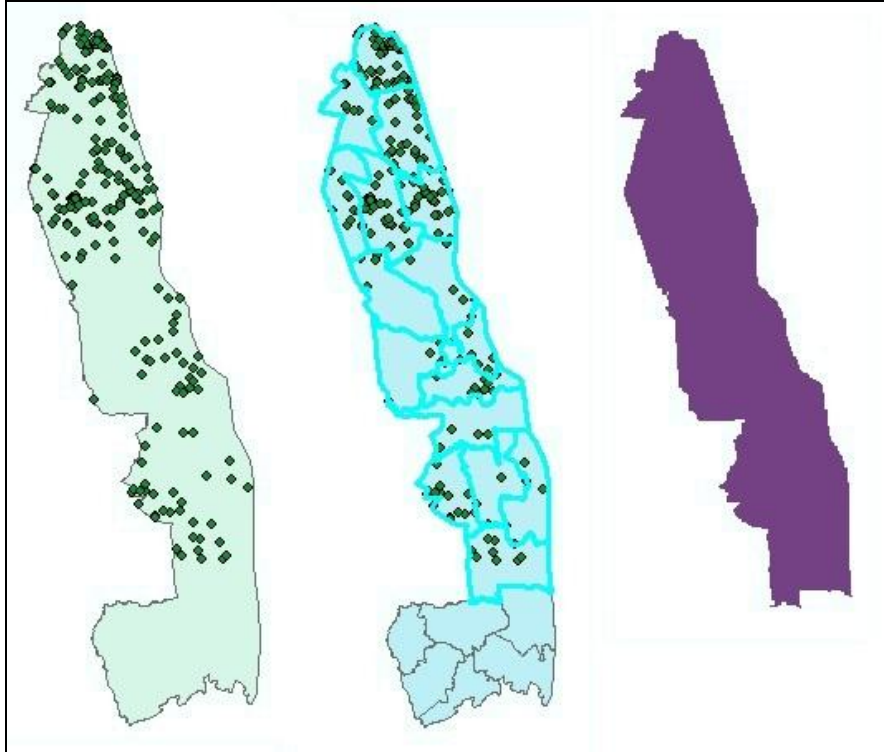


Figure 12: Background bias layer creation in Maxent. The purple map on the right indicates the bias layer which Maxent used to create background points from.

2.3.4. Null Model

ENMTools was used to generate the null-model datasets (Figure 31, Appendix D). A total of 1000 pseudo-presence location sets with 597 records each were created. A frequency histogram of all the AUCs was created and used for model AUC comparison.

2.3.5. Simple multiple regression

Simple multiple regression was used to eliminate linearly correlated variables (Herkt, 2007). Regression was done using StataSE 12.1 software (Statacorp, 2001).

2.3.6. Feature selection

The model was run 40 times, with removal of the least important variable after each run, until only one variable remained. A maximum of 5000 iterations per model run were used

to ensure convergence. The training gain, as well as the drop in training gain when the variable was omitted from the full model, was calculated for each variable. The variable with the lowest decrease in average training gain when omitted was the variable that was removed. The weakest of the remaining variables was again determined by repeating the above process. This procedure continued until only one variable remained.

Ninety-five percent confidence intervals were constructed for the training gain values associated with each model. The model that exhibited overlap in confidence intervals with the model of highest training gain was selected as the best model (Yost *et al.*, 2008).

Maxent has a built-in jackknife function to evaluate predictor importance by eliminating one variable at a time while recomputing the gain, thus eliminating variables that lead to overfitting (Peterson and Cohoon, 1999). A correlation matrix was derived from the set of all variables and this matrix was used to eliminate highly correlated variables (Herkt, 2007).

The results of the variable selection process were evaluated using the Maxent model surveyor (MMS). MMS is a stand-alone program that evaluates different sets of predictors for Maxent modelling (Verbruggen, 2012). The program determines the best subset of feature combinations based on maximum AUC. There are three different selection algorithms – best subset selection, forward stepwise search and backward stepwise search. The evaluation criteria can be AUC, Akaike Information Criterion (AIC), Corrected AIC (AICc) or the Bayesian Information Criterion (BIC).

The best subset selection procedure evaluates each possible combination of features and determines the AUC (or any evaluation criterion selected). The model with the subset of features that yields the highest AUC score is selected (Table 6).

The forward stepwise search procedure starts with only one variable and adds a variable on each model run until all variables have been added. The model with the highest evaluation criterion is selected.

The backward stepwise search procedure starts with all the variables and removes a variable on each model run until all variables have been removed. Again, the model with the highest evaluation criterion is selected.

Table 6: Best subset MMS procedure. Variables are added to the Maxent variable set until the AUC value for all possible combinations have been calculated.

Variable indicator	AUC	Variables
100000000000	0.614	altitude
010000000000	0.606	caventer
110000000000	0.681	altitude, caventer
001000000000	0.627	ephdist
Continued until all possible variable combinations have been evaluated		
001111111111	0.881	ephdist, geologyventer, indivi, landscapegert, ltypeventer, permdist, precdryq, seasdist, sotersoild, tempseasonality
101111111111	0.882	altitude, ephdist, geologyventer, indivi, landscapegert, ltypeventer, permdist, precdryq, seasdist, sotersoild, tempseasonality
011111111111	0.886	caventer, ephdist, geologyventer, indivi, landscapegert, ltypeventer, permdist, precdryq, seasdist, sotersoild, tempseasonality
111111111111	0.887	altitude, caventer, ephdist, geologyventer, indivi, landscapegert, ltypeventer, permdist, precdryq, seasdist, sotersoild, tempseasonality

2.3.7. Probability classes

Four arbitrarily defined probability classes were used to classify the ecological suitability of the modelled prediction:

- 1) high suitability: 80-100%;
- 2) moderate suitability: 60-80%;
- 3) low suitability: 30-60% ;
- 4) not suitable: 0-30%.

A threshold of 80% was used to identify the area as suitable for the occurrence of anthrax in the environment. This threshold was chosen based upon the assumption that areas with a very high probability (>80%) of anthrax occurrence have the potential to be ecologically suitable for harbouring spores, thus anthrax is endemic. Areas with a lower probability of anthrax occurrence are most likely propagating epidemic occurrences.

Potential anthrax endemic areas were defined as any area with a probability of anthrax occurrence over 80%. Potential anthrax epidemic areas were defined as any area with a probability of anthrax occurrence less than 80%.

2.3.8. Gap Analyses

A gap analyses was performed in Diva-GIS on the presence data and Maxent output to create a map where sampling should be prioritized in future surveys. This method uses the 10 percentile training presence map from the Maxent output (Appendix B). Areas where it is likely to encounter anthrax positive cases, but where there are currently few or no records of observations, can be identified by comparing maps of observed and potential diversity (Scheldeman and Van Zonneveld, 2010).

3. Results

Six hundred and sixty one anthrax positive cases were recorded, but sixty four points had to be excluded due to ineligible writing on the original records. The exclusion of these points should however have minimal effect on the Maxent outputs.

The modelling process reached convergence prior to the maximum iteration setting of 5000. The ROC curve had an average AUC of 0.9372 for training data and an average AUC of 0.909 for test data and was significantly different from a line of no information ($p < 0.01$).

A set of preliminary Maxent runs were performed for feature selection. The process of selecting the best model out of a subset of potential models is known as model selection. The process of eliminating or adding variables to a model is called variable or feature selection.

The variables 'meantcoldq', 'preccoldq' and 'preqwarmq' were excluded based on their perfect linear relationship with 'meantdryq', 'precdryq' and 'precwetq' respectively (Table 7). The latter three variables were selected over the former, based on the fact that anthrax typically occurs after a period of drought that is followed by heavy rain (De Vos, 1990). Due to the thorough variable removal process, only variables with a perfect linear relationship were excluded in this study. However, variables with a Pearson correlation coefficient above 0.7 should be considered for exclusion due to the problem of multicollinearity (Dormann *et al.*, 2012).

Table 7: Simple multiple regression on continuous variables

	MEANDR~0	MEANTC~0	MEANDT~0	M~MQN100	M~TQN100	MINT~100	PRECCO~0	PRECDR..	P~YQN100	PRECSE~0
MEANDRAN~100	1.0000									
MEANTCOL~100	0.9462	1.0000								
MEANTDRY~100	0.9462	1.0000	1.0000							
MEANTWAR~100	0.9736	0.9206	0.9206	1.0000						
MEANTWET~100	0.9752	0.9195	0.9195	0.9995	1.0000					
MINTCOLD~100	0.7458	0.9177	0.9177	0.7139	0.7106	1.0000				
PRECCOLD~100	-0.3775	-0.5174	-0.5174	-0.2316	-0.2316	-0.6327	1.0000			
PRECDRYM~100	-0.4250	-0.5242	-0.5242	-0.2816	-0.2815	-0.5830	0.9577	1.0000		
PRECDRYQN100	-0.3775	-0.5174	-0.5174	-0.2316	-0.2316	-0.6327	1.0000	0.9577	1.0000	
PRECSEAS~100	0.8635	0.9676	0.9676	0.8082	0.8059	0.9573	-0.6733	-0.6471	-0.6733	1.0000
PRECWARM~100	0.0543	0.1416	0.1416	0.2012	0.1960	0.2240	0.5346	0.5870	0.5346	0.0640
PRECWETM~100	-0.0062	0.1050	0.1050	0.1630	0.1528	0.2095	0.4922	0.5606	0.4922	0.0452
PRECWETQN100	0.0543	0.1416	0.1416	0.2012	0.1960	0.2240	0.5346	0.5870	0.5346	0.0640
TEMPANNR~100	0.8798	0.7160	0.7160	0.9262	0.9278	0.4000	0.0229	-0.0722	0.0229	0.5486
TEMPSEAS~100	0.5936	0.3566	0.3566	0.6912	0.6925	0.0023	0.3705	0.2620	0.3705	0.1494
	PRECWA~0	PRECWE..	P~TQN100	TEMPAN~0	TEMPSE~0					
PRECWARM~100	1.0000									
PRECWETM~100	0.9527	1.0000								
PRECWETQN100	1.0000	0.9527	1.0000							
TEMPANNR~100	0.1288	0.0827	0.1288	1.0000						
TEMPSEAS~100	0.1602	0.1375	0.1602	0.8999	1.0000					

Values represent r^2 . All calculations were done in StataSE12. Values of -1 or +1 represent perfect correlation.

Table 8: Simple multiple regression on categorical variables

	basalt~e	geolog~r	landsc~t	ltypev~r	soters~d
basaltgran~e	1.0000				
geologyven~r	0.3686	1.0000			
landscapeg~t	-0.2737	-0.6006	1.0000		
ltypeventer	0.1383	-0.5169	0.5043	1.0000	
sotersoilid	-0.1908	0.5983	-0.4972	-0.8554	1.0000

Values represent r^2 . All calculations were done in StataSE12. Values of -1 or +1 represent perfect correlation.

Some of the information contained in the categorical variables was nested in other categorical variables. Due to the big difference in the number of classes for each categorical variable, each variable was considered unique enough to include in the model.

Figure 13 displays the procedure of removing the variable with the lowest decrease in average training gain when omitted (Yost *et al.*, 2008). Table 9 illustrates the variable removal order.

Table 9: Variable selection procedure. The variable removed from the model based on the jackknife method is listed in the right column. Model gain and AUC values are displayed in columns 2 and 3 to illustrate the change (decrease) in value as more variables are removed.

Variable number	Gain	AUC	Variable Removed
40	1.864	0.936	basaltgranite
39	1.857	0.935	mintcoldmonth
38	1.857	0.935	boreholedist
37	1.856	0.935	precseasonality
36	1.856	0.935	soterph
35	1.856	0.935	annualmtemp
34	1.852	0.935	meandirange
33	1.852	0.935	meantdryq
32	1.856	0.935	isothermality
31	1.853	0.935	clay
30	1.85	0.935	meantwetq
29	1.848	0.935	meantwarmq
28	1.848	0.935	tempannrange
27	1.847	0.935	annualprecipitation
26	1.845	0.934	landcover
25	1.843	0.933	precdrymonth
24	1.839	0.933	maxtwarmmonth
23	1.839	0.932	precwetq
22	1.836	0.931	aspect
21	1.829	0.93	silt
20	1.82	0.93	tawc
19	1.812	0.929	phventer
18	1.806	0.928	precwetmonth
17	1.799	0.926	slope
16	1.79	0.925	cec
15	1.78	0.921	lithosoter
14	1.763	0.919	maxndvi
13	1.745	0.918	sand

12	1.726	0.915	preqdryq
11	1.698	0.91	caventer
10	1.678	0.907	altitude
9	1.647	0.906	ephdist
8	1.612	0.9	geologyventer
7	1.574	0.898	landscapegert
6	1.538	0.889	permdist
5	1.487	0.888	seasdist
4	1.425	0.877	ltypeventer
3	1.284	0.861	tempseasonality
2	1.213	0.849	indvi
1	1.295	0.832	sotersoilid

The gain in the second column is the regularized training gain. The variable removal order is represented in the right column. Figure 19 was constructed from Table 9 for illustrative purposes.

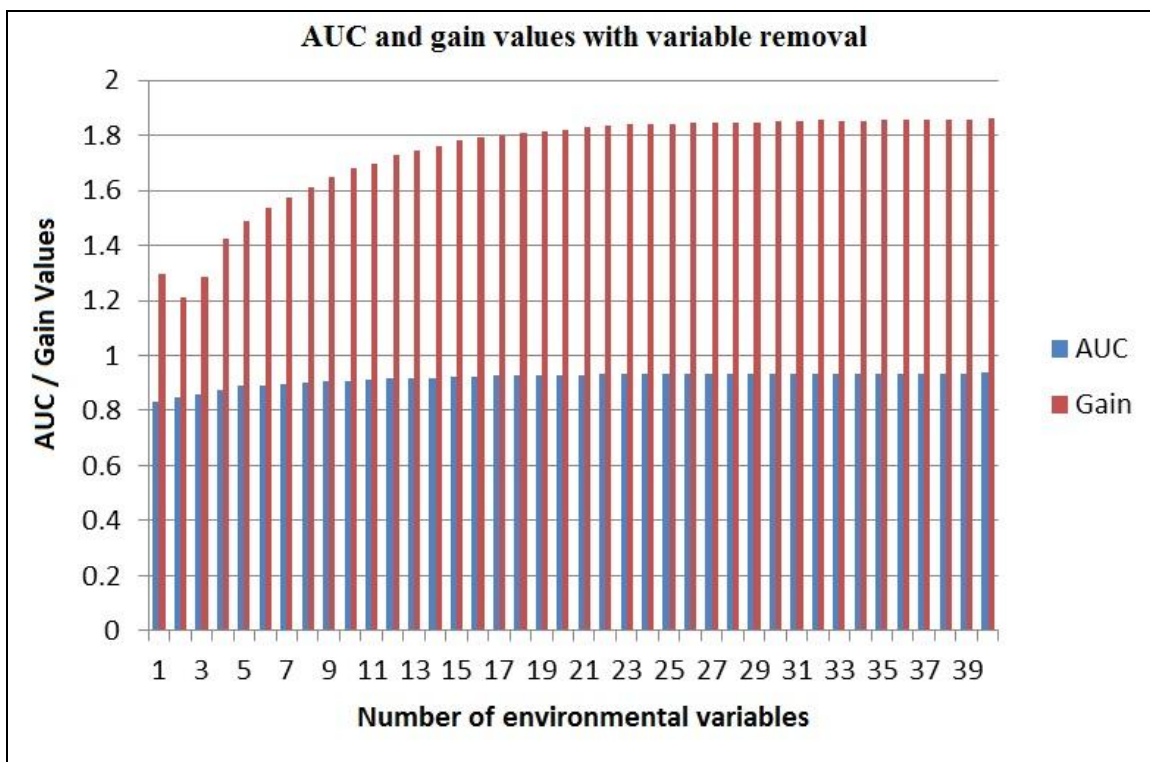


Figure 13: Variable selection procedure indicating the loss of gain and associated decrease in AUC value when variables are removed.

The training gain in the model with 12 variables was not significantly worse from the training gain of the models with 13 or more variables, but was significantly better from the models with fewer variables. The top 12 variables were used to build the final suitability model.

Results of the jackknife procedure for variable selection were evaluated by using the Maxent response curves (Appendix A) and the MMS algorithm (Verbruggen, 2012).

One disadvantage of using this (MMS) method is computation time. It took a total of 65536 model runs to select the best subset of 16 variables, making evaluation with 40 variables unrealistic. The best subset selection of 12 variables took 4095 model runs.

The best result of the MMS method was the 10 variable set of caverter, ephdist, geologyventer, indivi, landscapegert, ltypeventer, permdist, seasdist, sotersoilid and tempseasonality (Figure 32, Appendix D). The best subset AUC was 0.888 which is significantly lower than the Maxent model tested. This is because MMS uses a 50/50 partition of test and training data, making less data available for training. There is only a difference of two variables between the best subset MMS procedure output and the confidence interval jackknife method output. In this study the AUC was used as the MMS evaluation criterion for feature selection.

The Null Model Maxent procedure consisted of 1000 pseudo-presence observations with a mean AUC of 0.649 (Std. Dev. 0.0089). The model predicted presence significantly better than random based on the null model AUC 95% CI, $p < 0.0001$.

Results of the 10 model training partitions were evaluated by using a one-tailed t-test. The average regularized cumulative training gain of the ten model partitions was 1.9254 (H_0 : cumulative gain ≤ 1.5 and H_A : cumulative gain > 1.5). **The result rejects the null hypothesis and accepts the alternative hypothesis ($t = 21.05$ and $p < 0.0001$ at 95% CI).**

The model identified the following areas (Figure 14) as most suitable for the occurrence of anthrax namely Northern Pafuri depression, Mpongolo-Shingwedzi confluence and the Letaba-Olifants confluence.

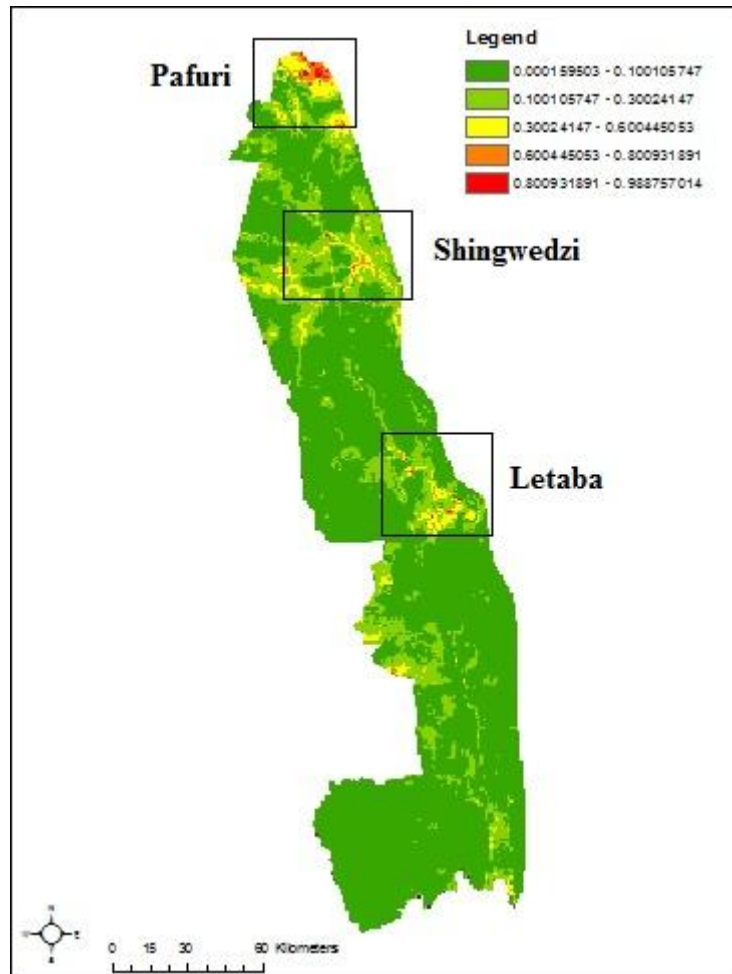


Figure 14: Suitable anthrax areas within Kruger National Park. Black rectangles indicate core areas (greater than 80% probability).

Table 10 contains estimates of the relative contributions of the 12 selected environmental variables in the final Maxent model. To determine the first estimate in each iteration of the training algorithm, the increase in regularized gain is added to the contribution of the corresponding variable (or subtracted if change is negative). This is a heuristic approach to model importance in which the contribution values are determined by the increase in gain in the model provided by each variable. Caution must be used when employing this

method as strong collinearity can influence results by indicating more importance for one of two or more highly correlated variables (Baldwin, 2009). For the second estimate, for each environmental variable in turn, the values of that variable on training presence and background data are randomly permuted. The model is re-evaluated on the permuted data, and the resulting drop in training AUC is shown in the table, normalized to percentages (Phillips *et al.*, 2006). Maxent performs only a single permutation of the predictor to calculate the metric. Performing many such permutations would be more reliable (but that would take a lot more time). Both these estimates serve as guidelines for variable importance and should be interpreted in combination with other methods such as jackknife.

Table 10: Variable contribution table including percent contribution and permutation importance of the variable listed.

<i>Variable</i>	<i>Percent contribution</i>	<i>Permutation importance</i>
sotersoilid	36.3	17.1
precdryq	15.9	23.2
landscapegert	9	13.7
indvi	8.6	2.4
altitude	7.3	2.5
ltypeventer	6	5
ephdist	4.6	9.5
permdist	3.3	5.4
seasdist	2.9	5.8
tempseasonality	2.7	5
geologyventer	2.2	2
caventer	1	8.4

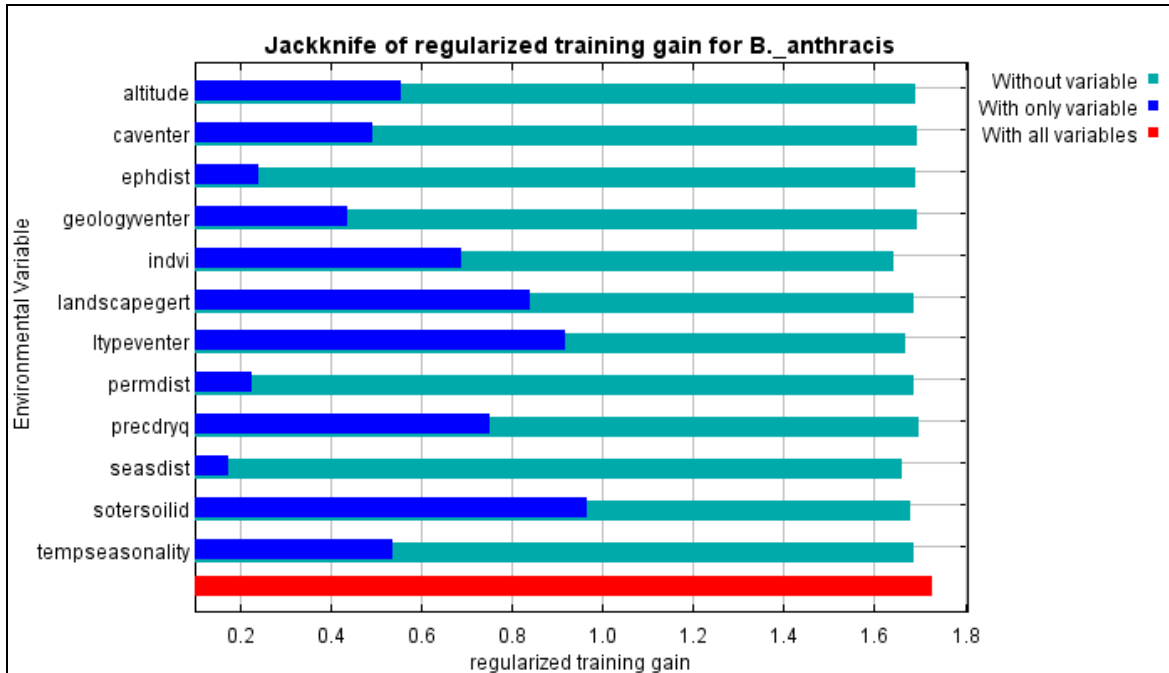


Figure 15: Jackknife results of the twelve variables used to construct the final model.

Figure 15 displays the variable importance as determined by the jackknife procedure.

Important variables can either have:

- 1) large dark blue bars, indicating strong (but perhaps non-unique) contribution to presences (see sotersoilid, Figure 15); or
- 2) short turquoise bars, indicating no other variable contains equivalent information (see indvi, Figure 15); or
- 3) both 1 and 2, indicating the variable is independently predictive (see ltypeventer, Figure 15).

The variable that had the highest gain when the jackknife procedure was applied was the SOTER classified soil class, indicating that this variable had the strongest contribution to presences.

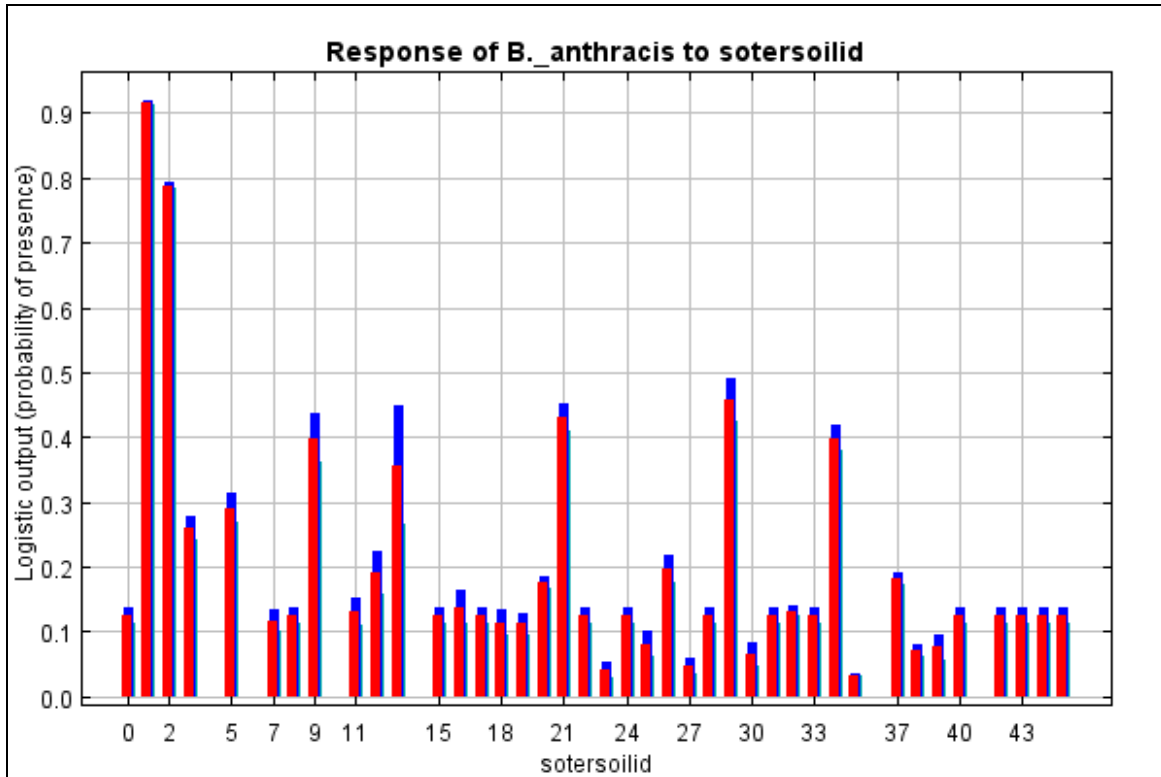


Figure 16: Probability of presence of the different classes that make up the sotersoilid variable. Red bars indicate the average value over the 10 model runs, dark blue bars indicate the maximum values and the turquoise bars indicate the minimum values.

Table 11: Sotersoilid class and name as listed in the SOTER database

0	1	2	3	4	5	6	7	8	9
ZW74	ZA21	ZA22	ZA29	ZA32	ZA41	MZ22	ZA55	ZA56	ZA62
10	11	12	13	14	15	16	17	18	19
ZA65	ZA80	ZA83	MZ223	ZA87	ZA98	ZA101	ZA115	ZA33	ZA145
20	21	22	23	24	25	26	27	28	29
MZ16	ZA160	ZA170	ZA189	ZA190	ZA202	ZA233	ZA245	ZA258	ZA262
30	31	32	33	34	35	36	37	38	39
ZA282	ZA283	ZA291	ZA314	ZA335	ZA357	ZA387	ZA450	ZA491	ZA533
40	41	42	43	44	45				
ZA548	ZA564	ZA604	ZA610	ZA634	ZA641				

According to Figure 16, the most important soil classes within the sotersoilid variable were 1 and 2, which are ZA21, followed by ZA22 (Table 11).

The variable that decreased the gain the most when omitted was the integrated NDVI value, which means that it had the most information that was not present in other variables.

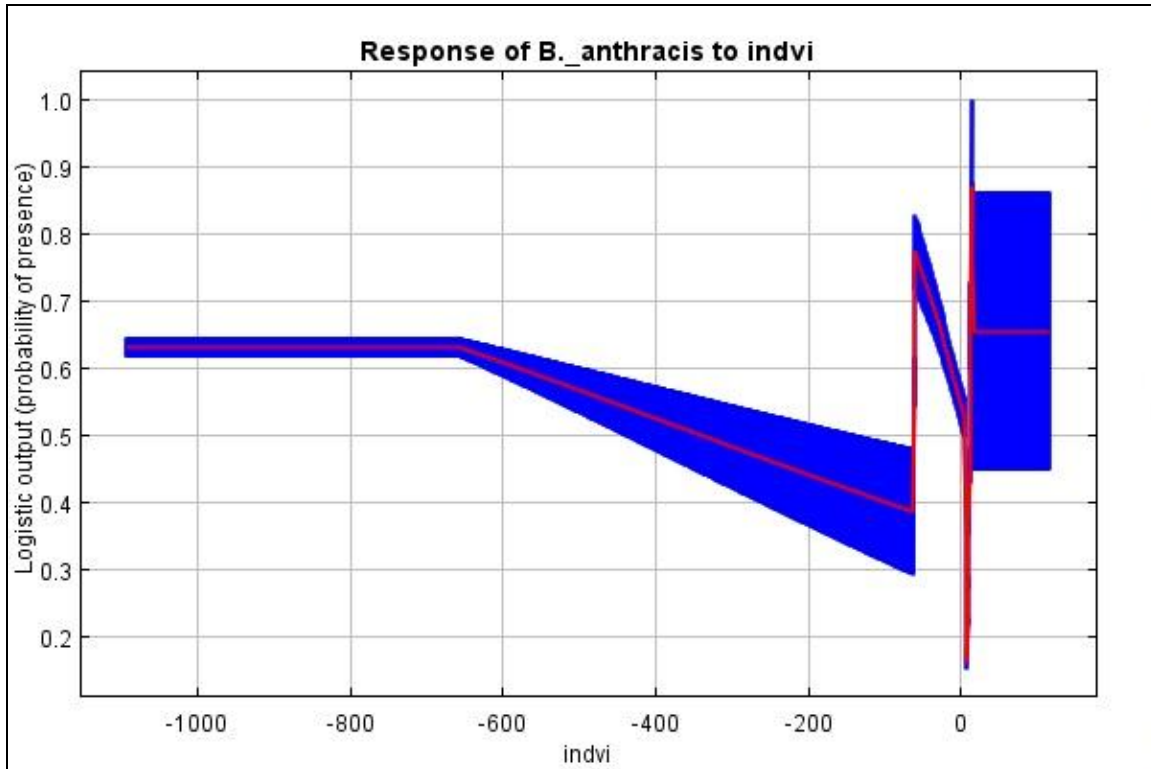


Figure 17: INDVI values for the 10 model runs. Red line indicates average and blue areas are deviations.

Figure 17 has a very well defined peak value corresponding to a high anthrax probability. This INDVI value is 8-10 (with 80-100% probability). INDVI values smaller than 8 display a lot of variation and INDVI values larger than 10 stay constant with an average probability of 65%.

Other important variables were the landscape as classified by Gertenbach (1983), land type as classified by Venter (1990) and driest quarter precipitation. The six most important variables were identified using unpartitioned data (all positive points used for training) to utilize maximum information.

According to the jackknife procedure, the six most important environmental predictors, (in descending order) were:

1. SOTER soil ID (sotersoilid)
2. Land type as defined by Venter (ltypeventer)

3. Landscape as defined by Gertenbach (landscapegert)
4. Precipitation of the driest quarter (precdryq)
5. Integrated NDVI value (INDVI)
6. Altitude (altitude)

This study found that the modelled niche for the *B. anthracis* spore can be defined by a narrow index of precipitation, NDVI and soil type. Soil type considers the land type, landscape and soil values. In addition to the above, Maxent response curves can be used to evaluate variable values for increased suitability (Appendix A, Figure 16 and Figure 17). Annual rainfall in the modelled areas ranged between 400-500mm and the integrated NDVI values ranged from 8 in the north to 11 in the south. This indicates a higher plant biomass toward the south (Pettorelli *et al.*, 2005). Thus a higher number of potential vectors for the dissemination of spores would most likely be found in the southern modelled areas.

According to this study the most suitable ecological conditions for anthrax were a combination of low rainfall, soil type, high calcium in the soil and high animal biomass as reflected by the NDVI variables.

Presence points from the 1959 outbreaks in KNP were plotted on the model output map (Figure 18). Figure 25 Map (a) contains the points of the first 1959 outbreak and upon visual inspection the model predicts the majority cases very well. Figure 18 Map (b) contains the points of the second 1959 outbreak and upon visual inspection only predicts some of the locations correctly. This model aims to find areas suitable for anthrax endemicity and because this was an epidemic outbreak it is unlikely to accurately predict all the epidemic areas.

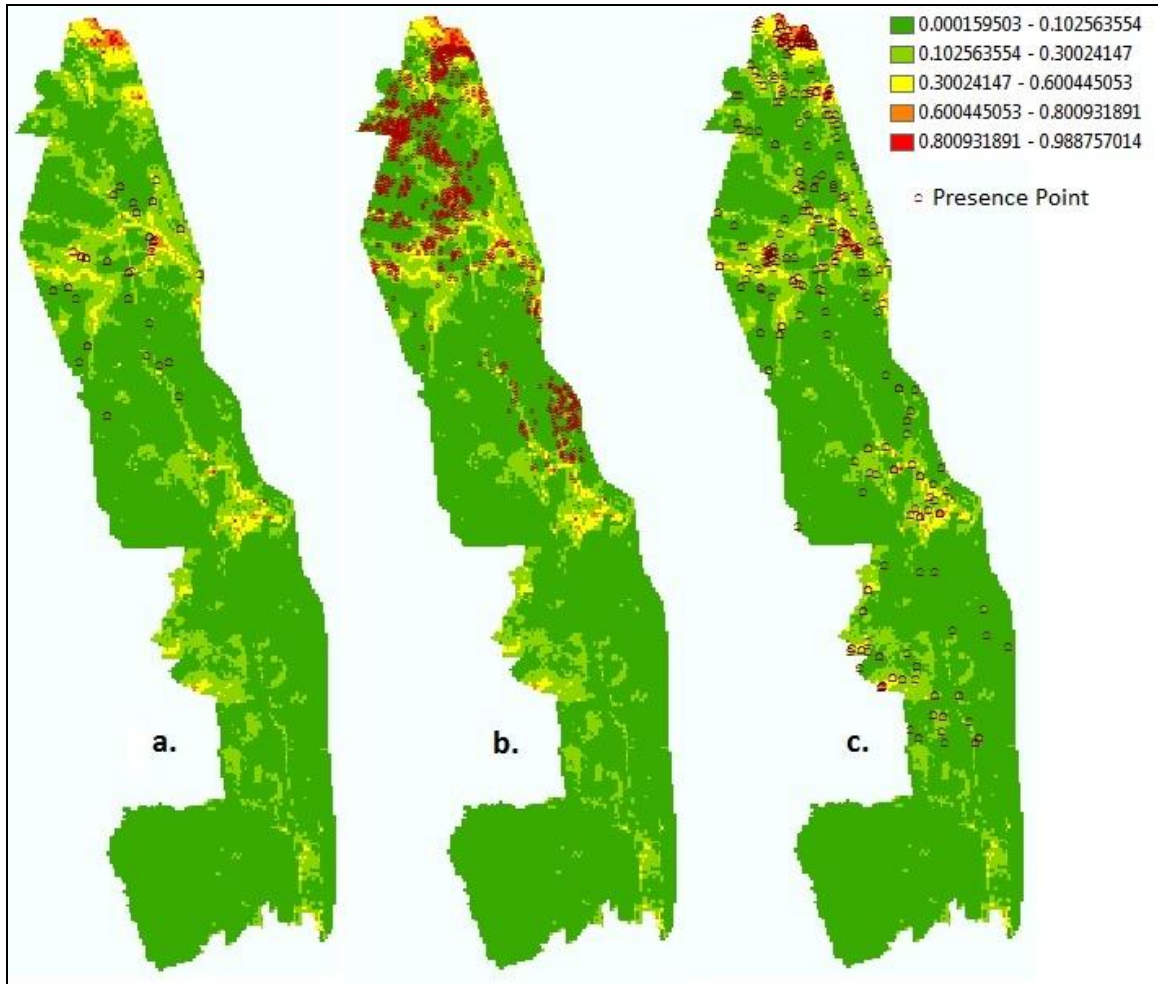


Figure 18: 1959 Outbreak data plotted against modelled suitability map. a. First outbreak of 1959 (Pienaar, 1960) b. Second outbreak of 1959 (Pienaar, 1961) c. Presence points used to train the model (1988 - 2011)

As mentioned earlier, two models were developed in this study – a 40 variable model and a final 12 variable model. In Figure 19, the output maps of the two models are displayed.

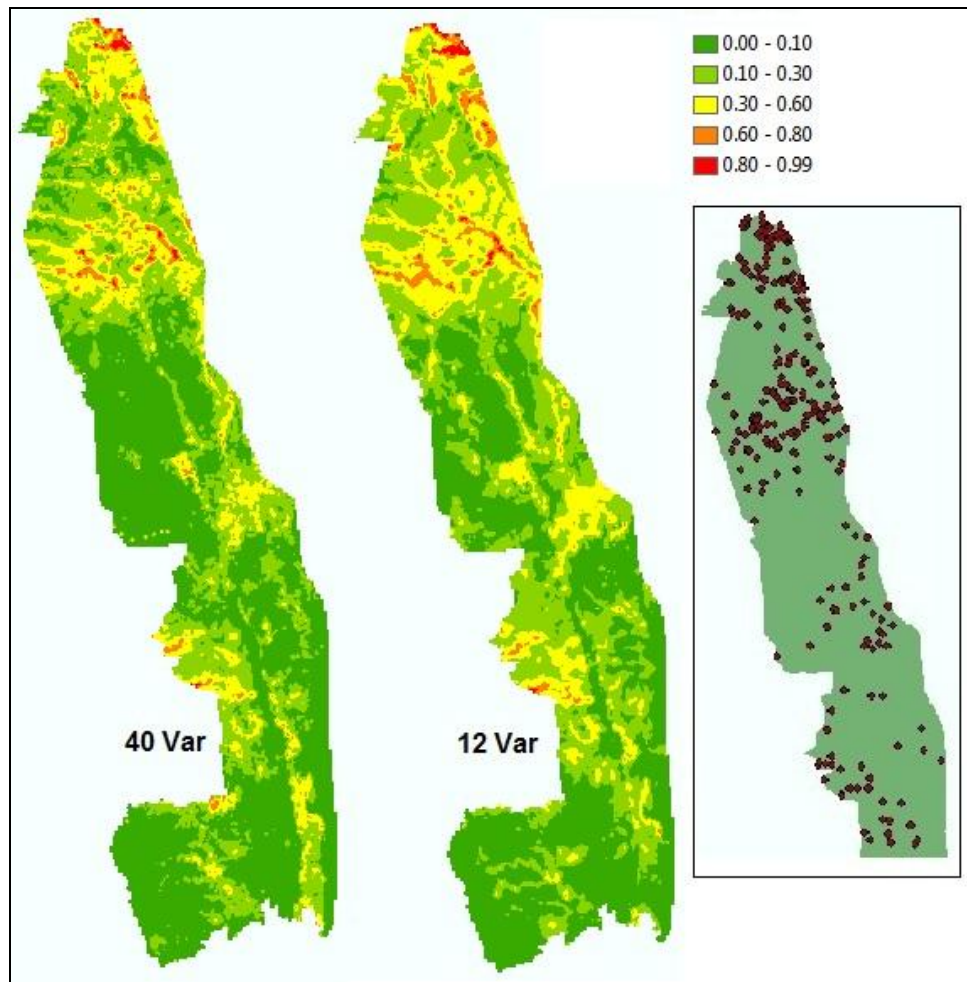


Figure 19: Two models were developed in this study – a preliminary 40 variable model and a final 12 variable model, 5000 iterations, with all presence points.

Upon visual inspection there is not much difference between the 40 and 12 variable model output maps. Both models predict the same areas as suitable and also predict the same threshold areas. Note the increase in suitability in the southern KNP with the 12 variable model. Based to the fact that the average suitability value in the southern area was still less than 30%, it can be concluded that there is no significant difference between the two output maps.

To determine the individual variable values of importance for ecological suitability, all points with a probability of occurrence higher than 80% were selected and the environmental variable values extracted (Table 12). According to Table 12, a low altitude, high soil calcium content and low driest quarter precipitation seem to play important roles in the suitability of the environment for anthrax. See Figure 14 for the map indicating these three areas.

Table 12: Threshold point values for the top 12 variables and the three areas identified by the model (Figure 14) as having the highest suitability for anthrax occurrence.

Section	Pafuri	Shingwedzi	Letaba	Study area mean (range)
<i>n</i>	186	2	30	597
Altitude (altitude)	228.8064516	281	225.9333333	292 (145 – 570)
Calcium (caventer)	171.835714	184.6258	198.0494133	148 (10 – 282)
Ephemeral Distance (ephdist)	2069.046753	0	13181.526	3618 (0 – 20384)
Seasonal Distance (seasdist)	3169.587167	0	7038.007	4124 (0 – 14764)
Permanent Distance (permdist)	2190.000172	46840.2	733.3333333	12461 (0 – 53450)
Geology Venter (geologyventer)	LB, CS, AL	AL	EC	LB, CS, AL, EC
INDVI (indvi)	8.58999014	9.36794	8.867802667	7.34 (-775 – 15)
Landscape (landscapegert)	25,28	35	21,22	21, 22, 25, 28, 35
Landtype	Pa04,Pa05	Le05	Le01	Le01, Le05, Pa04, Pa05

(ltypeventer)				
Precipitation Driest Quarter (precdryq)	9	14	23.73333333	14.7 (8 – 36)
SOTER Soil ID (sotersoilid)	ZA21,ZA22	ZA115	ZA282	ZA21, ZA22, ZA115, ZA282
Temperature Seasonality (tempseasonality)	3389.096774	3687	3759.933333	3439 (2838 – 3828)

Geology Legend (Gertenbach, 1983):

LB – Karoo system. Olivine rich basalts, sub-ordinate alkali basalts, shoshonites.

CS – Karoo system. Fine grained sandstone, mudstone, chert (Cave sandstone and redbed stages).

AL – Quarternary. Alluvium.

EC – Karoo system. Shale with coal seams, mudstone, grit (Ecca series).

Landscape Legend (Venter, 1990):

15 – *Colophospermum mopane* forest

21 – *Combretum* / *Acacia nigrescens* rugged veld

22 – *Combretum* / *Colophospermum mopane* rugged veld

25 – *Adansonia digitata* / *Colophospermum mopane* rugged veld

28 – Limpopo / Luvuvhu floodplain

35 – *Salvadora angustifolia* floodplains

SOTER Soil ID Legend (Dijkshoorn *et al.*, 2008):

ZA21, ZA115 – Eutric cambisols

A cambisol (CM) can be defined as having either a cambic or a mollic horizon.

A cambic horizon is a weakly developed mineral soil horizon and a mollic horizon is a surface horizon of mineral soil that is dark in colour, relatively deep and contains (dry weight) at least 1% organic matter or 0.6% organic carbon.

ZA22 – Eutric leptosols

A leptosol (LP) can be defined through (1) a limit in depth by continuous hard rock within 25 cm from the soil surface, (2) overlying material with a calcium carbonate equivalent of more than 40 percent within 25 cm from the soil surface or (3) less than 10 percent (by weight) fine earth to a depth of 75 cm or more from the soil surface.

ZA282 – Leptic phaeozems

Continuous rock starting between 50 and 100 cm from the soil surface with a mollic horizon and (1) a base saturation of 50 percent or more and no secondary carbonates, at least to a depth of 100 cm from the soil surface and (2) with no diagnostic horizons other than an albic, argic, cambic or vertic horizon.

Finally, the result of the gap analyses is displayed in Figure 20 below:

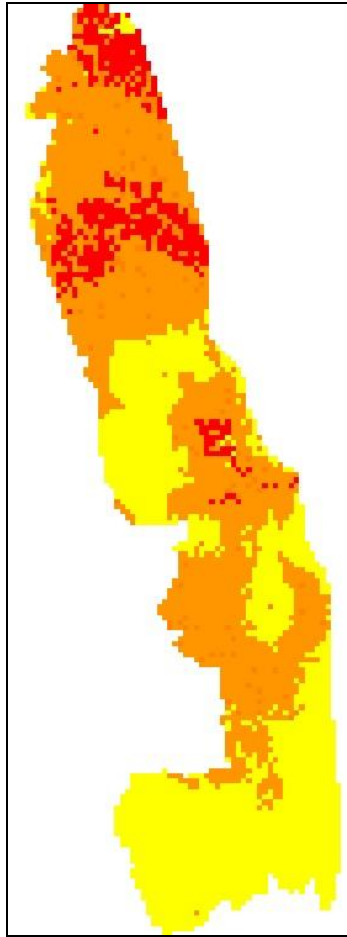


Figure 20: Gap analyses on Maxent model output. Red areas indicate areas where no presence points were used as training data, but that have a very high likelihood of producing positive cases. Thus red indicates highest sampling priority, followed by orange and lastly yellow.

4. Discussion

This study presents an estimate of ecologically suitable areas for anthrax occurrences throughout the KNP. According to the model, there are at least three geographically distinct regions within the Park – Pafuri, Shingwedzi and Letaba – that are highly suitable for spore dormancy and survival. The Pafuri ranger region is the only region to be described in the literature as endemic to anthrax. Since most of the positive occurrences from the dataset were from Pafuri, it was expected that the region will have a very high suitability for anthrax. The other two identified areas were predicted with a significantly smaller amount of positive occurrences. This can reflect reality, overestimate (for Pafuri), or underestimate (for the rest of KNP) anthrax occurrences due to bias in sampling effort and location.

In addition to sampling bias, lack of information from the occurrence data can have a negative impact on the model performance through the influence of environmental variables which contains large gradients in their values. For example, the environmental variable altitude will have a much higher value on a mountain. If the data point location is erroneously reported as on top of a mountain instead of at its foothills, it can have a negative (or false positive) impact on the predictive contribution of the altitude variable.

According to Van Ness (1971), the conditions favourable for the persistence of anthrax spores would be i) calcareous soils which were rich in nutrients and contained a high moisture content, ii) high soil pH levels (>7), iii) low lying depressions where water stagnated and organic matter decayed, iv) rocklands that were dried up rivers and v) hillside seeps where organic matter accumulated during run-off. Dragon and Rennie (1995), referred to these areas as 'storage areas'. The modelled predictions appear to follow associated rivers eastward, with the highest suitability points at river confluences. This is most likely due to the alluvial deposition of spores after flooding. Spores have a high surface hydrophobicity, allowing clumping in water and a high buoyant density, allowing clumped organic matter to float (Dragon and Rennie, 1995). This also contributes to spore concentration during run-off into stagnant pools.

Calcium (Ca) has been shown to be important for both spore germination and the maintenance of dormancy. The Pafuri, Shingwedzi and Letaba regions have high pH and Ca levels, corresponding to the requirement for the dormancy of soil spores as mentioned by Dragon and Rennie (1995). High levels of Ca in the soils may act as a buffer to the internal spore Ca supply and greatly extend its survivability (Dragon and Rennie, 1995). Anthrax spores have increased survivability in alkaline soil because the free Ca is readily available. As the soil pH increases above 7.2, due to additional soil Ca, the "free" Ca is not absorbed into the soil and can bind with other compounds (Spectrum Analytic Inc, 2012). Table 13 lists soil sample values for four locations in KNP indicating that high levels of Ca, pH and CEC (cation exchange capacity) in soil might support anthrax spores.

Table 13: Selected parameters from soil sample points in Kruger National Park indicating anthrax presence or absence (Dijkshoorn *et al.*, 2008).

Parameter*	Pafuri	Shingwedzi	Letaba	Pretoriuskop
Ca	255.3	254	221	30
pH	7.7	9	7.7	5
CEC	324.7	440	313	150
Na	1.7	10.9	2.7	10
Anthrax	Present	Present	Present	Absent

* Ca: Calcium; CEC: cation exchange capacity; Na: Sodium

Lower CEC soils hold less Ca, and high CEC soils hold more (Spectrum Analytic Inc, 2012). Note the very low CEC value for Pretoriuskop in Table 13 and the correlation to the absence of anthrax. Abnormally high levels, or application rates of other cations, in the presence of low to moderate soil Ca levels tends to reduce the uptake of Ca. Excess sodium (Na) in the soil competes with Ca, and other cations to reduce their availability (Spectrum Analytic Inc, 2012). Thus pH, CEC and Na parameters all influence the Ca concentration. These soil parameters correspond to results from previous studies which stated that anthrax spores need a high calcium level in the soil to persist for extended

periods (Van Ness, 1959a; Van Ness, 1971; De Vos, 1990; Dragon and Rennie, 1995; Hugh-Jones and Blackburn, 2009).

Contrary to the findings of Blackburn (2007) altitude did not play a major role in model prediction, which is likely due to the small variation in altitude within the KNP.

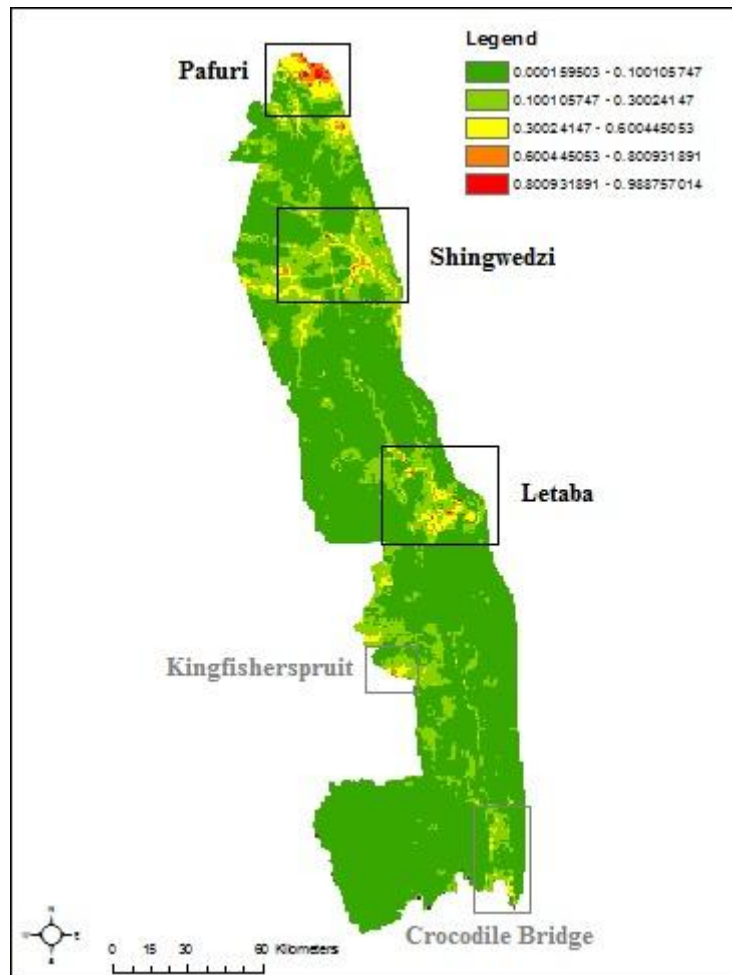


Figure 21: Suitable anthrax areas within Kruger National Park. Black rectangles indicate core areas (greater than 80% probability). Grey rectangles indicate notable areas (60% - 80% probability).

Core Sections (Figure 21)

The final model indicated areas with a probability of 80 – 100% of having suitable ecological conditions for anthrax. These areas were:

1. Northern Pafuri

2. Mpongolo-Shingwedzi confluence
3. Letaba-Olifants area

The Northern Pafuri depression

This area (Figure 22) is classified by Venter (1990) as the Pafuri Land Type (Pa05) and is located in the Pafuri ranger section of KNP. It's endemicity to anthrax has been described by De Vos (1990).

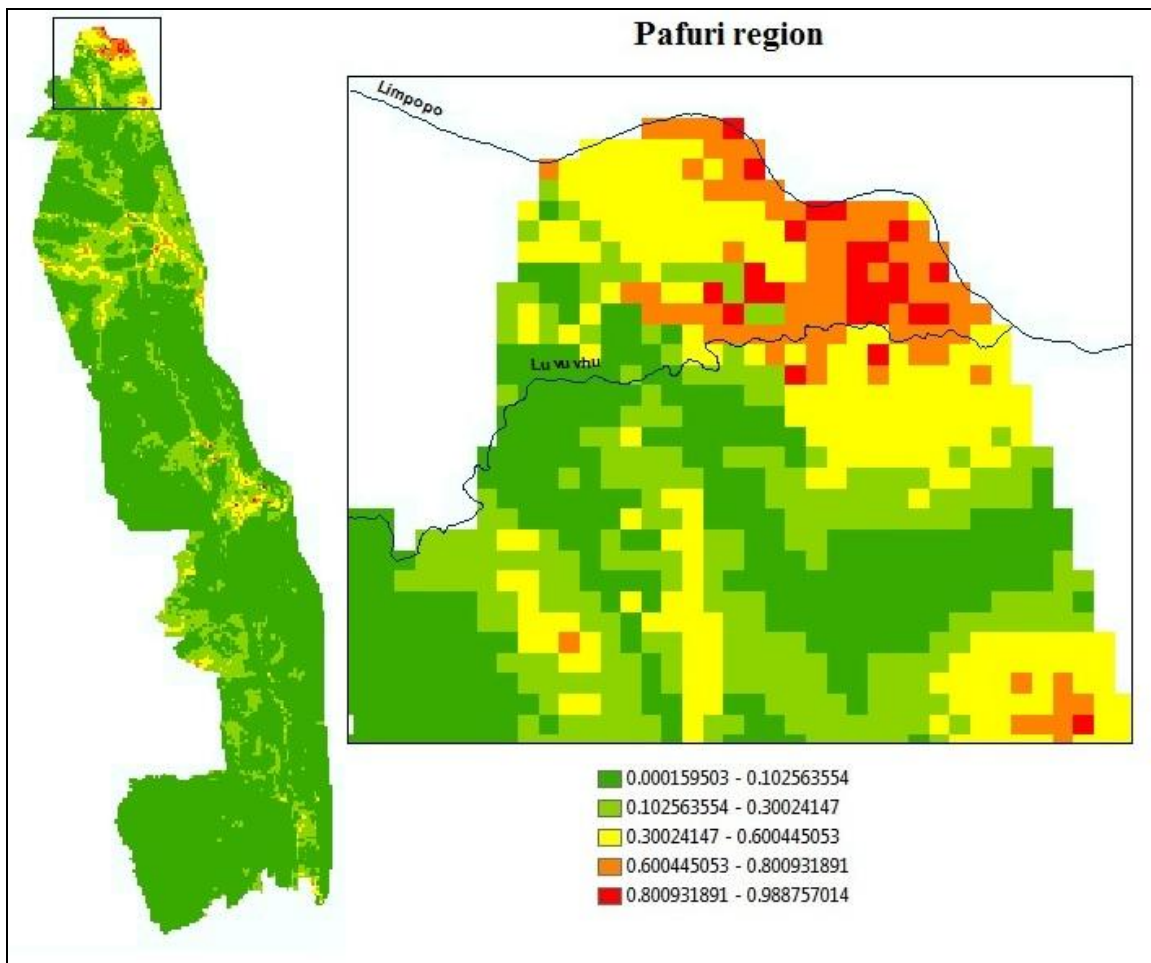


Figure 22: Pafuri region suitability map for *Bacillus anthracis*. This map depicts the predicted area in more detail.

Alluvial lowlands flank the lower Luvuvhu and Limpopo rivers with sandy to deep red silt sediments. Several large seasonal pans occur in this area and large floodplains are present. Pans are replenished by run-off water after heavy rains. Oakleaf and Valsrivier

soil forms dominate these floodplains. The average rainfall for the area is 400mm per year. Table 14 indicates favourable soil parameters for the survival of *B. anthracis* spores where the soil properties of anthrax endemic Pafuri land type is compared with unfavourable soil properties on non-anthrax Pretoriuskop land type (Sk01).

Table 14: Selected soil properties of Pafuri land type (Pa05) where anthrax is endemic compared with anthrax unfavourable Pretoriuskop land type (Sk01) (Venter, 1990)

Property*	Value Pa05 (Pafuri)	Value Sk01 (Pretoriuskop)
Clay (%)	31 – 33	20 – 40
CEC (me/kg)	275 – 450	150
pH	8.25 – 10.5	4.5 – 5.5
K (me/kg)	3 – 15	1 – 1.5
Ca (me/kg)	260 – 280	20 – 50
Mg (me/kg)	36 – 46	10 – 30
Na (me/kg)	0 – 3	5 – 15
Phosphorus (mg/kg)	8 – 12	5 – 9

* CEC (Cation exchange capacity), K (Potassium), Ca (Calcium), Mg (Magnesium), Na (Sodium).

Mpongolo-Shingwedzi confluence

This area (Figure 23) is classified by Venter (1990) as the Shingwedzi Land Type (Le05) and includes the Shingwedzi river and its major tributaries Bububu, Phugwane, Mpongolo and Nkokodzi. Significant alluvial deposits occur along the Shingwedzi river system. The area is characterized by peripheral incised areas with shallow soils, which slope down gently towards the rivers. Undulating landforms along rivers are frequent, with flat alluvial plains alongside drainage channels. Very high pH and Ca levels occur along valley bottoms with predominantly dense and heterogeneous riverine vegetation. The Valsrivier soil form is most prevalent and *Colophospermum mopane* is the most dominant woody plant species in the valley bottom hillslope unit.

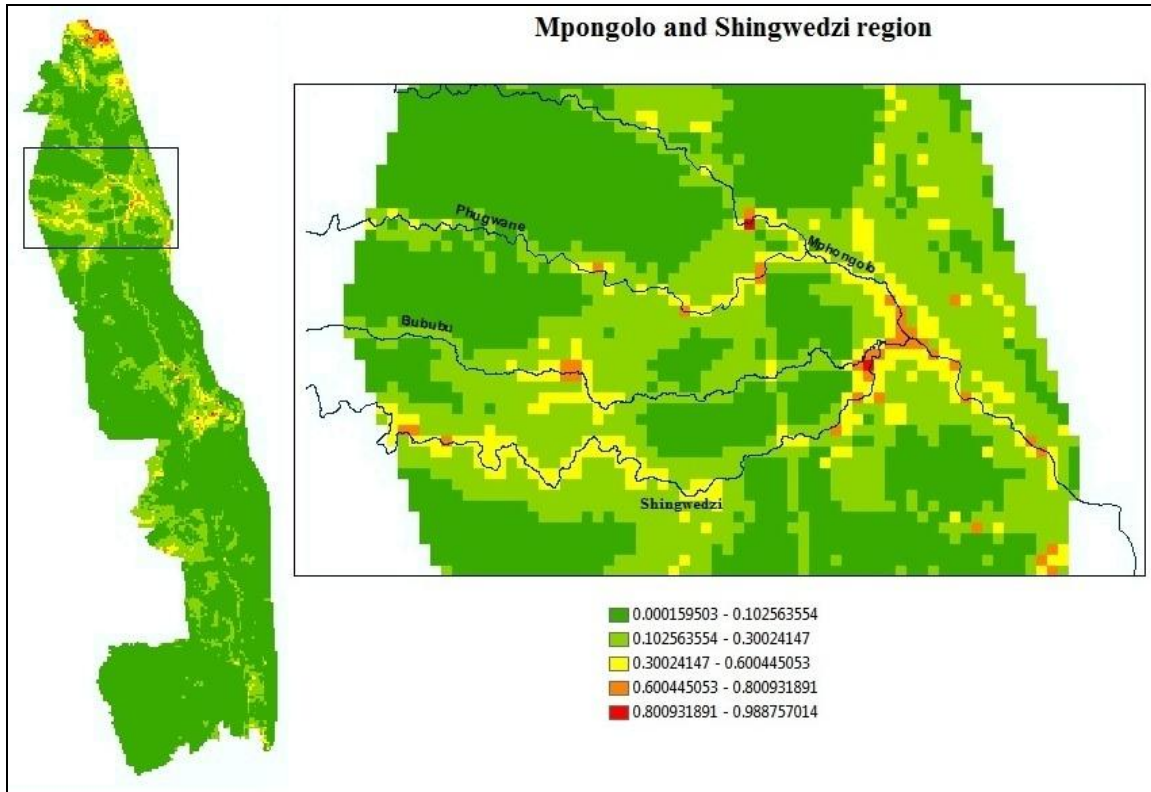


Figure 23: Shingwedzi region suitability map for *Bacillus anthracis*. This map depicts the predicted area in more detail.

The Shingwedzi land type with its high pH and Ca levels (Table 15), corresponding to the requirement for the occurrence of soil spores compared to the low pH and Ca level of unfavourable conditions of the Pretoriuskop land type (Sk01) as indicated by the model.

Table 15: Selected soil properties of Shingwedzi land type (Le05) favourable for anthrax compared to unfavourable anthrax Pretoriuskop land type (Sk01) (Venter, 1990)

Property	Value Le05 (Mpongolo-Shingwedzi confluence)	Value Sk01 (Pretoriuskop)
Clay (%)	33 – 35	20 – 40
CEC (me/kg)	440 – 550	150
pH	8.8 - 10.3	4.5 – 5.5
K (me/kg)	18 – 23	1 – 1.5
Ca (me/kg)	255 – 275	20 – 50
Mg (me/kg)	35 – 55	10 – 30

Na (me/kg)	10 – 65	5 – 15
Phosphorus (mg/kg)	8 – 70	5 – 9

Letaba-Olifants area

This area (Figure 24) is classified by Venter (1990) as the Letaba landtype (Le02) and is characterized by extensive, flat to strongly undulating plains with calcareous soils, dominated by *C. mopane* woody vegetation.

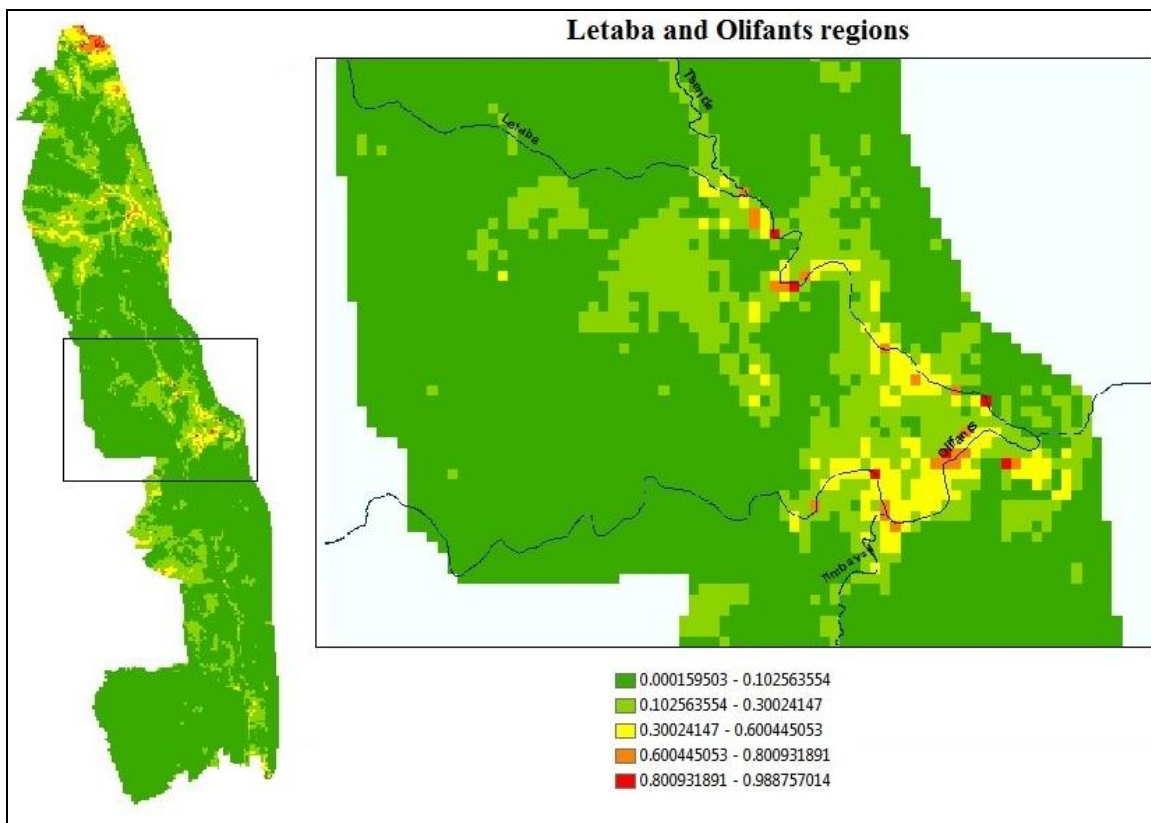


Figure 24: Letaba-Olifants region suitability map for *Bacillus anthracis*. This map depicts the predicted area in more detail.

Fairly extensive alluvial deposits occur along the Letaba river, consisting of mainly Oakleaf and Valsrivier forms. Once again the high pH, Ca and CEC of Letaba land type indicates favourable soil parameters for the survival of *B. anthracis* spores (Table 16).

Table 16: Selected soil properties of Letaba land type (Le02) favourable for anthrax compared to unfavourable anthrax Pretoriuskop land type (Sk01) (Venter, 1990)

Property	Value Le02 (Letaba-Olifants Area)	Value Sk01 (Pretoriuskop)
Clay (%)	33 – 35	20 – 40
CEC (me/kg)	425 – 470	150
pH	7.8 – 8	4.5 – 5.5
K (me/kg)	3 – 4.5	1 – 1.5
Ca (me/kg)	230 – 275	20 – 50
Mg (me/kg)	52 – 60	10 – 30
Na (me/kg)	4.4 – 4.6	5 – 15
Phosphorus (mg/kg)	0 – 3	5 – 9

A striking feature of all three these suitable areas is the high Ca and pH values (similar to De Vos (1990); Dragon and Rennie (1995)), indicating potentially favourable soil conditions for anthrax spores. The pH layers used implicitly in this model, surprisingly did not play a significant role in predicting suitability, but this could be due to lack of information at the spatial scale at which they were employed in this model.

Noteworthy sections (Figure 21)

The model indicated areas with a probability of 60 – 80% of having suitable ecological conditions for anthrax in the following ranger sections:

1. Kingfisherspruit
2. Lower Sabie and Crocodile Bridge

Kingfisherspruit

This area (Figure 25) is classified by Venter (1990) as the Orpen landtype (Sa05). Soils in this area are shallow to moderately deep, black, occasionally calcareous clay (Mayo, Bonheim, Milkwood and Arcadia soil forms).

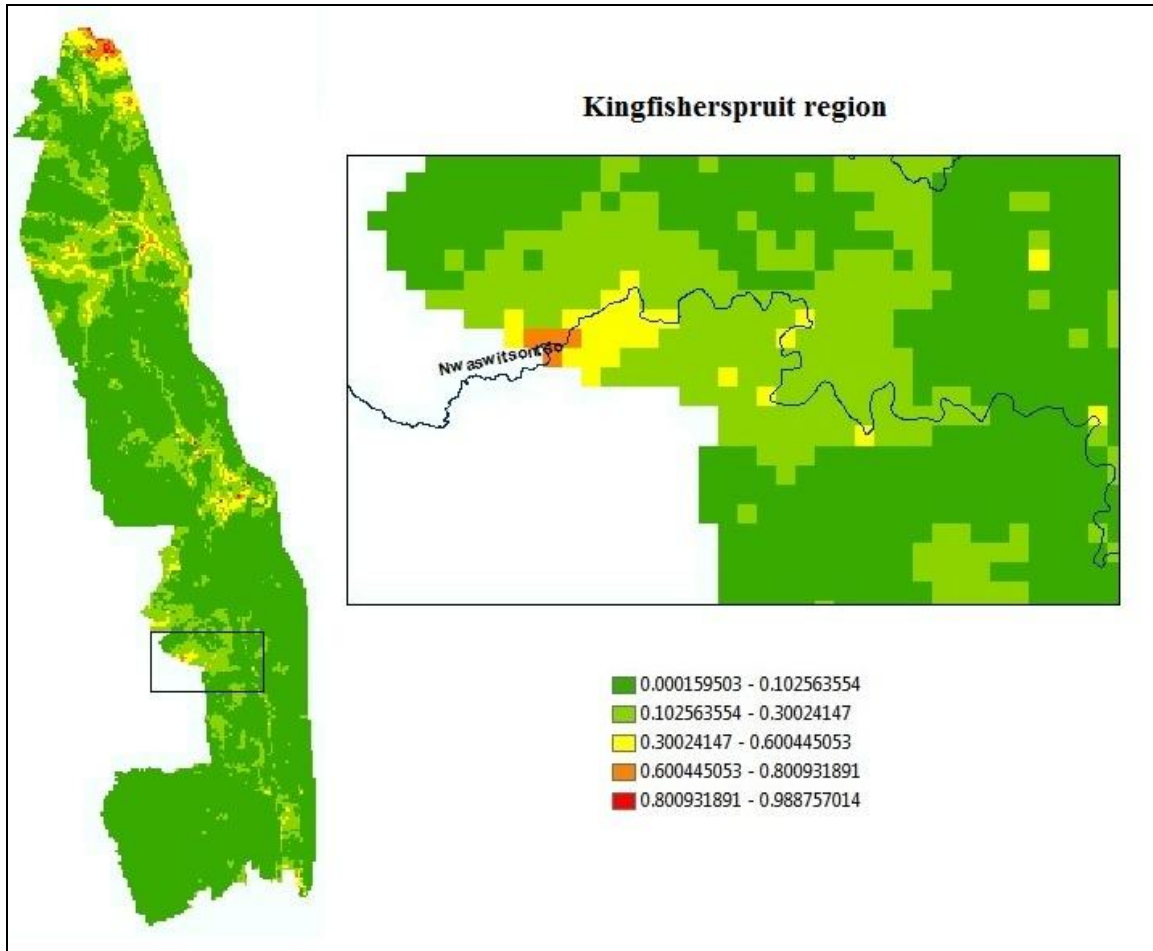


Figure 25: Kingfisherspruit region suitability map for *Bacillus anthracis*. This map depicts the predicted area in more detail.

The black soils are associated with olivine gabbro and *Acacia nigrescens* is the dominant woody plant species. Table 17 indicates high pH, Ca and CEC of Orpen land type compared to Pretoriuskop land type that is unsuitable for anthrax.

Table 17: Selected soil properties of Orpen land type (Sa05) suitable ecological conditions for anthrax compared to unfavourable anthrax Pretoriuskop land type (Sk01) (Venter, 1990)

Property	Value (Kingfisherspruit) Sa05	Value (Pretoriuskop) Sk01
Clay (%)	25 – 40	20 – 40
CEC (me/kg)	250 – 350	150
pH	7.0 – 8.0	4.5 – 5.5

K (me/kg)	5 – 10	1 – 1.5
Ca (me/kg)	150 – 250	20 – 50
Mg (me/kg)	40 – 45	10 – 30
Na (me/kg)	2 – 4	5 – 15
Phosphorus (mg/kg)	2 – 5	5 – 9

Lower Sabie and Crocodile Bridge

This area (Figure 26) is classified by Venter as the Satara landtype (Sa01) and is characterized by Olivine poor areas with shallow red and brown, paraduplex clay (Shortlands and Swartland forms).

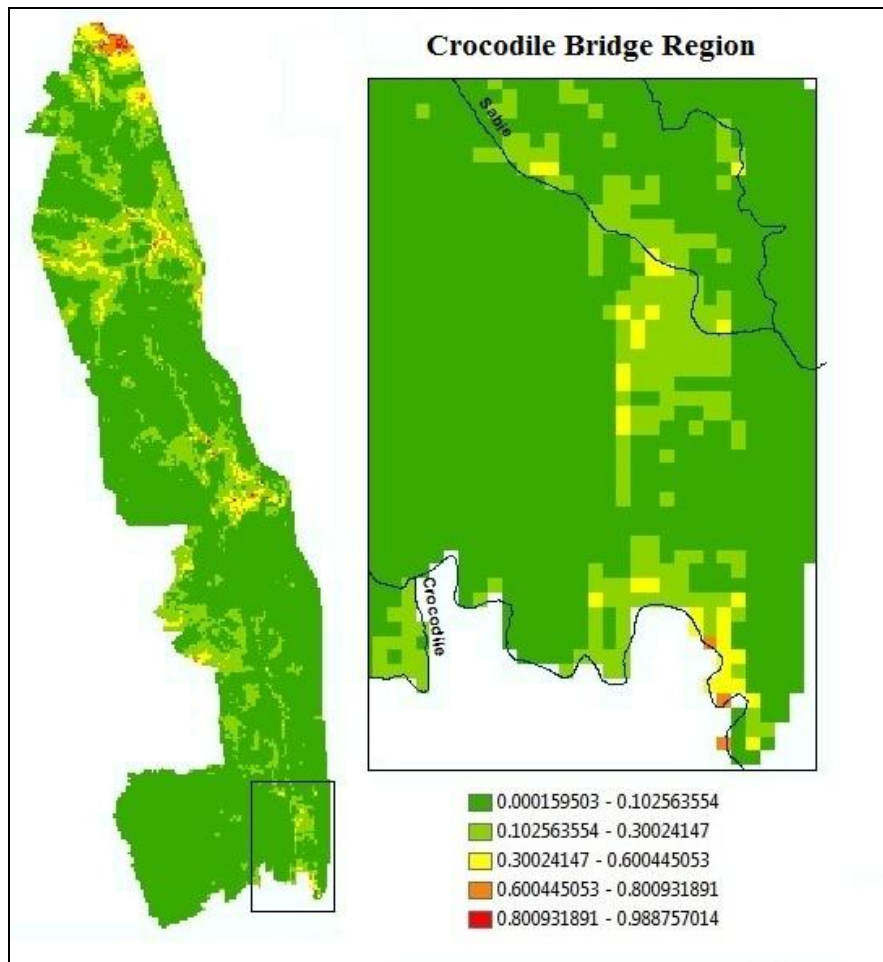


Figure 26: Crocodile Bridge region suitability map for *Bacillus anthracis*. This map depicts the predicted area in more detail.

Vegetation is dominated by open *Sclerocarya birrea* / *Acacia nigrescens* tree savanna. Calcareous soils are restricted to depressions and valley bottoms in this land type. The altitude of the area ranges from 140 m to 310 m. Table 18 shows that the area contains very high soil Na values. Since Na competes with Ca, it decreases the probability of extended spore survival (Spectrum Analytic Inc, 2012).

Table 18: Selected soil properties of Satara land type (Sa01) suitable ecological conditions for anthrax compared to unfavourable anthrax Pretoriuskop land type (Sk01) (Venter, 1990)

Property	Value Sa01 (Lower Sabie / Crocodile Bridge)	Value Sk01 (Pretoriuskop)
Clay (%)	50 – 60%	20 – 40
CEC (me/kg)	35 - 45	150
pH	7 – 8.2	4.5 – 5.5
K (me/kg)	5 - 6	1 – 1.5
Ca (me/kg)	185 - 240	20 – 50
Mg (me/kg)	58 - 95	10 – 30
Na (me/kg)	100 - 225	5 – 15
Phosphorus (mg/kg)	5 - 10	5 – 9

Another factor to consider in the distribution of anthrax within KNP is the specific strain involved (Smith *et al.*, 2000). The A type strain is more prevalent in the north of the Park and also less pathogenic, while the B type is more prevalent in the central KNP. The A type is able to withstand adverse environmental conditions for longer periods of time than type B. Type B is also associated with significantly higher soil Ca levels than type A. Determining the spore type along the Letaba-Olifants regions can give an indication of endemicity of the type B strain. A separate model for ecological suitability should be created for the two types since type B requires even more specific environmental conditions.

None of the evaluation methods employed in this study can be used in isolation to measure model performance (Pearce and Boyce, 2006). Several methods, including AUC, null model evaluation and threshold dependant binomial statistics, were used to assess

whether the model predicts better than random. Results of all the methods indicated that the model predicts suitable habitat for anthrax survival significantly better than random ($p < 0.05$).

A prominent feature of all the models was the little difference variable removal made to the model outcome, suggesting that only a few of the variables made a high contribution to the final prediction. These variables were precipitation during the driest quarter (precdryq), landscape as defined by Gertenbach (landscapegert), land type as defined by Venter (ltypeventer) and the SOTER soil class (sotersoilid).

5. Synthesis

5.1. Conclusion

The aim this study was to identify and map areas within the KNP that were ecologically suitable for the harbouring of *B. anthracis* spores within the soil. This was achieved using maximum entropy as a statistical model, a range of environmental predictors and provided anthrax occurrence data. A regularized training gain of 1.9254 was achieved and a bootstrapped AUC of 0.9372. This research yielded a distribution model with a good fit to the sample data and a good performance on test data.

Three regions within the KNP have been identified and described as ecologically suitable for the long term survival of *B. anthracis* spores. The only area that has historically been described as endemic to anthrax in the KNP was the northern Pafuri (De Vos, 1990). Comparison of the environmental conditions in Pafuri with those at the other suitable sites revealed very similar ecological parameters.

The most useful predictors were found to be land type as classified by Venter, the large integral NDVI values as indicators of plant biomass, the soil class and the precipitation during the driest quarter of the year.

Finally, a gap analyses revealed the areas where future surveillance efforts should receive priority. The results of this study concurs that the northern Pafuri region is endemic to anthrax, but in addition, also provides at least two further potential areas for anthrax endemicity, namely the Mpongolo-Shingwedzi confluence and the Letaba-Olifants river region.

5.2. Recommendations and future work

- Compare the results of the Maxent model with the results of other modelling techniques, e.g. GARP (Stockwell and Peters, 1999).
- Use location data from negative cases in the dataset as “absence” records to build presence/absence models, using e.g. DOMAIN (Carpenter *et al.*, 1993)

- Strain typing on future cases from the Letaba region to determine if the type B strain is endemic in the area.
- Collection of soil Ca and pH data as part of anthrax surveillance data will significantly enhance the accuracy of this model. Soil data is needed for every sampling point.
- Use the proposed sampling areas as a starting point for future anthrax surveillance efforts.
- Model the potential distribution of anthrax in KNP under multiple climate change scenarios (Joyner *et al.*, 2010)

The most important factor for improving the model accuracy is improvement of locality data of occurrence records, which includes specific GPS coordinates and intensification of passive surveillance efforts.

6. References

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

ANDERSON, R.P., GÓMEZ-LAVERDE, M. & PETERSON, A.T. 2002. Geographical distributions of spiny pocket mice in South America: insights from predictive models. *Global Ecology and Biogeography*, 11, 131–141.

APACHE SOFTWARE FOUNDATION. 2012. Apache OpenOffice 3.4.1., Apache Software Foundation, Available at: www.openoffice.org.

ARAÚJO, M.B. & GUISAN, A. 2006. Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, 33, 1677-1688.

AUSTIN, M. 2007. Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Journal of Ecological Modeling*, 200, 1-19.

BALDWIN, R.A. 2009. Use of maximum entropy modeling in wildlife research. *Entropy*, 11, 854-866.

BATJES, N.H. 2004. *SOTER-based soil parameter estimates for Southern Africa*. 04. Wageningen: ISRIC - World Soil Information.

BLACKBURN, J.K., MCNYSET, K.M., CURTIS, A. & HUGH-JONES, M.E. 2007. Modeling the geographic distribution of *Bacillus anthracis*, the causative agent of anthrax disease, for the contiguous united states using predictive ecological niche modeling. *American Journal of Tropical Medicine Hygiene*, 77(6), 1103-1110.

- BRAACK, L.E.O. & RETIEF, P.F. 1986. Dispersal, density and habitat preference of the blowflies *Chrysomya marginalis* Diptera Calliphoridae. *Onderstepoort Journal of Veterinary Research*, 53, 13-18.
- BRAACK, L.E.O. & TESKE, R.T. 1997. *A Visitor's guide to Kruger National Park*. South Africa: Struik Publishers.
- BURNHAM, K.P. & ANDERSON, D.R. 2001. Kullback-Leibler information as a basis for strong inference in ecological studies. *Wildlife Research*, 28, 111-119.
- BUSBY, J.R. 1986. A biogeographical analysis of *Nothofagus cunninghamii* (Hook.) Oerst. in southeastern Australia. *Australian Journal of Ecology*, 11, 1-7.
- CARPENTER, G., GILLISON, A.N. & WINTER, J. 1993. DOMAIN: a flexible modeling procedure for mapping potential distributions of plants, animals. *Biodiversity and Conservation*, 2, 667-680.
- CHENG, J. 2007. Modelling and understanding multi-temporal land use changes. *Archives*, 1955, 2-7.
- COSTANZA, R. & RUTH, M. 2001. Dynamic systems modeling. In: *Ecosystems and Sustainability*. Boca Raton, FL, USA: Lewis Publishers.
- CRAMER, J.S. 2003. *Logit models: from economics and other fields*. Cambridge University Press.
- DE VOS, V. 1990. The Ecology of Anthrax in the Kruger National Park, South Africa. *Salisbury Medical Bulletin*, 68, 19-23.
- DE VOS, V. & BRYDEN, H.B. 1996. Anthrax in the Kruger National Park: temporal and spatial patterns of disease occurrence. *Salisbury Medical Bulletin*, 87, 26-30.
- DE VOS, V. & TURNBULL, P.C.B. 1994. Anthrax. In: COETZER, J.A.W., THOMSON, G.R. & TUSTIN, R.C. (Eds). *Infectious diseases of livestock with special*

reference to Southern Africa. 2nd edn. Cape Town: Oxford University Press Southern Africa, 1788-1818.

DIJKSHOORN, J.A., VAN ENGELEN, V.W.P. & HUTING, J.R.M. 2008. *Soil and landform properties for LADA partner countries (Argentina, China, Cuba, Senegal and The Gambia, South Africa and Tunisia)*. Wageningen: ISRIC - World Soil Information and FAO.

DORMANN, C.F., ELITH, J., BACHER, S., BUCHMANN, C., CARL, G., CARRÉ, G., MARQUÉZ, J.R.G., GRUBER, B., LAFOURCADE, B., LEITÃO, P.J., MÜNKEMÜLLER, T., MCCLEAN, C., OSBORNE, P.E., REINEKING, B., SCHRÖDER, B., SKIDMORE, A.K., ZURELL, D. & LAUTENBACH, S. 2012. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, doi: 10.1111/j.1600-0587.2012.07348.x

DORMANN, C.F., MCPHERSON, J.M., ARAÚJO, M.B., BIVAND, R., BOLLIGER, J., CARL, G., DAVIES, R.G., HIRZEL, A., JETZ, W., KISSLING, W.D., KÜHN, I., OHLEMÜLLER, R., PERES-NETO, P.R., REINEKING, B., SCHRÖDER, B., SCHURR, F.M. & WILSON, R. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, 30, 609-628.

DRAGON, D.C. & RENNIE, R.P. 1995. The ecology of anthrax spores: tough but not invincible. *The Canadian Veterinary Journal*, 36, 295-301.

DUDÍK, M., PHILLIPS, S.J. & SCHAPIRE, R.E. 2005. Correcting sample selection bias in maximum entropy density estimation. In: *Advances in neural information processing systems*. Massachusetts, USA: MIT Press, Cambridge.

ELITH, J., GRAHAM, C.H., ANDERSON, R.P., DUDÍK, M., FERRIER, S., GUISAN, A., HIJMANS, R.J., HUETTMANN, F., LEATHWICK, J.R., LEHMANN, A., LI, J., LOHMANN, L.G., LOISELLE, B.A., MANION, G., MORITZ, C., NAKAMURA, M., NAKAZAWA, Y., OVERTON, J.M., PETERSON, A.T., PHILLIPS, S.J., RICHARDSON, K.S., SCACHETTI-PEREIRA, R., SCHAPIRE, R.E., SOBERON, J.,

WILLIAMS, S., WISZ, M.S. & ZIMMERMANN, N.E. 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29, 129-151.

ELITH, J., PHILLIPS, S.J., HASTIE, T., DUDÍK, M., CHEE, Y.E. & YATES, C.J. 2011. A statistical explanation of Maxent for ecologists. *Diversity and Distributions*, 17, 43-57.

ELLER, S.P. & SEIFU, Y. 2002. Using spatial statistics to select model complexity. *Journal of Computational & Graphical Statistics*, 11, 348-369.

EPISTIS. 2012. Epi-STIS Project. Research programme for earth observation “STEREO II”. Support to the exploitation and research in earth observation, SR/00/102. Available at: <http://www.belspo.be/belspo/fedra/proj.asp?l=en&COD=SR/00/102#docum>

ESRI. 2012. ArcGIS Desktop, Redlands, CA: Environmental Systems Research Institute.

FAWCETT, T. 2006. An introduction to ROC analysis. *Pattern Recognition Letters*, 27, 861-874.

FIELDING, A.H. & BELL, J.F. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, 24, 38-49.

FRANKLIN, J. 2009. *Mapping Species Distributions. Spatial Inference and Prediction*. New York: Cambridge University Press.

GERTENBACH, W.P.D. 1983. Landscapes of the Kruger National Park. *Koedoe - African Protected Area Conservation and Science*, 26, 9-121.

GRAHAM, C.H., ELITH, J., HIJMANS, R.J., GUISAN, A., PETERSON, A.T., LOISELLE, B.A. & THE NCEAS PREDICTING SPECIES DISTRIBUTIONS WORKING GROUP. 2008. The influence of spatial errors in species occurrence data used in distribution models. *Journal of Applied Ecology*, 45, 239-247.

GRENDÁR, M. & GRENDÁR, M. 2001. Maximum Entropy: Clearing up Mysteries. *Entropy*, 3, 58-63.

GUISAN, A., EDWARDS, T.C. & HASTIE, T. 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling*, 157, 89-100.

GUO, Q. & LUI, Y. 2010. ModEco: an integrated software package for ecological niche modeling. *Ecography*, 33, 1-6.

HANLEY, J.A. & MCNEIL, B.J. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143, 29-36.

HANNART, P. & HUGHES, D.A. 2003. A desktop model used to provide an initial estimate of the ecological instream flow requirements of rivers in South Africa. *Journal of Hydrology*, 270, 167-181.

HERKT, M. 2007. *Modelling Habitat Suitability To Predict The Potential Distribution Of Erhard's Wall Lizard Podarcis erhardii On Crete*. MSc Thesis, International Institute for Geo-Information Science and Earth Observation (ITC).

HIJMANS, R.J., CAMERON, S.E., PARRA, J.L., JONES, P.G. & JARVIS, A. 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25, 1965-1978.

HIJMANS, R.J., GARRETT, K.A., HUAMAN, Z., ZHANG, D.P., SCHREUDER, M. & BONIERBALE, M. 2000. Assessing the geographic representativeness of genebank collections: the case of Bolivian wild potatoes. *Conservation Biology*, 14, 1755-1756.

HIJMANS, R.J., GUARINO, L. & MATHUR, P. 2012. Diva-GIS, University of California.

HIJMANS, R.J., PHILLIPS, S., LEATHWICK, J. & ELITH, J. 2012. Species distribution modeling with R. Available at: cran.r-project.org/web/packages/dismo/vignettes/sdm.pdf [Accessed on 9/15, 2012].

HILL, T. & LEWICKI, P. 2007. *STATISTICS: Methods and applications*. Tulsa, OK: StatSoft.

HIRZEL, A.H., HAUSSER, J., CHESSEL, J. & PERRIN, N. 2002. Ecological niche factor analysis: how to compute habitat suitability maps without absence data? *Ecology*, 87, 2027-2036.

HUGH-JONES, M. & BLACKBURN, J. 2009. The ecology of *Bacillus anthracis*. *Molecular aspects of medicine*, 30, 356-367.

HUGH-JONES, M. & DE VOS, V. 2002. Anthrax and wildlife. *Rev. sci. tech. Off. int. Epiz.*, 21(2), 359-383.

HUNTLEY, B., BERRY, P.M., CRAMER, W. & MCDONALD, A.P. 1995. Modelling present and potential future ranges of some European higher plants using climate response surfaces. *Journal of Biogeography*, 22, 967-1001.

HUTCHINSON, G.E. 1957. Concluding remarks. *Cold Spring Harbor symposia on quantitative biology*, 22, 415-427.

JÖNSSON, P. & EKLUNDH, L. 2004. TIMESAT - a program for analyzing time-series of satellite sensor data. *Computers & Geosciences*, 30, 833-845.

JOYNER, T.A., LUKHNOVA, L., PAZILOV, Y., TEMIRALYEVA, G., HUGH-JONES, M.E., AIKIMBAYEV, A. & BLACKBURN, J.K. 2010. Modeling the Potential Distribution of *Bacillus anthracis* under multiple climate change scenarios for Kazakhstan. *PLoS ONE*, 5(3), e9596. doi:10.1371/journal.pone.0009596.

KEITT, T.H., BJORNSTAD, O.N., DIXON, P.M. & CITRON-POUSTY, S. 2002. Accounting for spatial pattern when modeling organism-environment interactions. *Ecography*, 25, 616-625.

KLOPPERS, J.J. & BORNMAN, H. 2005. *A Dictionary of Kruger National Park Place Names*. SA Country Life.

LEGENDRE, P. 1993. Spatial autocorrelation - trouble or new paradigm. *Ecology*, 74, 1659-1673.

LIU, C., BERRY, P.M., DAWSON, T.P. & PEARSON, R.G. 2005. Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, 28, 385-393.

LOBO, J.M., JIMENEZ-VALVERDE, A. & REAL, R. 2007. AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17, 145-151.

MANEL, S., DIAS, J.M. & ORMEROD, S.J. 1999. Comparing discriminant analysis, neural networks and logistic regression for prediction species distributions: a case study with a Himalayan river bird. *Ecological Modelling*, 120, 337-347.

MANEL, S., WILLIAMS, H.C. & ORMEROD, S.J. 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, 38, 921-931.

MERCKX, B., STEYAERT, M., VANREUSEL, A., VINCX, M. & VANAVERBEKE, J. 2011. Null models reveal preferential sampling, spatial autocorrelation and overfitting in habitat suitability modelling. *Ecological Modelling*, 222, 588-597.

MICROSOFT CORPORATION. 2010. Microsoft Excel, Part of Microsoft Office 2010 Professional Edition.

NAIMI, B., SKIDMORE, A., GROEN, T. & HAMM, N. 2011. Spatial autocorrelation in predictors reduces the impact of positional uncertainty in occurrence data on species distribution modelling. *Journal of Biogeography*, 38(8), 1497-1509.

NIX, H. 1986. *A biogeographical analysis of Australian elapid snakes*. Canberra, Australia: Australian Government Publishing Service.

NIX, H., MCMAHON, J. & MACKENZIE, D. 1977. Potential areas of production and the future of pigeon pea and other grain legumes in Australia. In: WALLIS, E.S. & WHITEMAN, P.C. (Eds). *The potential for pigeon pea in Australia. Proceedings of Pigeon Pea (Cajanus cajan (L.) Millsp.) Field Day*. Queensland, Australia: University of Queensland.

PEACE PARKS FOUNDATION. 2008. *Greater Limpopo Trans-frontier Park Priority, Banhine and Kruger National Park (West) Dataset. Final Data Report and Meta Data (version 1)*.

PEARCE, J. & BOYCE, M. 2006. Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology*, 43, 405-412.

PEARSON, R.G. 2007. Species' distribution modeling for conservation educators and practitioners. Synthesis. American Museum of Natural History. Available at: <http://ncep.amnh.org>.

PEARSON, R.G., DAWSON, T.P. & LIU., C. 2004. Modelling species distributions in Britain: A hierarchical integration of climate and land-cover data. *Ecography*, 27, 285-298.

PEARSON, R.G., THUILLER, W., ARAÚJO, M.B., MARTINEZ-MEYER, E., BROTONS, L., MCCLEAN, C., MILES, L., SEGURADO, P., DAWSON, T.P. & LEES, D. 2006. Model-based uncertainty in species' range prediction. *Journal of Biogeography*, 33, 1704-1711.

PETERSON, A.T. & COHOON, K.P. 1999. Sensitivity of distributional prediction algorithms to geographic data completeness. *Ecological Modelling*, 117, 159-164.

PETTORELLI, N., VIK, J., MYSTERUD, A., GAILLARD, J.M., TUCER, C. & STENSETH, N. 2005. Using the satellite-derived NDVI to assess ecological responses to environmental change. *Trends in Ecology & Evolution*, 20, 503-510.

PHILLIPS, S.J., ANDERSON, R.P. & SCHAPIRE, R.E. 2006. Maximum entropy modeling of species geographic distributions. *Journal of Ecological Modeling*, 190, 231-259.

PHILLIPS, S.J., DUDÍK, M., ELITH, J., GRAHAM, C.H., LEHMANN, A., LEATHWICK, J. & FERRIER, S. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, 19(1), 181-197.

PHILLIPS, S.J., DUDÍK, M. & SCHAPIRE, R.E. 2004. A Maximum Entropy Approach to Species Distribution Modeling. *Proceedings of the Twenty-First International Conference on Machine Learning*, 662-665.

PIENAAR, U.De V. 1960. \n Uitbraak van miltsiekte onder wild in die Nasionale Kruger Wildtuin 28.9.59 tot 20.11.59. *Koedoe - African Protected Area Conservation and Science*, 3(1), 238-251.

PIENAAR, U.De V. 1961. A Second Outbreak Of Anthrax Amongst Game Animals in the Kruger National Park. 5th June to 11th October, 1960. *Koedoe - African Protected Area Conservation and Science*, 4(1), 4-17.

R CORE DEVELOPMENT TEAM. 2008. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

RAES, N. & TER STEEGE, H. 2007. A null-model for significance testing of presence-only species distribution models. *Ecography*, 30, 727-736.

REESE, G.C., WILSON, K.R., HOETING, J.A. & FLATHER, C.H. 2005. Factors affecting species distribution predictions: a simulation modeling experiment. *Ecological Applications*, 15, 554-564.

ROBERTSON, M.P., CAITHNESS, N. & VILLET, M.H. 2001. A PCA-based modeling technique for predicting environmental suitability for organisms from presence records. *Diversity and Distributions*, 7, 15-27.

SCHAPIRE, R. 2003. The boosting approach to machine learning - an overview. In: DENISON, D.D., HANSEN, M.H., HOLMES, C., MALLICK, B. & YU, B. (Eds). *MSRI Workshop on Nonlinear Estimation and Classification, 2002*. New York: Springer, 1-23.

SCHELDEMAN, X. & VAN ZONNEVELD, M. 2010. *Training Manual on Spatial Analysis of Plant Diversity and Distribution*. Rome, Italy: Bioersivity International.

SMITH, K.L., DE VOS, V., BRYDEN, H., PRICE, L.B., HUGH-JONES, M.E. & KEIM, P. 2000. *Bacillus anthracis* diversity in Kruger National Park. *Journal of clinical microbiology*, 38(10), 3780-3784.

SPECTRUM ANALYTIC INC. 2012. Calcium (Ca⁺⁺). Spectrum Analytic Inc. Available at: http://www.spectrumanalytic.com/doc/library/articles/ca_basics [Accessed on 9/15, 2012].

STATA CORP. 2001. STATA/SE STATISTICAL, 12.1. College Station, Texas: Stata Corporation, Available at: www.stata.com.

STOCKWELL, D. & PETERS, D. 1999. The GARP modeling system: problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science*, 13, 143-158.

STOCKWELL, D.R.B. & NOBLE, I.R. 1992. Induction of sets of rules from animal distribution data: a robust and informative method of data analysis. *Mathematics and Computers in Simulation*, 33, 385-390.

TRAVERS, R. 2005. Borehole closures in the Kruger National Park. *Borehole Water Journal*, 59, 8-15.

USGS. 2012. United States Geological Survey. Science for a changing world. Available at: <http://earthexplorer.usgs.gov>. [Accessed on 8/12, 2012].

VAN NESS, G.B. 1959a. Anthrax - a soil borne disease. *Soil Conservation*, 21, 206-208.

VAN NESS, G.B. 1959b. Soil relationship in the Oklahoma-Kansas anthrax outbreak of 1957. *Journal of Soil and Water Conservation*, 1, 70-71.

VAN NESS, G.B. 1971. Ecology of Anthrax. *Science*, 172, 1303-1307.

VAN NESS, G.B. & STEIN, C.D. 1956. Soils of the United States favorable for anthrax. *Journal of the American Veterinary Medical Association*, 128, 7-9.

VENTER, F.J. 1990. *A classification of land for management planning in the Kruger National Park*. PhD Thesis, University of Pretoria.

VENTER, F.J., SCHOLES, R.J. & ECKHARDT, H.C. 2003. The abiotic template and its associated vegetation pattern. In: DU TOIT, J.T., ROGERS K.H. & BIGGS H.C. (Eds). *The Kruger Experience: Ecology and management of savanna heterogeneity*. Washington, DC: Island Press. 83-129

VERBRUGGEN, H. 2012. RasterTools: MMS.pl. Available at: <http://www.phycoweb.net/software> [Accessed on 9/28, 2012].

WARREN, D.L., GLOR, R.E. & TURELLI, M. 2008. Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. *Evolution*, 62, 2868-2883.

WINTLE, B.A., ELITH, J. & POTTS, J. 2005. Fauna habitat modelling and mapping in an urbanising environment: a case study in the Lower Hunter Central Coast region of NSW. *Austral Ecology*, 30, 729-748.

YEE, T.W. & MITCHELL, N.D. 1991. Generalized additive models in plant ecology. *Journal of Vegetation Science*, 2, 587-602.

YOST, A.C., PETERSEN, S.L., GREGG, M. & MILLER, R. 2008. Predictive modeling and mapping sage grouse (*Centrocercus urophasianus*) nesting habitat using Maximum Entropy and a long-term dataset from Southern Oregon. *Ecological Informatics*, 3, 375-386.

YOUNG, N., CARTER, L. & EVANGELISTA, P. 2011. A Maxent Model v3.3.3e Tutorial (ArcGIS v10). [ColoradoView], [Online]. Available at: http://ibis.colostate.edu/WebContent/WS/ColoradoView/TutorialsDownloads/A_Maxent_Model_v7.pdf [Accessed on 6/12, 2012].

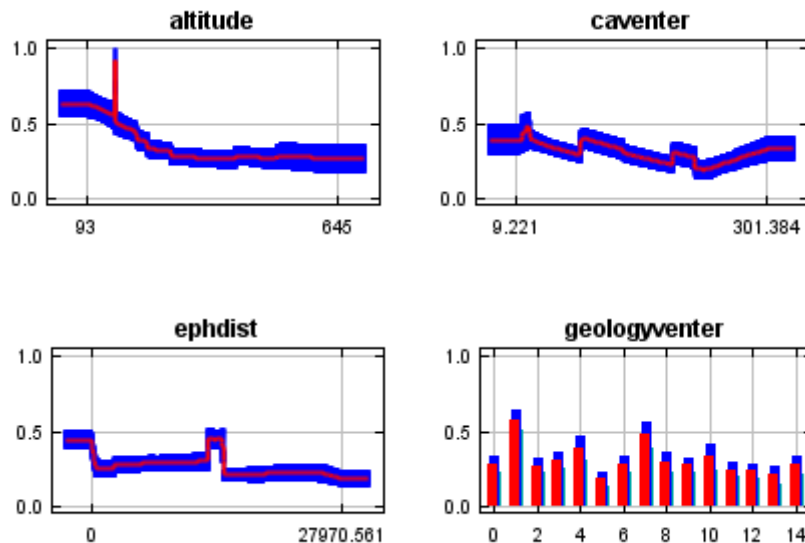
7. Appendices

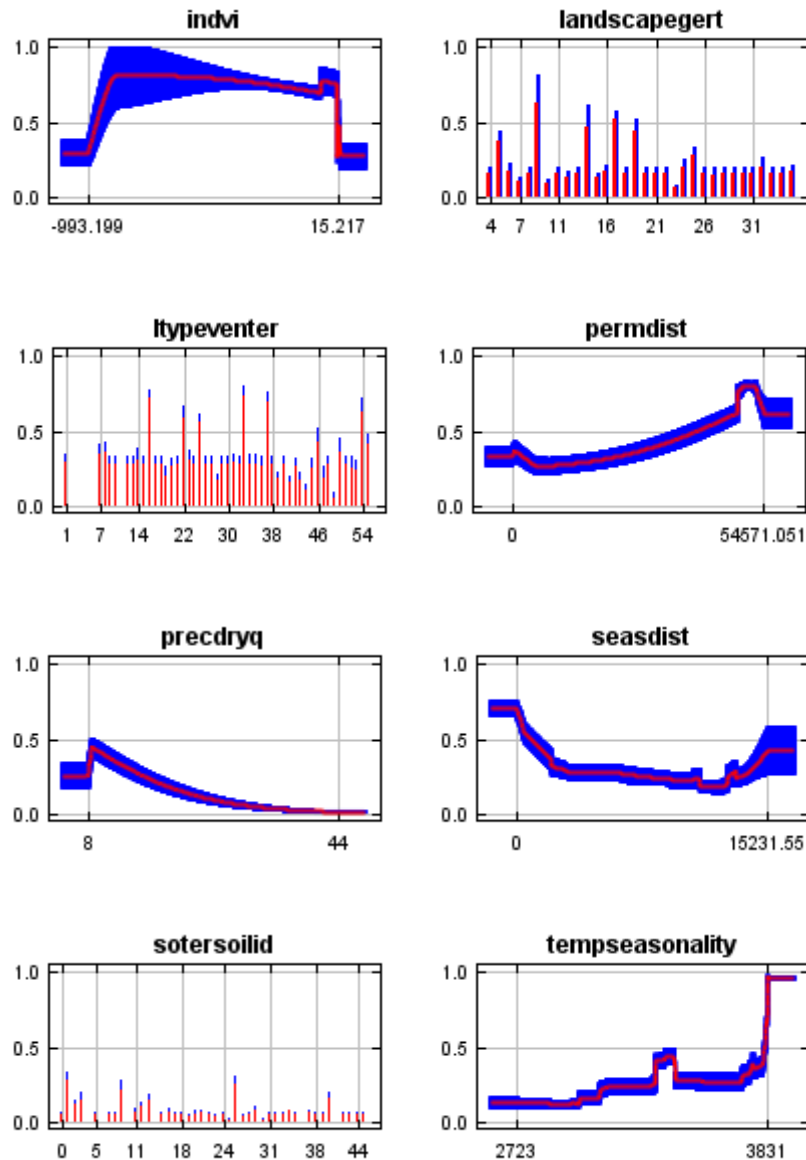
Appendix A

Maxent Variable Response Curves

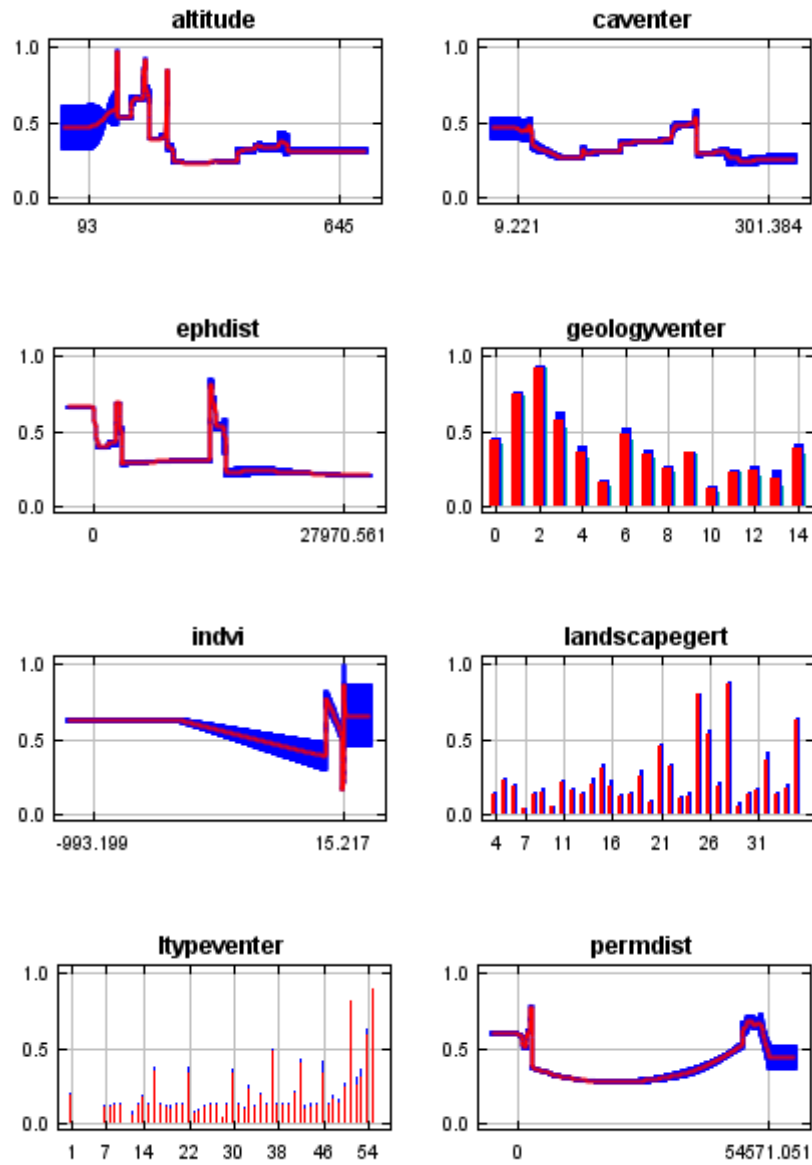
Results of B._anthracis.html

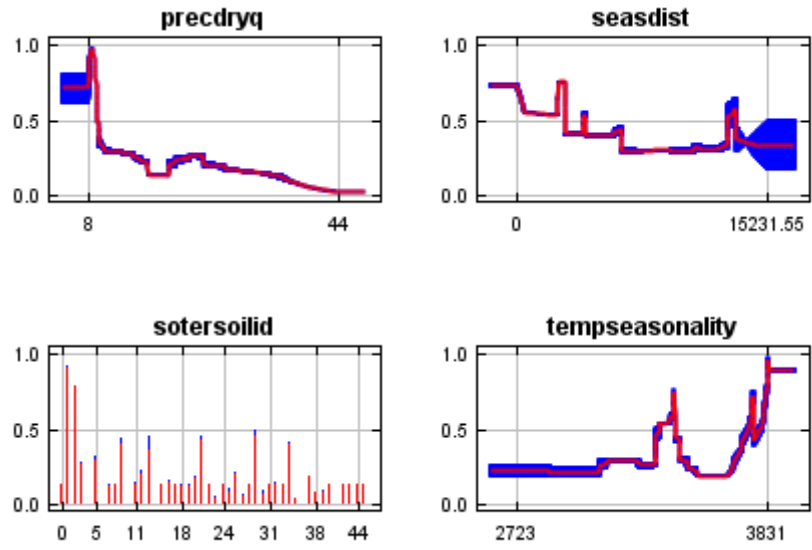
These curves show how each environmental variable affects the Maxent prediction. The curves show how the logistic prediction changes as each environmental variable is varied, keeping all other environmental variables at their average sample value. Note that the curves can be hard to interpret if you have strongly correlated variables, as the model may depend on the correlations in ways that are not evident in the curves. In other words, the curves show the marginal effect of changing exactly one variable, whereas the model may take advantage of sets of variables changing together (Phillips, 2006). The red in the figures below indicate the mean value, while the blue indicate variation around the mean.





In contrast to the above marginal response curves, each of the following curves represents a different model, namely, a Maxent model created using only the corresponding variable. These plots reflect the dependence of predicted suitability both on the selected variable and on dependencies induced by correlations between the selected variable and other variables. They may be easier to interpret if there are strong correlations between variables (Phillips, 2006).



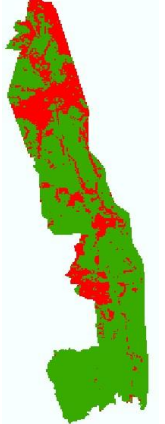



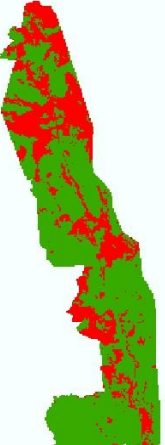





Appendix B

Model output maps for different threshold values

Binary output maps after applying a threshold. Red indicates suitable, green indicates unsuitable. For an explanation of each value see Table 2.

			
80% Threshold	0% Threshold	10% Threshold	
Fixed value	Lowest predicted value (sensitivity = 1)	Fixed sensitivity	Sensitivity – specificity equality
			
	0.54 Kappa with 56% threshold, calculated using ModEco (Guo and Lui, 2010)	16% Threshold	50% Threshold
Sensitivity-specificity sum maximization	Maximize Kappa	Average probability/suitability	Equal prevalence

Appendix C

Borehole Closures in KNP

BOREHOLE	SECTION	DRILL DATE	DATE CLOSED
Mavumbye	Satara	1950-01-01	1972
Machayipan	Pafuri	1961-01-01	1980
Nsemane	Satara	1950-01-01	1982
Rhilazeni	Satara	1950-01-01	1982
Sweni	Nwanetsi	1950-01-01	1982
Olienhoutfontein	Pretoriuskop	1976-11-02	1989
Bvumanyundu	Pafuri	1964-09-08	1990
Rhidonda	Phalaborwa	1975-01-01	1991
Mack	Crocodile Bridge	1976-08-31	1992
Rietpan	Tshokwane	1983-01-01	1994
Metsimetsi	Tshokwane	1971-07-12	1995
Ribbokrand	Tshokwane	1973-01-01	1995
Ruigtevlei	Skukuza	1975-04-25	1995
Koorsboom	Pafuri	1980-03-01	1996
Kremetart	Pafuri	1975-09-13	1996
Buffeldoring	Crocodile Bridge	1973-01-01	1998
Bejane	Skukuza	1971-08-01	1999
Biyamite	West Pretoriuskop	1965-07-01	1999
Jock	Malelane	1973-01-01	1999
Kirkman	Pretoriuskop	1950-01-01	1999
Lushof	Tshokwane	1976-08-17	1999

Manyahule	Skukuza	1970-08-01	1999
Mavukani	Stolsnek	1965-10-01	1999
Mikstok	Stolsnek	1976-10-16	1999
Mlambane West	Stolsnek	1973-01-01	1999
Môrester	Pretoriuskop	1976-10-28	1999
Newu	Stolsnek	1965-11-01	1999
Ngwenyeni	Stolsnek	1965-07-01	1999
Nkombanine	Stolsnek	1969-07-01	1999
Nwatindlopfu N	Tshokwane	1962-07-01	1999
Peru North	Houtboschrand	1958-01-01	1999
Sithungwane	Pretoriuskop	1965-07-01	1999
Vutomi Boloop	Tshokwane	1969-07-01	1999
Shitlhave	Pretoriuskop	1965-07-01	1999

Appendix D

Software images

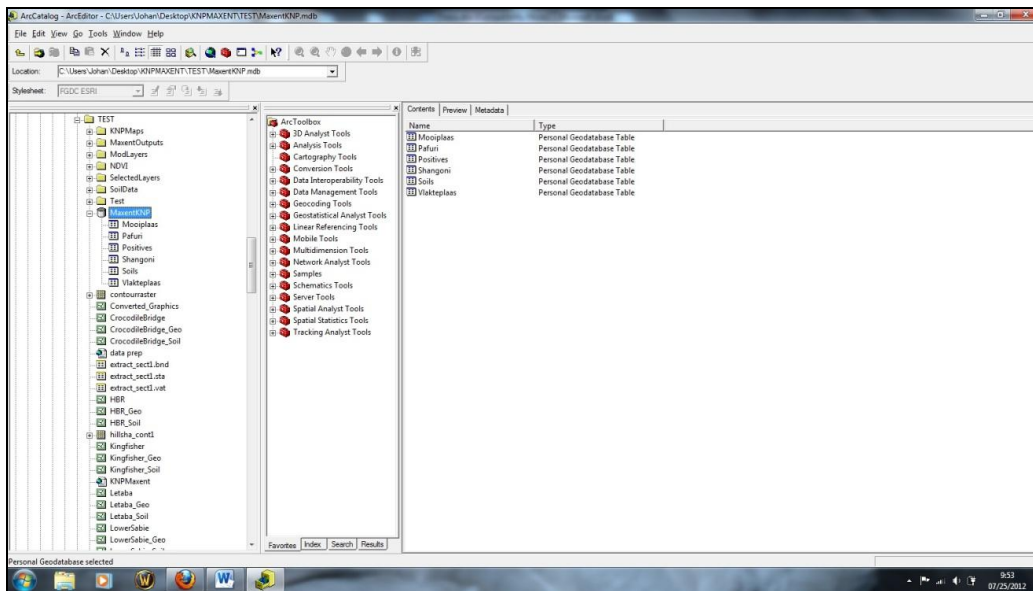


Figure 27: Creation of a personal geodatabase in ArcCatalog to enable addition of individual records of anthrax cases.

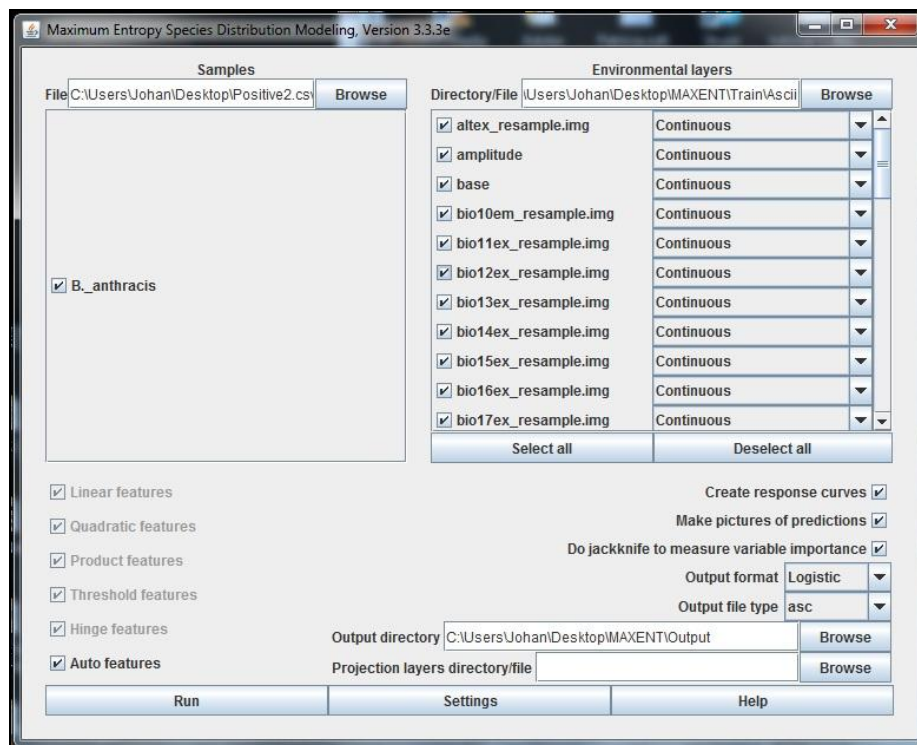


Figure 28: Maxent GUI, main screen indicating features selected.

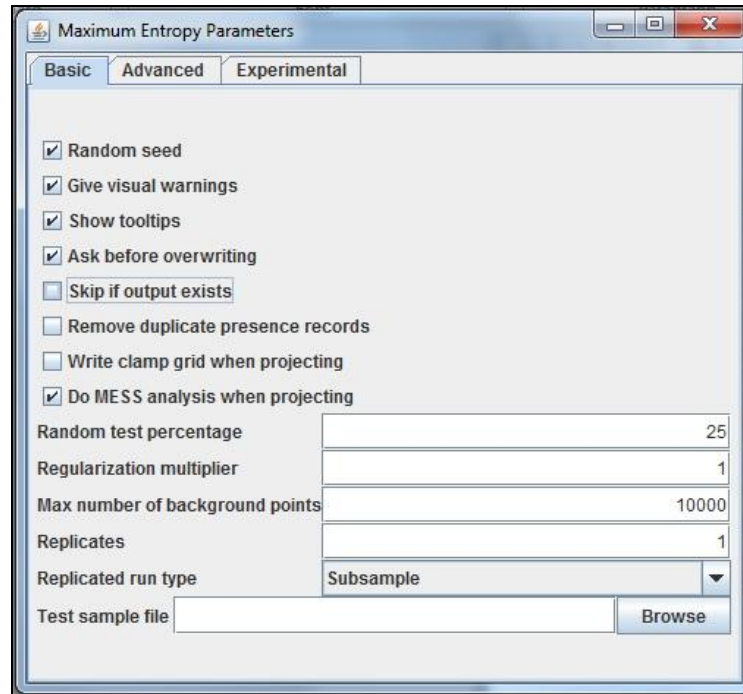


Figure 29: Basic options selected in Maxent. Note the 25 random test percentage.

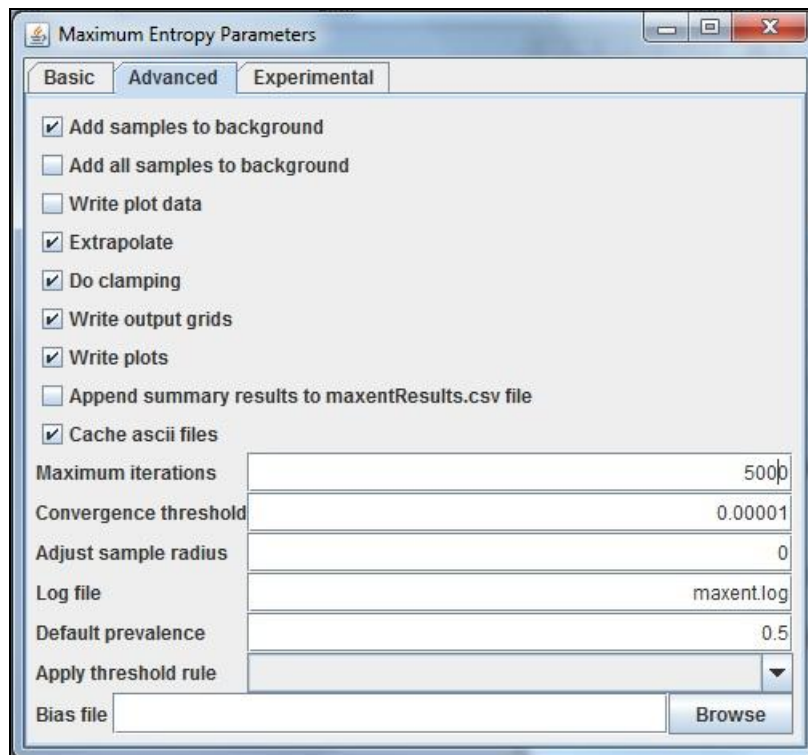


Figure 30: Advanced options in Maxent. Note the number of maximum iterations to ensure convergence.

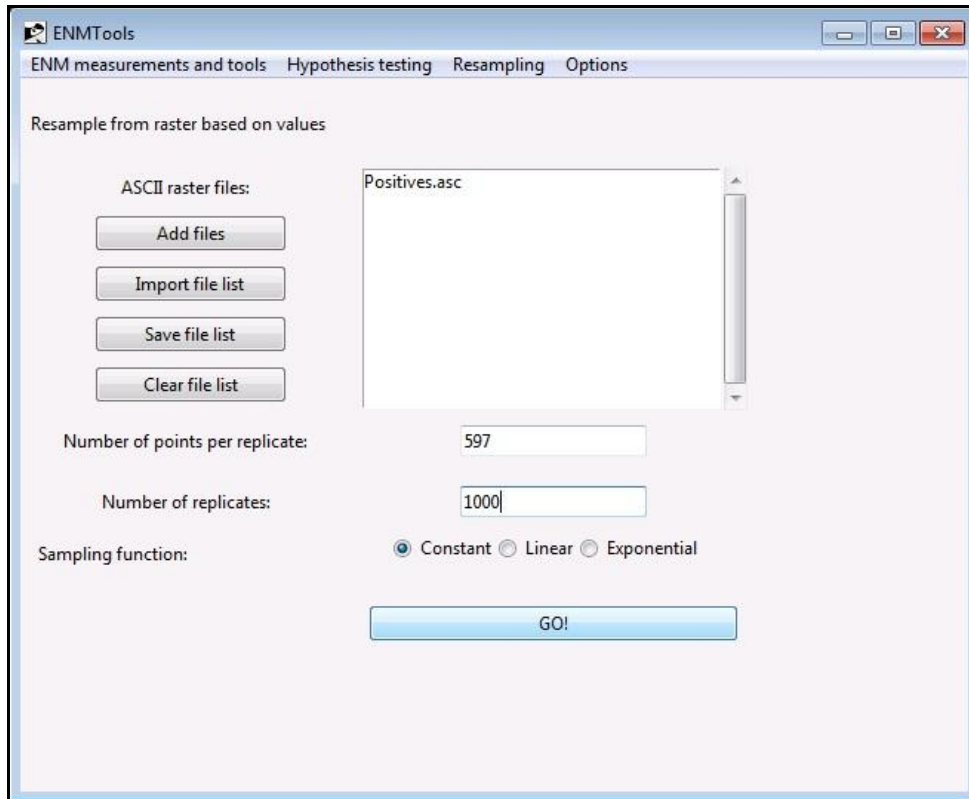


Figure 31: Null Model creation with ENMTools.

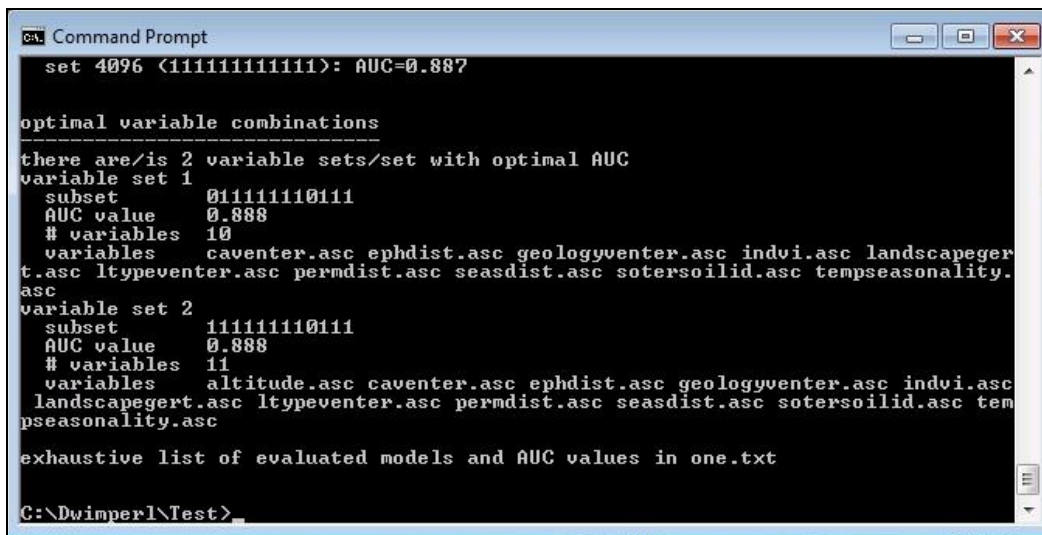


Figure 32: Maxent model surveyor output. Two variable sets with optimal AUC are displayed containing 10 and 11 variables respectively.