Modelling Habitat Suitability To Predict The Potential Distribution Of Erhard's Wall Lizard *Podarcis erhardii* On Crete

Matthias Herkt
March 2007

Course Title:	Geo-Information Science and Earth Observation for Environmental Modelling and Management
Level:	Master of Science (MSc)
Course Duration:	September 2005 - March 2007
Consortium partners:	University of Southampton (UK) Lund University (Sweden) University of Warsaw (Poland) International Institute for Geo-Information Science and Earth Observation (ITC) (The Netherlands)
GEM thesis number:	2005-15

Modelling Habitat Suitability To Predict The Potential Distribution Of Erhard's Wall Lizard *Podarcis erhardii* On Crete

by

Matthias Herkt

Thesis submitted to the International Institute for Geo-information Science and Earth Observation in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation, Specialisation: Environmental Modelling and Management.

Thesis Assessment Board

Prof. Dr. A.K. Skidmore (Chair), NRS Department, ITC, The Netherlands
Prof. P. Pilesjö (External Examiner), GIS Centre, Lunds Universitet, Sweden
Dr. A. G. Toxopeus (1st supervisor), NRS Department, ITC, The Netherlands
Dr. M. Schlerf (2nd supervisor), NRS Department, ITC, The Netherlands

ITC International Institute for Geo-Information Science and
Earth Observation, Enschede, The Netherlands

Disclaimer

This document describes work undertaken as part of a programme of study at the International Institute for Geo-information Science and Earth Observation. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the institute.

Abstract

Predictive species distribution modelling is a valuable tool for decision-makers in biodiversity conservation, invasive species monitoring and other natural resource management fields. This study employs one recently proposed modelling technique – Maxent – to investigate the curious geographic distribution pattern of Erhard's wall lizard *Podarcis erhardii* on Crete and surrounding islets. The main objective is to find out if this distribution can be explained using a set of environmental variables only. A secondary objective is to test the usefulness of an ASTER-derived land cover variable. Thirdly, the effect of replacing the single point occurrence data with representative 'natural habitat' polygons created during fieldwork in the immediate vicinity based on expert knowledge is investigated.

A set of 19 environmental predictors is employed together with 75 presence-only records, obtained from the National History Museum of Crete. Results are evaluated using the threshold-dependent True Skills Statistic (TSS), a binomial test and the threshold-independent ROC analysis with AUC. Relative variable importance is assessed based on Maxent's built-in Jacknife functionality.

Multi-annual NDVI is found to be the most important predictor, matching not only areas with high presence but also areas of apparent absence of *P. erhardii*. While the climate variables cloud cover and actual evapotranspiration rank next, the ground variables altitude and CORINE land cover also contribute significantly to the overall 'cumulative gain' of 1.86. The resulting distribution fits the provided occurrence data very well (AUC of *test* partition = 0.86) and results are highly significant at the sensitivity-specificity-equality threshold ($p < 0.001$).

Western Crete serves as subset for testing the usefulness of ASTER imagery for the purposes of this study at regional scale. The ASTER-derived land cover variable is found to contribute as much unique information to the distribution as NDVI, ranks second in individual 'cumulative gain' and increases the overall 'cumulative gain by over 20%. The replacement of single occurrence points with more representative plot data increases the 'cumulative gain' by an additional almost 20%, primarily because this allows to better exploit the discriminative power of continuous climatic variables with 1x1km resolution.

The study concludes with the observation that present environmental conditions alone may 'explain' the observed curious geographic distribution of *P. erhardii* on Crete. Furthermore, it recommends the use of ASTER imagery for similar studies, because overlay analysis reveals not only a fairly strong association between ASTER, NDVI and CORINE classes preferred by *P. erhardii*, but also a much more concise identification.

Keywords: species distribution models, presence-only, Maxent, ASTER, NDVI, herpetofauna, habitat suitability, *Podarcis erhardii*, Crete.

Acknowledgements

This thesis is an outcome of the BIOFRAG–ITC internal research project in collaboration with the National History Museum of Crete (NHMC) at the University of Crete, Heraklion, Greece and the International Institute for Geo-Information Science and Earth Observation (ITC), Enschede, The Netherlands.
I owe special thanks to Dr. Petros Lymberakis (NHMC) for sharing his expert knowledge on *Podarcis erhardii* and introducing me to Greek cuisine, to Dr. Manolis Nikolakakis (NHMC) for his dedicated assistance in data acquisition and to Professor Dr. M. Mylonas (University of Crete) for facilitating this collaborative project. Without their cooperation and willingness to share data and assist in the field this research would not have been possible.

A sincere thank you to my first supervisor Dr. Bert Toxopeus. Fieldwork with you was both very instructive and fun, including the scorpion encounter. I very much appreciated your stimulating questions and guidance throughout this thesis project. You have dared to let me work on my own for a while and yet you always ensured that I didn't lose track. Thank you!
I wish to express gratitude also to my second supervisor Dr. Martin Schlerf for his helpful comments and to the whole team at ITC who facilitated and evaluated my internet-based mid-term presentation. Special thanks also to Dr. Kees de Bie for his excellent GPS instructions and NDVI processing suggestions, and to André Kooiman for listening and support during the thesis proposal phase.
Further I would like to extent my sincere appreciation to Professor Pete Atkinson (University of Southampton), Professor Petter Pillesjö (Lunds Universitet), Professor Kasia Dąbrowska-Zielińska (Warsaw University) and Professor Andrew Skidmore (ITC) for having made the GEM MSc programme possible. It has truly been a rewarding and enjoyable time!

Außerdem ein dickes Dankeschön an meine Eltern Roswitha und Mike dafür, dass sie stets daran glauben dass ich weiß was ich tue und jederzeit mit Rat und Tat zur Seite sind. Besonders bedanken möchte ich mich auch bei Dietlind und Hermann Keller sowie Heilwig und Michel Parys für die finanzielle Unterstützung die mir dieses Studium erst ermöglicht hat.
Finally, thanks to all my fellow GEM students – it's been a really good time! – and to Jelle Harms, who has been a great fieldwork partner. Dude, I miss the dust. We should go for another fieldwork trip soon!

Table of Contents

List of Figures

List of Tables

List of Appendices

1. Introduction

1.1. Research Background

1.1.1. Context

Knowledge of potential species geographic distribution is of practical relevance to many disciplines. Epidemiologists and invasive species management require information on potential habitat suitability to focus their resources (Buckley et al., 2006, Meentemeyer et al., 2004, Peterson and Robins, 2003). Cultivation of crops under changing climatic conditions is of significant interest to the agricultural sector and associated industries (Rounsevell et al., 2006). In light of its unprecedented speed in geological-historic terms (IPCC, 2007), climate change also increases the pressure on many less common wildlife species, which are already struggling with relatively sudden habitat loss, degradation and fragmentation due to the unabated expansion of areas dominated by anthropogenic land use (Corsi et al., 2000. The challenges biodiversity conservationists face, may further illustrate the significance of such research.

When determining optimum localities for reserve establishment, obviously a wider range of factors needs to be taken into account – not just single-species habitat suitability. Instead, many of these factors are aspects characterizing the ecosystem community level, e.g. species richness and evenness indices or the principles of complementarity, persistence and vulnerability. Their consideration is essential as reflected in the vast body of literature on this topic (Magurran, 1988, Rebelo, 1994, Margules and Pressey, 2000, Lévêque and Mounolou, 2001, Sarkar and Margules, 2002, Magurran, 2005, Raffaelli, 2004, Henle et al., 2004).

Without a thorough understanding of *species-specific* habitat requirements however, sound estimates of future species-specific habitat suitability cannot be made. Reserves established on this basis may prove unable to support some of the initially targeted species (van Teeffelen et al., 2006, Moilanen and Wintle, 2006). Should a keystone species be among this group of species facing unsuitable future conditions,

the worst-case scenario is a drastic decline in the reserve's ecosystem functionality (Catterall et al., 2003). Given typically very limited financial and political resources in the field of protected area management, the consequences of such poor investment decisions may be considerable (Pressey et al., 1993). This example from the field of biodiversity conservation further illustrates the need to make models spatially explicit, as geometrical configuration of reserves is generally critical to success van Teeffelen et al., 2006). In conclusion, testing and advancing current species distribution models has a wide range of potential benefits for decision-makers in natural resource management.

1.1.2. Species Distribution Modelling

To predict species potential distribution, a range of models has been developed. While major differences exist regarding the statistical algorithms used and their species occurrence data type requirements, all models generate predictions in multidimensional ecological space. Species distribution models therefore do not predict species geographic occurrences as such, but produce a spatially explicit probability surface (sometimes as binary output only) that represents habitat suitability in ecological hyperspace after factoring in some specified constraints (sometimes including variable interactions).

According to ecological niche theory (Hutchinson, 1957), each species depends on the existence of a specific set of environmental conditions for its long-term survival. This concept refers to not only the abiotic environment but also to biotic factors of the respective ecosystem determining the abundance of resources as well as trophic chain interactions. As a consequence of such biotic interactions (competition, predation), but also of geographic barriers to dispersal and colonization as well as anthropogenic pressures, species in reality never fully occupy their fundamental niche, i.e. the ecological-geographic space which meets their requirements (Anderson et al., 2003). Instead, a species almost always occupies a subset of its fundamental niche only, which is termed the realized niche (Brown and Lomolino, 1998). The important implication of this concept for species distribution modelling is that occurrence records by definition can only be sampled from the realized niche. Predicted results therefore tend to underestimate the "potential" distribution (Phillips et al., 2006).

Predictive distribution models generally contain three components: (1) an ecological model which serves to select and prepare the environmental variables for input, (2) a

data model which defines type and collection method of occurrence records used for input, and (3) a (non)-statistical model that integrates and analyzes these data (Austin, 2002). Identification of the relevant environmental variables and quantification of their interaction terms are the main challenges in determining the ecological model. Within the data model it is essential to consider whether species occurrence information is available in presence / absence format or as presence-only data, as this distinction results in different potential biases, suitable evaluation techniques and strongly influences the choice of the (non-) statistical model (Phillips et al., 2006, Guisan and Zimmermann, 2000).

For binary response variables (presence/absence) a range of species distribution models has been developed, many of which use logistic regression approaches, most notably GLMs (McCullagh and Nelder, 1989) and GAMs (Hastie and Tibshirani, 1990).

For presence-only data – as needed in this study – a similar variety of (non)-statistical models is available, although the proportion of recently developed models seems higher (Elith et al., 2006). While for instance BIOCLIM (Busby, 1991) and DOMAIN (Carpenter et al., 1993) as environmental envelope techniques work exclusively with presence-only data, most other models employ background samples. By using 'pseudo-absences' some methods, e.g. GAMs and GLMs, which were initially built to work with presence-absence data, have successfully been modified to work with presence-only data (Elith et al., 2006). GARP (Stockwell and Peters, 1999), which combines rule-based and iterative random elements to infer species distribution, has been a reference model for some time. Recently however a large comparative study (Elith et al., 2006) has produced further evidence that Maxent (Phillips et al., 2004) may generally outperform GARP, especially when sample size is small (Hernandez et al., 2006), but not necessarily at local scales (García Márquez, 2006). Another promising adequate model for the purpose of this study recently developed by the machine learning community is BRT (Friedman et al., 2000), while ENFA (Hirzel et al., 2002) has received increasing attention for its application of principal component and marginality aspects into the conventional ecological niche concept. Community-based models are another area of active research with promising results for both new approaches such as GDM (Ferrier et al., 2002) which focuses on compositional dissimilarity, and for modifications of established approaches such as MARS-COMM (Friedman, 1991). These models use presence information of other species as surrogates to enhance predictions for the target species (Elith et al., 2006). Major differences between models exist regarding integration of the response variable (species occurrence), the selection, weighting

and fitting of individual predictors as well as provision for interactions among variables and output format of predictions (Elith et al., 2006). Some of these differences as well as potential biases related to presence-only data will be discussed in more detail in later chapters when describing Maxent and fieldwork design.

There are at least three reasons why research testing models for presence-only data is very valuable: Firstly, a large number of geo-referenced presence-only data resides with herbaria and museums and becomes increasingly accessible via the internet. Secondly, for most faunal studies species information is limited to presence-only records, because of the spatial and temporal mobility which characterizes wildlife biology. Full enumeration sampling techniques are virtually impossible in this case, as a species may be absent only temporarily ('false absences') from the site due to e.g. weather conditions, population dynamics and trespassing humans (Hirzel et al., 2006). Note that the degree of certainty attached to this 'absence' claim depends mainly on the mobility of individuals of the species, on how abundant / detectable the species is locally (Kéry, 2002), and on the survey design (Mackenzie and Royle, 2005). Thirdly, as a consequence of this lack of complimentary absence information for modelling, less mature evaluation techniques are available for presence-only models (Fielding and Bell, 1997). This thesis intends to contribute to this area of research by testing the performance of a newly developed distribution model (Maxent) on a presence-only dataset.

1.1.3. Study Area

The island of Crete is located in the Eastern Mediterranean sea and belongs politically to Greece since 1913. Stretching about 245 km from West to East (slightly North of the 35° N parallel) and between 12 and 56 km from North to South, it covers an area of approximately 8,336 km^2, thus representing the fifth largest island of the Mediterranean. An overview map of the study area is provided in chapter 1.2.
A high mountain range forms the geological 'backbone' of Crete and spans the island from West to East reaching peak altitudes of 2456 m (Psiloritis) and 2452 m (Lefka Ori). The four main mountain massifs are separated by undulating lowlands; the Mesará in the South constitutes the island's major fertile plain. Four large peninsulas structure the North coast of Crete. There are numerous offshore islands, some of them only a few hundred square meters in size.
Geologically, much of Crete's terrain is characterized by karst formations whose limestone and dolomite rocks have also allowed for the carving of deep gorges and underground drainage systems. As a result, large rivers are rare and most streams fall

dry during summer. Fertile soils on Crete are mostly rendzinas on calcareous parent material. In higher altitude areas, this material decomposes mostly into red clays, and yellow clays below 1000m (Fielding and Turland, 2005).

Crete enjoys a Mediterranean climate. Summer from June to August is hot and dry except for thunderstorm rainfalls in the mountains induced by northerly winds from the Aegean. Winter ranges from November to March and receives most of the annual rainfall due to moist, westerly winds from the Atlantic. Air temperatures however remain below those of the sea and snow cover is common above 1600 m during this period (Fielding and Turland, 2005). Geographic variations in rainfall and temperature will be discussed in more detail further on.

Species richness on Crete is high for both flora and fauna, as is the degree of endemism. This is due to both it's recent history as an island in isolation for five million years as well as to Crete's equidistance to three ecological regions of continental dimension – Asia (Turkey), Africa (Lybia) and Europe (Greece) – to which it has been linked by land bridges temporally from time to time. Vegetation composition however has been heavily influenced by human activities over the last 5000 years, in particular by widespread deforestation in the 17th century aggravated by an increasingly drier climate (Fielding and Turland, 2005, Legakis and Krypriotakis, 1994).

Home to Europe's oldest urban civilization – the Minoan culture – until its earthquake-induced collapse in 1450, Crete is now home to about 600,000 inhabitants and the destination for 2 million tourists every year. About 55% of the population lives in urban, the considerable rest in rural areas. The agricultural sector, including grazing cattle, has only recently been replaced by the service industry (mainly tourism) as the basis of the Cretan economy. Wine, raisins, olives, fresh fruits, horticultural products like tomatoes and eggplants, as well as honey and herbal pharmaceuticals remain important exports to – mostly – mainland Greece (Wikipedia contributors, 2007).

1.1.4. Target Species

Podarcis erhardii (Squamata: Lacertidae) is a small lizard with about 7cm in body length and a tail twice as long. The skin is smooth and ranges in colour between grey and brown, occasionally green. The dark side lines covering its back are always thicker than the dorsal line if it exists at all. Island populations however tend to have spotted patterns on the back instead of lines. Occasionally *P. erhardii* displays a row of blue spots in the lower abdomen area which is otherwise bright in colour. It's diet consists mainly of insects and other arthropods. When threatened, the species usually

runs and hides in small cracks, holes and shrub (Engelmann, 1986, Arnold, 2002). As it's name suggests, the species is a good climber, but it is not a good swimmer and unable to cross even narrow stretches of water (Poulakakis et al., 2005).

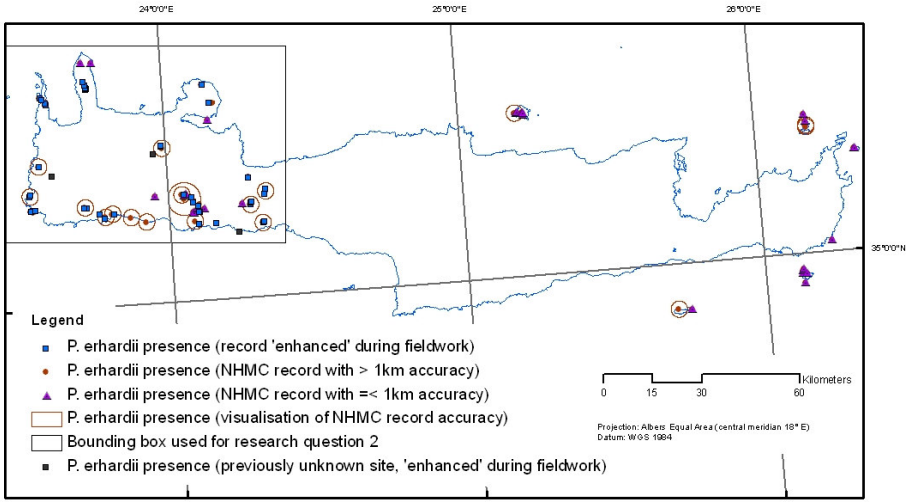
[Photos: Jelle Harms, Crete 2006]

Figure 1 Target Species *Podarcis erhardii*

P. erhardii populations are relatively common throughout the Balkan region extending South into mainland Greece, the Eastern rim of the Peloponnesus and many of the Aegean islands west of the Mid Aegean trench including Crete (Poulakakis et al., 2005). Although there is a certain degree of taxonomic uncertainty at subspecies level (Poulakakis et al., 2005), the Crete population of *P. erhardii* seems to have lived in isolation for about 5.2 million years (Poulakakis et al., 2005). Coastlines with rocky, sandy or pebble shores, rocky areas and Mediterranean shrub lands represent suitable habitat for *P. erhardii*, as well as garden and urban environments (Poulakakis et al., 2005). The adaptive capability of *P. erhardii* may be considerable, as island populations have been found to inhabit also open spaces like sand dunes. Furthermore, perhaps in response to greater spatial and seasonal clustering of food availability on less hospitable islands, these *P. erhardii* populations also display faster digestive abilities (Pafilis et al., 2007).

1.2. Problem Statement and Justification

Apart from the above biogeography studies based on phylogenetics, there appears to be a lack of literature on what environmental variables constitute the main gradients specifying the niche requirements of *P. erhardii* and to what extent they determine the species current geographic distribution on Crete. This problem is all the more intriguing as the observations of *P. erhardii* are not uniformly distributed geographically across the island.



[Source: NHMC, 2006 and own fieldwork].

Figure 2 P. erhardii presence records across Crete

While some islands and islets surrounding Crete are inhabited by *P. erhardii*, others are not – despite the occasionally very small distance of less than 100 m. Even stranger, while occurring on near-coastal islets of Eastern Crete, *P. erhardii* has not been observed at all on the Eastern and Central parts of mainland Crete. Note that significant sampling of herpetofauna has taken place all over Central and Eastern Crete, which may not have been *focusing* on *P. erhardii* but *could* have still yielded occasional observations of *P. erhardii*. In the Western third of Crete however the species is abundant between sea level and 2000 m altitude. It has been observed on coastal, open, and rocky areas as well as scrublands. This geographic pattern provokes the question whether *P. erhardii* can be considered a generalist rather than a specialist regarding habitat requirements. It also raises the question of whether there is a particular preferential or prohibitive (set of) environmental variable(s) that affect *P. erhardii* in Western Crete, and the remaining parts of mainland Crete respectively.

Owing to the advances in modelling techniques described earlier and the provision of access to *P. erhardii* occurrence records stored at the National History Museum of Crete (NHMC), this knowledge gap can be – at least partially – addressed within this thesis.

From a biogeography perspective, the significance of such research lies in a better understanding of whether *P. erhardii* populations on Crete and the near-coastal islets

are in a stage of expansion or retreat. If no evidence emerges for a hostile environment that likely prevented *P. erhardii* from (recent) colonization of mainland of Central and Eastern Crete, it may prove worthwhile to investigate historical factors more closely, such as island fragmentation due to tectonic changes, the presence of parasites and certain early anthropogenic land use practises.

From a more utilitarian perspective, the justification for this research lies in the demonstration as to what extent environmental data of different scale and detail – including high-resolution remote sensing data – together with a suite of GIS tools can be employed to identify and extrapolate species-specific habitat suitability. Both more evidence for the potential of remote sensing data in fauna species distribution modelling (De Leeuw et al., 2002), so as well as further insights into issues of adequate scale for predictor variables (García Márquez, 2006, Murwira et al., 2003) may result from this approach. Should it produce promising and transferable results, it may represent a more cost-effective way of identifying and monitoring areas of high importance for a given target species. Researchers and managers from a variety of related disciplines may benefit from this knowledge, e.g. epidemiologists, invasive-species managers and reserve planners for biodiversity conservation (see also chapter 1.1.1).

1.3. Research Objectives

The general objective of this research is to investigate to what extent the currently observed geographical distribution of *P. erhardii* on Crete and its near-shore islands, can be explained by (current) environmental conditions. More specifically, this research sets out to

1a) create a potential species distribution map with associated information about its prediction strength.

1b) identify the set of most significant environmental predictors.

2a) evaluate the usefulness of an ASTER-derived land cover map as an environmental predictor.

2b) assess the sensitivity of the distribution model created in (2a) to changes in the way species occurrence data are expressed.

3) conduct a preliminary investigation into land and vegetation cover types preferred by *P. erhardii*.

1.4. Research Questions

1a) Can a probability distribution for *P. erhardii* be modelled across Crete that 'explains' the species' curious geographic occurrence pattern, using Maxent and a given set of high accuracy species observation points derived from both fieldwork and the NHMC database?

1b) Which are the most relevant environmental predictors for this distribution model and what is the relative loss in prediction strength when only the most important one-third are used for modelling?

2a) Can the probability distribution in Western Crete be improved by adding an ASTER-based land cover classification to the suite of predictor variables?

2b) How does the probability distribution model created for Western Crete (which includes the ASTER-derived variable) change, if the initial species observation 'points' are replaced by a set of points representing a wider area presumed to feature "natural habitat conditions" as assessed in-situ during fieldwork?

3) What are the land cover and vegetation conditions preferred most by *P. erhardii*?

1.5. Research Hypotheses

1a - H_1: A probability distribution for *P. erhardii* across Crete can be modelled that differs significantly from random and features a regularized training gain higher than 1.5 (using the Maxent algorithm and predictor data with a resolution of 1km or finer).

1a - H_0: A probability distribution for *P. erhardii* across Crete that differs significantly from random and features a regularized training gain higher than 1.5 (using the Maxent algorithm and predictor data with a resolution of 1km or finer) can <u>not</u> be modelled.

1b – H_1: Given the pattern of documented *P. erhardii* observation localities on Crete and the scale of the study area, the most important environmental predictor identified in (1a) – measured as individual training gain using Jacknife – is a variable representing a ground feature.

1b – H_0: Given the pattern of documented *P. erhardii* observation localities on Crete and the scale of the study area, the most important environmental predictor identified in (1a) – measured as individual training gain using Jacknife – is not a variable representing a ground feature.

2a – H_1: The probability distribution in Western Crete can be improved in terms of AUC by including an ASTER-based land cover classification variable.

2a – H_0: The probability distribution in Western Crete can not be improved in terms of AUC by including an ASTER-based land cover classification variable.

2b – H_1: The training gain of the probability distribution modelled for Western Crete (including the ASTER-derived variable) increases, if the initial species observation 'points' are replaced by a set of points representing a wider area presumed to feature "natural habitat conditions" as assessed in-situ during fieldwork.

2b – H_0: The training gain of the probability distribution modelled for Western Crete (including the ASTER-derived variable) does not increase, if the initial species observation 'points' are replaced by a set of points representing a wider area presumed to feature "natural habitat conditions" as assessed in-situ during fieldwork.

3 – H_A: *P. erhardii* presence records coincide significantly with 'open areas', i.e. the most frequent NDVI class among presence sites belongs to the lowest third of all NDVI classes generated for Crete (using an ISODATA classification).

3 – H_0: *P. erhardii* presence records do not significantly coincide with 'open areas', i.e. the most frequent NDVI class among presence sites does not belong to the lowest third of all NDVI classes generated for Crete (using an ISODATA classification).

1.6. Research Approach

a) Pre-process remotely sensed environmental variables where necessary, e.g. NDVI and ASTER-based land cover. For the latter this includes the full sequence of orthorectification, relative correction between granules, mosaicing, cloud removal and classification.

b) Prepare data for all environmental variables to fit study area regarding extent and spatial resolution, apply identical projection and convert to ascii format for Maxent.

c) Conduct fieldwork to enhance provided species data by determining representative "natural habitat polygons" (mapping units) in immediate vicinity of the species observation point as listed in the NHMC databank. Complement species observation data by own sightings made during fieldwork (optional, for analysis outside this thesis).

d) Prepare visited species observation sites as training data for subsequent modelling by converting it into target projection, rasterizing both the full area and the centre point of each mapping unit. Finally, project, rasterize and integrate additional species observation points provided by NHMC.

e) With Maxent software, model potential species distribution across Crete using all environmental variables (except ASTER-based land cover) and species occurrence data represented by (1) the centre points of recorded "natural habitat polygons", (2) the above plus additional NHMC records with accuracy of better than 500m radius and who have not been visited ('replaced') during fieldwork.

f) Analyse (2) and use (1) to compare sensitivity of individual variables to different extent; run (2) again using only the most important variables; overlay results with remaining (less accurate) NHMC records for visual analysis; respond to hypotheses and research questions (1a) and (1b).

g) Evaluate results by partitioning occurrence data and use of (1) threshold-dependent binomial z / t statistic, (2) TSS and (3) threshold-independent ROC/AUC.

f) Include the ASTER-based land cover variable as model input and run Maxent again using all variables, but for comparison reasons with species data as in (e1).

Limit projection to Western Crete. Compare (e1) and use this run's AUC values to address hypothesis and research question (2a).

g) Randomly select 10 points inside each mapping unit created during fieldwork (minimum distance 5 m); run Maxent again using these 'plots' instead of previous mapping unit central points together with all predictors as in (f). Compare resulting AUC value with the one created in (f). Address hypothesis and research question (2b).

h) Explore frequency statistics of original predictor files, sample values at 'enhanced' mapping unit central point locations plus NHMC presence records if accuracy better than 1 km; overlay NDVI and ASTER predictors with CORINE variable and extract associated names for land / vegetation cover variables. Consider the relative class frequency of total background when interpreting count-based predictor class preferences of *P. erhardii*. Acknowledge limitations of this descriptive analysis and address hypothesis and research question (3).

1.7. Assumptions

It is assumed that the available species presence data – albeit 'static' in character – reflect the equilibrium state of the long-term species-environment relationship (Hirzel and Guisan, 2002), i.e. that the sample is derived from source not sink populations. As to environmental predictor data, the assumption is made hat present conditions represent past conditions to the extent that changes in past conditions did not surpass the adaptive capability of *P. erhardii*.

Looking at the geographical distribution of confirmed *P. erhardii* observations on Crete and near-coast islets alone, 'distance to shoreline' may turn out – if tested – to be a significant environmental predictor. However, as there are confirmed species observations throughout the Balkans and mainland Greece in locations at least 50 km away from the coast (Poulakakis et al., 2005), this study assumes that 'distance to shoreline' does not matter on Crete either. Several assumptions are made with respect to the acquired ASTER imagery. They are outlined in the respective Data Preparation chapter.

1.8. Limitations

Due to time and budget limitations, fieldwork was restricted to visiting *P. erhardii* occurrence sites located West of Heraklion. In-situ data related to habitat suitability

could therefore not be generated for Eastern and much of Central Crete. Consequently, not all species occurrence records provided in the NHMC database could be enhanced (as described in chapter 2.1.4). Thus, when species occurrence records were used to model a distribution across *all* of Crete, these data were not identical in terms of positional accuracy and descriptive quality. This inconsistency however is unlikely to have caused a significant degradation of species occurrence records, because for the purpose of this study, positional accuracy was only adjusted to better coincide with presumed *natural* habitat conditions (as opposed to a village centre for instance).

Unfortunately, ASTER imagery could only be compiled for Western Crete. Although relative atmospheric correction was successfully performed granules covering Western Crete, it proved impossible to mosaic and append granules covering central and Eastern Crete due to the presence of band-specific haze and a particular narrow image overlap in central Crete.

Finally, generated probability distributions display some NoData areas – mostly along the coast and the South-Eastern mainland including offshore islands. This is due to the varying spatial extent and generalization of the environmental variables used (and in the case of ASTER also a consequence of cloud cover).

2. Methods and Materials

2.1. Species Observation Data

2.1.1. Provided NHMC Data

Species occurrence data were obtained from the National History Museum of Crete (NHMC, 2006) as part of a collaborate research project between the NHMC at the University of Crete, Greece, and ITC, The Netherlands. The data were used both as reference for partial 'enhancement' procedures during fieldwork as well as direct input into the species distribution model.
To ensure temporal correspondence with the suite of environmental predictors, topicality of the NHMC data was checked and considered sufficient, as *P. erhardii*

observations dated back on average to 1999 with no observation older than 1990. Data were provided as presence-only records in EGSA projected coordinate system as well as geographical Latitude / Longitude. As many occurrence sites had been revisited several times, only one (the most recent) *P. erhardii* observation record for each identical XY locality was extracted from the database and included in this project. After removing all presence-only records with an associated accuracy of lower than 1km, a total of 51 (out of 69) were used for further analysis. Eventually, after fieldwork allowed for the 'replacement' of most locations in Western Crete (see next chapter), a total of 29 species observation records from this database remained for *direct* input into the distribution model.

2.1.2. Fieldwork Objectives

The principal objective of fieldwork was to enhance provided species observation data by recording areas of representative natural habitat for each provided point location. The replacement of general location information (e.g. XY coinciding with a road crossing or village centre) was expected to improve results when modelling species-environment relationships. While this information may or may not improve the outlook for research objective 1, it was a requirement for research objective 2: the assessment of the usefulness of an ASTER-based land cover classification map with 15m resolution as environmental predictor.
A secondary objective of fieldwork was to record in-situ macrohabitat information as auxiliary data to support (a) the ASTER-based land-cover classification and (b) to develop a better understanding of potential ecological surface cover variables influencing habitat suitability for *Podarcis erhardii*. The data collected represents an opportunity to apply vegetation-based cluster analysis and ordination techniques (as described e.g. in Jongman et al., 2005) to infer a new set of proxy variables to further improve the habitat suitability model for *P. erhardii*. An analysis of these records beyond descriptive statistics however is not intended within this thesis.

2.1.3. Fieldwork Design and Limitations

To achieve both fieldwork objectives above, fieldwork design had to address at least five challenges associated with the provided species observation data: extent, sampling strategy, associated biases, sample size and seasonal timing.

Firstly, due to time and budget limitations, the Eastern part of the island could not be covered during fieldwork, although that region contains confirmed observations of *P.*

erhardii on several offshore islets. The limited geographical extent of fieldwork meant that for subsequent modelling data, 'enhanced' and 'non-enhanced' occurrence data had to be combined (see next chapter for details).

Secondly, it was not possible to set up a sampling strategy for the specific purpose of this study prior to sampling, because the existing NHMC database was used as species occurrence dataset – and fieldwork 'enhancement' was intentionally restricted to marginally modifying *each location*. Although 3 new *P. erhardii* observations were made during fieldwork and 9 occurrence sites replaced by two sites in the vicinity (which further increased 'site selection bias', see below), this did not fundamentally change the predetermined regional stratification of the NHMC dataset. As in this case any random sampling among NHMC data would suffer from a bias whose magnitude would be unknown (Hirzel et al., 2002), the solution was an (almost) full enumeration of provided occurrence sites within the predefined extent, i.e. Western Crete.

Thirdly, despite refraining from sampling within the provided observation records, any *existing* biases present in the original NHMC dataset obviously *continued* to exist. Researchers who collected the NHMC data in the first place, almost certainly introduced site selection biases as a function of e.g. proximity to roads and different sampling methods due to different fieldwork purposes (Anderson et al., 2003). Spatial autocorrelation in areas of intense sampling (Segurado et al., 2006), may also exist, e.g. South of Levka Ori. Although the lack of *P. erhardii* observations in mainland Eastern Crete may in fact ***not*** be a consequence of low sampling effort, because NHMC and partners have carried out substantial fieldwork in this area yielding a multitude of other species observation (Lymberakis, 2006), considerable site selection bias is likely to exist and will have to be accounted for when interpreting predicted distribution. Computing density for all presence records across Crete (e.g. on a 100 km^2 grid) and subsequently randomly selecting an equal number from each cell could have perhaps reduced spatial autocorrelation, but would also have created potential bias in ecological space. The other two main types of bias apart from site selection – poor recording techniques and measurement flaws (Hirzel and Guisan, 2002) – were even more beyond the control of this study and *assumed* to be negligible if existent.

Fourthly, 'sample size', i.e. the number of species occurrence records available for modelling, was essentially predetermined by the NHMC database. This can constitute a major problem if (a) the data come from heterogeneous sources with varying degrees of compatibility (Margules and Austin, 1994) or when (b) the number of observations is relatively low with respect to the study area covered

(Jaberg and Guisan, 2001). As the data provided by NHMC were subject to an expert standardization upon integration into the database, the assumption is made that it meets the first criterion 'homogeneity of source'. The total number of unique observation sites however (47 located in Western Crete, of which 44 were 'enhanced' during fieldwork plus 3 new observation sites; plus another 25 located on the islets of Central and Eastern Crete) is relatively low compared to the large study area. On the other hand, a study by Stockwell and Peterson, 2002, found that 50 data points can be sufficient to produce near-maximal occurrence predictions employing coarse surrogates and a machine-learning model (note: GARP not Maxent). The number of species records in this study is hence likely to be just about large enough for modelling all of Crete; for modelling the subset of Western Crete alone however (46 'enhanced' points), any validation based on splitting these records into training and test data, must be interpreted with utmost care.

Finally, as fieldwork was conducted from October 2nd – 21st, 2006, conditions for recording habitat information and sighting *P. erhardii* in-situ were impacted by some bad weather. 'Natural habitat polygons' however, could still be identified with sufficient confidence. Hence, as little of the macrohabitat information recorded is being used in this thesis, an introduction of bias due to the timing of fieldwork is unlikely.

2.1.4. Fieldwork Procedure

To meet the primary fieldwork objective, the provided NHMC presence records were 'enhanced' (and possibly improved in terms of accuracy as well) by creating one or more 'natural habitat polygons' in the immediate vicinity of each visited presence record, representing the environmental conditions assumed to have ensured the species long-term survival at this location.

Some preparation of auxiliary data was required however prior to these recordings. As not all species observation records in the NHMC database referred to an exact location (XY point) but contained the accuracy information for each locality as an extra field, a corresponding buffer zone was created around each point, thereby indicating the approximate area within which the species was observed. This accuracy information was provided categorically with classes defined as "20-100 m", "101-300 m", "301-1000 m", "1-5 km" and "over 5 km". The buffer zones were given the maximum value of each category; the latter one a proxy value of 10,000m.

Next, each species observation locality in the NHMC databank was visually inspected and recoded in the field, following the procedure below:

- a compact polygon of about 100m^2 was created around the exact XY point unless the location was dominated (or tightly encroached) by man-made infrastructure (grazing accepted). A minimum distance of about 10 m was maintained between polygon borders and any non-natural features such as a road. If the exact location coincided with a road or turned out to be inaccessible, the polygon was created as close to the exact locality as possible.
- considering the point-specific accuracy of the observation, sometimes additional polygons were created for points with an accuracy of less than 100 m in order to capture the full range of apparently or potentially suitable habitat. A range limit of about 200 m away from the exact locality was established however, which equalled the average visibility and walking distance considered feasible given the time and budget constraints of the fieldwork.
- an additional polygon was created within this range if (a) a very heterogeneous surface cover was observed which the observation record obviously referred to, e.g. a patchwork of high shrubs, rocky outcrops, grassland and tall trees. In this case a polygon was created for each main natural surface cover type based on visual inspection. A unit was delineated around a distinct area where internal heterogeneity of the potential unit was visually smaller than the heterogeneity observed between potential units.
- an additional polygon was created outside this range – but within the buffer zone – if (b) it was not feasible to access the 200m range. In this case a polygon was placed as close as possible to the range zone, as soon as a fairly homogenous natural surface cover was found and deemed akin to the one of the exact locality.

The above procedure of placing polygons – also coined "mapping units" – around or near species observation localities was primarily based on visual analysis of surface cover. However, both a CORINE land cover map and ASTER imagery were displayed on the iPAQ handheld and used to assist in this process. The ASTER imagery used for this purpose was a simplified land cover classification map, created by first running principal component analysis (PCA) on the original image (orthorectified, all VNIR and SWIR bands) and then an ISODATA classification algorithm on the PCA image using the top five "bands" as classes (Jensen, 2005). It was displayed as two separate RGB images (4-3-2 and 3-2-1) and owing to its large-scale resolution it primarily assisted in ensuring homogenous mapping units. The CORINE layer primarily helped to find out if there was a major second or third natural land cover type present in the 300m range, which would justify the creation of an additional polygon for this particular species observation point.

For each of the mapping units identified, a number of macrohabitat parameters were recorded. Most of these described the area contained in the mapping unit, some characterized the overall landscape, while a third group contained species-specific ecological observations. Additionally, metadata such as weather conditions, time of the day, minutes spent searching for *P. erhardii* etc was recorded. It was beyond the scope of this study however to prepare these data using canonical ordination techniques (Jongman et al., 2005) and to employ them as additional predictors for modelling *P. erhardii* habitat suitability.

2.1.5. Pre-processing of Species Observation Data

Two different species observation data sets were prepared as input for modelling: The first one (46 points) contained only the central points of the mapping units created during fieldwork (F = fieldwork). The second one (+29 points) contained in addition to these 46 points, all non-visited observation points as listed in the NHMC databank, which had an accuracy of 1000m or better and had not been visited (and 'replaced') during fieldwork (FM = fieldwork + NHMC). Although it could not be verified during fieldwork whether the selected non-visited points in fact coincide with "representative natural habitat", the sheer geographic location of these points – remote mountain areas and uninhabited islets – was considered sufficient evidence to make this *assumption*. Thus, in the second input data set, all visited and selected non-visited observation sites were combined. The main reason for creating this second input dataset is that it includes 17 out of 21 of the observation points located in the non-visited Eastern part of Crete. It was expected that this would improve the fit of the probability distribution to known presence sites on Crete, as the few remaining sites not included in this 'sample' (due to their poor accuracy) seem to not be located at the extremity of any predictor and are geographically quite evenly distributed. Both observation datasets were prepared in ArcMap. XY data were extracted *after* conversion into the target projection, and eventually converted into .csv format (together with the species field) for input into Maxent modelling software.

2.2. Environmental Predictors

2.2.1. Selection Criteria

Environmental predictors were selected in light of the ecological processes assumed to influence *P. erhardii*, subject to data availability and the objective of this study. This represents a deductive element in this thesis' primarily inductive research

approach (for a related theoretical discussion see Corsi et al., 2000 and De Leeuw et al., 2002). Both direct (e.g. rainfall, temperature) and indirect variables (e.g. altitude, cloud cover) were incorporated along with a set of proximal resource variables (land cover standing for shelter and food availability). Isothermality and temperature seasonality are included as surrogates for climatic stress tolerance (Austin and Smith, 1989). Variables representing anthropogenic disturbance factors (e.g. distance to settlements) were not considered given the species literal preference for man-made walls. As neither the exact response curves of these variables nor the amount of (likely) interdependencies are known however, no attempt was made to establish a more solid ecological model prior to statistical modelling. Although this general approach is common in species distribution modelling (Austin, 2007) it is clear that outcomes would gain interpretability if predictor selection was based on a more comprehensive and explicit theoretical framework. The ASTER-derived land cover variable was selected as input primarily to investigate the potential usefulness of high-resolution remote sensing data in this context.

#	variable	units	orig. spat. resolution	dated	source
1	altitude	[m]	3 arc seconds (~90m)	2000	USGS / SRTM
2	aspect	[degrees]	3 arc seconds (~90m)	2000	USGS / SRTM
3	slope	[degrees]	3 arc seconds (~90m)	2000	USGS / SRTM
4	geology	[nominal]	n/a	n/a	NHMC
5	soil type WU	[nominal]	1:1,000,000 (1km)	1986	Wageningen University
6	soil type WRB full	[nominal]	1:1,000,000 (1km)	2004	ESBN ** (vector)
7	dominant parent material	[nominal]	1:1,000,000 (1km)	2004	ESBN ** (vector)
8	depth to rock	[ordinal]	1:1,000,000 (1km)	2004	ESBN ** (vector and raster)
9	volume of stones	[ordinal]	1:1,000,000 (1km)	2004	ESBN ** (vector and raster)
10	NDVI	[nominal]	1 km	2005	CNES / SpotImage
11	dominant land use	[nominal]	1:1,000,000 (1km)	2004	ESBN ** (vector)
12	land cover CORINE	[nominal]	1:100,000 (~300m) *	2000	EEA
13	cloudcover	[%]	0.5 degrees (~50km)	1996	USGS / NIEHS
14	potential evapotranspiration	[mm]	0.5 degrees (~50km)	1996	USGS / NIEHS
15	actual evapotranspiration	[mm]	0.5 degrees (~50km)	1996	USGS / NIEHS
16	annual precipitation	[mm]	30 arc seconds (~1km)	2005	Worldclim / Hijmans et al.
17	mean annual temperature	[°C] *10	30 arc seconds (~1km)	2005	Worldclim / Hijmans et al.
18	min temp of coldest month	[°C] *10	30 arc seconds (~1km)	2005	Worldclim / Hijmans et al.
19	temperature seasonality	[continous]	30 arc seconds (~1km)	2005	Worldclim / Hijmans et al.
20	isothermality	[continous]	30 arc seconds (~1km)	2005	Worldclim / Hijmans et al.
21	land cover ASTER	[nominal]	~ 15 m	2002-06	NASA

** explicit reference is made to ESBN as the owner of this dataset (see reference section for full details); I acknowledge that this data has been made available for research purposes only, and that no commercial activities or passing on of the data to third parties is allowed.

Table 1 Environmental variables tested for significance in this study

Only simple multiple regression was performed (on the continuous variables), which resulted in the removal of the predictor 'minimum temperature of coldest month', as it showed not only perfect correlation with 'mean annual temperature' (see Appendix D) but also a very low added 'cumulative gain' in preliminary Maxent runs (based on Jacknife). Also, despite its relatively strong predictive contribution,

the variable 'soil type WU' was excluded from subsequent analysis, because it theoretically duplicated information contained in the predictor 'Soil Type WRB Full' and its classes were more general than the alternative. The remaining 18 classes are referred to hereafter as "all predictors". Although multi-collinearity tests using e.g. the variance inflation factor (VIF) method are available for a more advanced correlation analysis (Jongman et al., 2005, Brauner and Shacham, 1998), this study abstains from further variable elimination, as correlations are frequent in ecology and essential information might thus accidentally be discarded in the process (Burnam and Anderson, 1998).

All data were either downloaded via the internet from the respective provider or obtained from ITC or NHMC. Details for each source are listed in chapter 'References'.

2.2.2. Spatiotemporal Framework

To ensure positional accuracy and attribute integrity of all environmental predictors intended as modelling input, the projected coordinate system, geographic extent (two versions: all of Crete and only Western Crete) and spatial resolution of all variables were set to coincide prior to converting all predictor data from raster into ascii format.

As target projection coordinate system, a modified Albers Equal Area Projection based on the WGS 84 ellipsoid was chosen, because preserving area characteristics was deemed the most critical aspect given the large size of the study area. To ensure compatibility with other projects conducted in the Mediterranean region by ITC and partner institutes, the central meridian was set to 18° E. The other parameters were specified as follows: linear unit: meters; false easting of 4,000,000 m; false northing: 0 m; standard parallel 1: 30; standard parallel 2: 50; latitude of origin: 0; datum: WGS 84.

Spatial resolution was set to 15 m, in order to maintain maximum information content of the ASTER-based land cover variable. Resampling was performed using Nearest Neighbour in order to avoid interpolating attribute values and introducing false precision.

To minimize computing requirements, all environmental data were clipped along the coastline of the study area, assigning NoData values to sea areas. The SRTM DEM provided the reference coastline, as it had been subject to expert coastline editing prior to being made available to the public.

Temporal correspondence between species occurrence records and selected

environmental data is essential (Anderson and Martinez-Meyer, 2004). With occurrence data collection dating back to 1999 on average, the selected environmental predictors were considered to meet this requirement in general, as older data represented more permanent features like geology and – arguably! – climate, although NDVI (average 2002) and high-resolution ASTER imagery (average 2004) should be interpreted with caution in this respect.
The following chapters describe each environmental data set in more detail and provide a summary of any variable-specific preparatory steps undertaken.

2.2.3. Pre-Processing of Topography, Soil and Land Use Variables

The environmental variable "altitude" was acquired via ITC from U.S.G.S. in SRTM "Finished" Format at a spatial resolution of 3 arc seconds (~90 m). The data were recorded in February 2000 during the Shuttle Radar Topography Mission. Note that the 90m data set was generated from the original data with 1 arc second resolution, which has been made only available so far for U.S. territory. The data were "finished" prior to distribution by delineating and flattening water bodies, correcting coastlines and removing extreme values representing likely errors. The resulting DEM was provided in geographic coordinates referenced to the EGM96 geoid. USGS states a horizontal and vertical accuracy of 20m (circular) and 16m (linear) respectively, at 90% confidence level (U.S.G.S., 2006).

To meet the objectives of this study, the following pre-processing steps were performed on the altitude data: 1) clipped out Crete and surrounding islets as study area, 2) resampled subset to 0,000138° (~15m) resolution using NN, 3) projected data into target projection, 4) resampled data to final 15m resolution using NN, 5) extracted only land (and lake) areas from study area as specified earlier, 6) set all remaining values < 0 to zero, 7) converted all values into integer format to reduce model run time and 8) converted file into ascii format given the data input format requirements of Maxent.
The sequence of these pre-processing steps merits a short justification. Although resampling the data prior to projecting it meant longer processing time, it also limited planar cell shifts to a maximum of ~10 m diagonally. If projecting had been done prior to resampling, the rotation and stretching (i.e. projecting) of each cell would have caused a shift in position of up to half the cell size (in up-down or / and right-left direction). This would have equalled up to ~ 45 meters for this data set. Given that aspect and slope information were derived from this data set – and "natural habitat polygons" were recorded with an average size of 30x30 m –

maintaining maximum positional accuracy was more important than processing time. Setting negative values to zero cells delineated as land, was done assuming (with confidence) that no significant sub-sea level areas exist on Crete.

The aspect variable was generated from the (clipped) DEM above. To avoid incorrect flat areas however, the respective tool box function in ArcMap was employed prior to resampling, projecting, land area extraction and ascii conversion (steps that were subsequently performed in this sequence).

Calculation of slope required a different pre-processing approach, because unlike aspect, slope could not be calculated prior to projection as meter units were required in all dimensions. Resampling to 15m with standard NN technique, prior to slope calculation however, would result in a "mesh" of large flat areas. To solve this accuracy dilemma, it was decided to project the data with Cubic Convolution (CC) while maintaining the initial resolution (~90m). This avoided the creation of meshed flat areas. Although a potential horizontal (diagonal) shift of ~64 m – i.e. $SQRT(45^2+45^2)$ – could not be prevented in this case, the use of CC offset some of the negative effects on accuracy, as it interpolated altitude values in response to horizontal cell shifts. CC calculates the weighted average of the value of the 16 surrounding neighbouring cells. Depending on the data, CC can smooth or sharpen the surface. It is critical to note that CC alters cell values when fitting the splines (Leica Geosystems, 2003). Given the relative large reduction in resolution (90 to 15m) however and the rugged terrain where positions of mountain ridges are important, the *reduced* loss in horizontal accuracy is considered worth the loss in vertical accuracy (found to be about 2 m near sea level and 20 m at 2000m altitude). Moreover, when resampling between such resolutions the use of CC rather than NN is not uncommon (Leica Geosystems, 2003).
After this projection / resampling process, the slope tool in ArcMap was used and the data resampled to 15m using NN. Finally, the reference land area was extracted and the data converted into ascii format.

Geology data were obtained from NHMC at the University of Crete (NHMC, 2006) from in vector format. After appending the English formation names, the data were projected from the original Greek Grid projection into the target projection of this study. This involved conversion of the geographic coordinate system and datum as well, because the original datum and GCS were in GGRS 1987 based on a GRS80 ellipsoid with central meridian at 24° E (NHMC, 2006). Next, the data were converted from vector to raster format with 15m spatial resolution, keeping the

geological code as z value. Finally the data were converted into ascii format. When interpreting results, the (unrealistic) abrupt change in cell values at former polygon boundaries must be kept in mind.

Soil type information was derived from two sources. Firstly, general soil type data were acquired via ITC from Wageningen University in form of a paper map. This "soil type WU" data is dated 1986 and original resolution is specified as 1:100,000, i.e. ~1 km cell size (Wageningen University, 1986). The data were digitized onscreen, projected into the target projection, resampled to a 15 m raster grid and converted into ascii format. When interpreting results, the (unrealistic) abrupt change in cell values at former polygon boundaries should therefore be kept in mind.

A more detailed set of soil variables was downloaded from the European Soil Database (ESDB) v 2.0 Raster Library 1x1km (EC-DGJRC, 2006). These variables are Dominant Parent Material, Depth to Rock, Volume of Stones and the full soil code of the Soil Typological Unit (STU) from the World Reference Base (WRB) for Soil Resources. As the provided grid however failed to cover all land areas along the shoreline, the original polygons delineating the Soil Mapping Units (SMUs) were extracted from the ESDB v 2.0, which is in vector format, owned by the European Soil Bureau Network and has been made available for this research purpose to ITC (ESBN, 2004). Surprisingly, the current vector database appeared to lack data for some soil variables, whereas the raster database included these. It was therefore decided to use the polygon SMUs as framework, to modify polygons where they represented shoreline by using the SRTM DEM as reference, and to manually fill all polygons with the respective nominal values provided in the raster library. Thus NoData areas were minimized in extent, shorelines for this variable matched the others used in this study, and information content was maximized.
All soil variables extracted from the ESDB share the following characteristics: they were generated at a scale of 1: 1,000,000 and values were usually estimated by expert judgement interpreting and synthesizing national and regional larger-scale maps. Given the coarse scale of the ESDB data, precision of variables is considered weak and SMU polygon delineations do not fully reflect the soil heterogeneity . Most of the data dates back to the 1980s (ESBN, 2007). Additional quality information (purity and confidence level) is provided in the database. For the ESDB variables used in this study, reference is therefore made to this source (ESBN, 2007), as it is identical to the quality maps associated with the specific thematic "Dominant Value" maps of the ESDB Raster Library.
Upon completion of the pre-processing steps detailed above, all four soil variables

were projected into the target projection, converted to raster format with 15m grid cells and finally converted into ascii format.

A rudimentary dominant land use variable was downloaded from the European Soil Database (ESDB) v 2.0 Raster Library 1x1km (EC-DGJRC, 2006) and subjected to the same pre-processing procedure as the soil variables from this source (described earlier).

2.2.4. Pre-processing of NDVI variable

To generate a vegetation variable, the free ten day synthesis product VGS-S10 was obtained via ITC from VITO Belgium (CNES, 2007). Only the Normalized Difference Vegetation Index (NDVI) data contained in that product was used in this study. Note that the ten-day-synthesis data are maximum value composites, i.e. cells show the highest value a given pixel displayed during these ten days, mostly to ensure minimisation of cloud coverage effects. NDVI is a ratio index that exploits the high reflectance of plant biomass in the Near InfraRed (NIR) region compared to the fairly low reflectance in the Red (R) region of the electromagnetic spectrum: NDVI = (NIR-R) / (NIR+R) (Jensen, 2005). Generally speaking, positive NDVI values indicate green vegetation, negative values indicate water surfaces, near zero values indicate surfaces dominated by soils, low positive values represent either coniferous (if absolute values high) or brown vegetation (if absolute values low) (Jensen, 2005).

The aim was however, to generate a vegetation variable that would be both *representative* and distinguish different vegetation *types*. To achieve this, the 252 ten-day-synthesis data sets covering the time period from April 1998 to March 2005 were stacked (using a batch file in ERDAS) to extract vegetation profiles for classification. Note that layer 4 and 7 proved to be clearly displayed, which was manually corrected: layer 4 was shifted +1000m and layer 7 -1000m along the X axis; all Y values remained unchanged. Next, the resulting multi-band image was classified using ISODATA with convergence threshold set to 1, and the number of iterations slightly higher than half the number of chosen classes. To determine an optimum number of classes, signature separability (in Euclidean distance) was calculated for each classified image in ERDAS' Signatur Editor and plotted in Excel. It was found that classification of this NDVI variable for Crete (including a 5km coastline buffer) would be both detailed enough and most robust if 30 classes were distinguished (Figure 3).

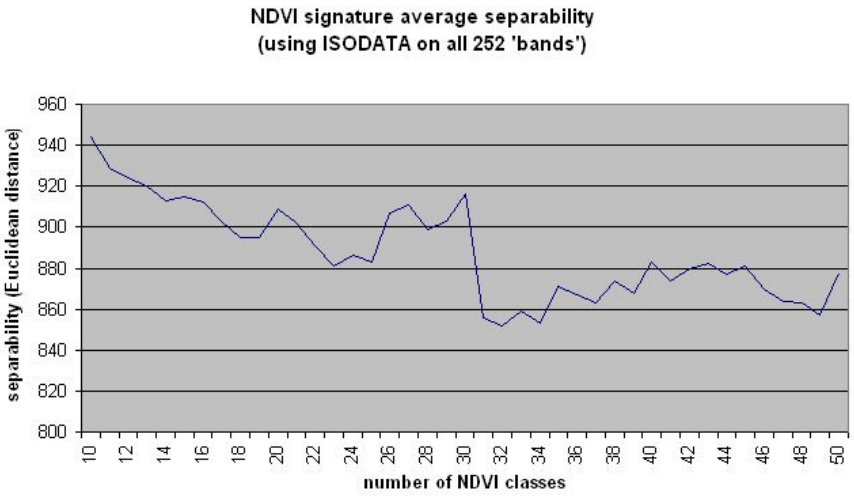


Figure 3 Determining optimum number of NDVI classes

Higher separability was only obtained for < 13 classes (including coastal waters!), which would have been too coarse. By copying the results from the Signature Editor in ERDAS into MS EXCEL, the mean 7-year-NDVI annual values for each of these vegetation classes could be plotted.

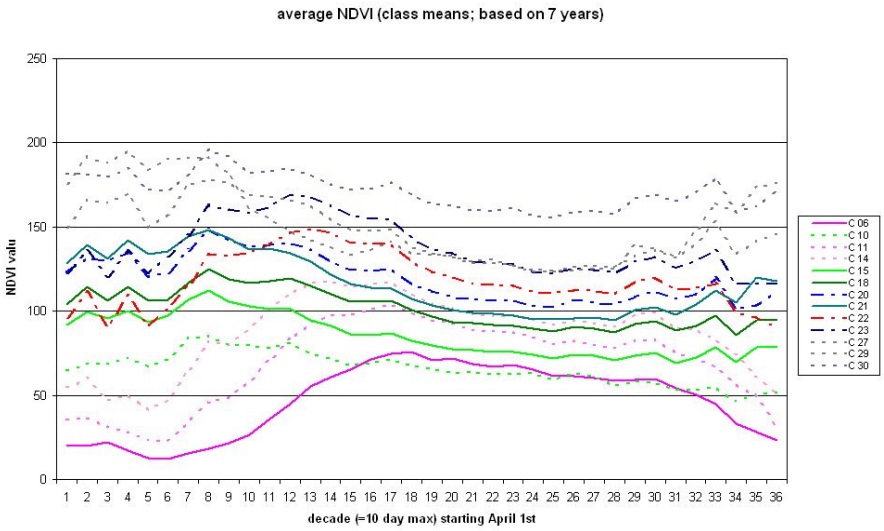


Figure 4 Classified seven-year-mean seasonal NDVI values

To preserve clarity, the figure above shows only some of the 30 classes. By visually analyzing and grouping classes with similar seasonal behaviour (shape) – albeit at different absolute NDVI levels – this classification could be improved further (note for instance the pink and green 'groups'; Figure 4 will be referred back to in the chapter 4.4). For the objectives of this study however this was not essential. Likewise, it was outside the scope of this project (and fieldwork) to assign vegetation cover names to these classes e.g. by in-situ validation, as well as to perform class-specific temporal trend analysis, which these data provide an excellent starting point for.

2.2.5. Pre-Processing of Land Cover Variables

Two different land cover variables were selected for this study; only one of them however – CORINE – was available for the complete study area and obtained via ITC from the European Environment Agency (EEA, 2006). CORINE is the result of a collaborative effort of 12 EU states. It is based on satellite data (Landsat TM, MSS and SPOT XS) as well as auxiliary data in form of national topographic and thematic maps, statistical land cover information and aerial photographs. These data were merged applying expert knowledge and following a standard procedure. Although input data varied in scale, a common working scale of 1:100,000 was adopted in the CORINE product with the smallest unit at least 25 ha in size (EEA, 2000). Procedures for updating CORINE are specified, but given the heterogeneity of initial sources no specific actuality information is available for the area of Crete. The nomenclature of CORINE comprises three levels. This study uses only the third and most detailed one.
As CORINE was provided in vector format, the following pre-processing steps were performed: 1) appended nomenclature to shapefiles, 2) projected data from its initial projection (ETRS1989) into target projection, 3) converted into raster format with 15 m resolution, 4) extracted only land areas, 5) converted data into ascii format. When interpreting modelling results, the initial scale and polygon structure of CORINE – and the resulting generalization and abruptness should be kept in mind.

The second land cover variable used in this study - although only for part of Crete – is an ASTER-based land cover classification generated from original Level_1A data. This variable required intensive pre-processing, which is outlined in the following paragraphs. Imagery was obtained via ITC from NASA's Distributed Active Archive Center (NASA, 2007). Although more were obtained and processed, only the following four ASTER granules – two pairs – could be mosaiced with sufficient

spectral accuracy and eventually be used in this study:

AST_L1A_00305142006091614_20061127102622_10621
AST_L1A_00305142006091623_20060825052229_20539
AST_L1A_00306042002091837_20061127102555_9891
AST_L1A_00306042002091845_20061127102607_10406

While the data for the 'main' part of Western Crete is dated 2006, the most recent data covering the 'far-west' part of Western Crete dates back to 2002. Although early summer is not an optimal time to acquire images for a land cover classification due to vegetation dynamics (Lewinski, 2005), the seasonal mismatch of little more than two weeks (mid-May vs. early June) was fortunately small – and no better alternative cloud free imagery available.
As a thorough analysis of the fairly wide overlap area showed, changes in surface reflectance between summer 2002 and 2006 were minimal – a precondition for the joint unsupervised classification intended.
The ASTER (Advanced Spaceborne Thermal Emission and Reflection radiometer) sensor consists of three instruments (for details see Abrams, 2000), of which only two were used in this study: the Visible and Near InfraRed (VNIR) comprising 4 bands with 15m resolution; band 3B however, the backward looking channel in the near infrared spectrum, which provides stereo vision and thus the potential for 3D imagery, was discarded. Although the 6 bands recorded in the Short Wavelength InfraRed (SWIR) spectrum at 30m resolution, were found to not improve land cover classification of urban areas (Stefanov and Netzband, 2005), for natural vegetation areas all nine bands have been used successfully (Marcal et al., 2005). Although some land cover studies have used all 14 bands (Wang and Zhang, 2006), Thermal InfraRed (TIR) bands were not considered in this study, because (a) its low resolution of 90m and (b) thermal irradiation conditions tend to exhibit stronger and more frequent temporal changes than VNIR and SWIR reflectance patterns (unnecessarily jeopardizing the representativeness of the land cover classification).

As shown in Table 2 only one of band 5 and 6 or 7 and 8 could have been used due to the high inter-band correlation. To preserve the potential to derive complete and meaningful ground feature signatures at a later stage (which however did not materialize), it was decided against this option.

Band	1	2	3	4	5	6	7	8	9
1	1,00000	0,98502	0,91928	0,94364	0,94975	0,95566	0,95895	0,95956	0,94851
2	0,98502	1,00000	0,92254	0,95641	0,95799	0,96524	0,96699	0,96731	0,95553
3	0,91928	0,92254	1,00000	0,97482	0,96365	0,96107	0,95424	0,95411	0,93965
4	0,94364	0,95641	0,97482	1,00000	0,99142	0,99283	0,98609	0,98672	0,97215
5	0,94975	0,95799	0,96365	0,99142	1,00000	0,99341	0,98342	0,98722	0,97507
6	0,95566	0,96524	0,96107	0,99283	0,99341	1,00000	0,99397	0,99049	0,97930
7	0,95895	0,96699	0,95424	0,98609	0,98342	0,99397	1,00000	0,99031	0,98211
8	0,95956	0,96731	0,95411	0,98672	0,98722	0,99049	0,99031	1,00000	0,98461
9	0,94851	0,95553	0,93965	0,97215	0,97507	0,97930	0,98211	0,98461	1,00000

Table 2 Correlation Matrix of ASTER bands (VNIR and SWIR)

Level 1A imagery was obtained in order to allow orthorectification after images had been converted into the target projection. Adjusting for vertical displacements was deemed necessary given the maximum elevation of above 2000 m in the study area and a generally rugged terrain. This step was performed in Leica's Photogrammetry Suite using the DEM interpolated with cubic convolution to 15m resolution. Next, the two granules from each year (2002 and 2006) were mosaiced together which posed no problem as they were both taken during the same overflight. Following the stacking of layers, manual horizontal shift corrections between bands were made (in ArcMap 9.1) using a detailed road layer as reference. To mosaic both paired granules, the relative atmospheric correction approach was chosen (Jensen, 2005), as data on exact atmospheric conditions for an absolute atmospheric correction method such as 6S (Vermote et al., 1997) were not available. After identification and delineation of Pseudo-Invariant Features (PIFs), all bands were unstacked once more and each band of the 2002 imagery multiplied with the PIF coefficients to adjust reflectance levels to the master image (Appendix I).

Finally, after stacking all mosaiced bands once again, the image was classified using the ISODATA algorithm. The optimum number of classes (35) was found following the procedure described above for the NDVI variable. The class with the best minimum (and accidentally best average) Transformed Divergence (TD) distance value was chosen (see Appendix J). After clipping the classified image to the common extent of all predictors and visually identifying and assigning NoData values to all clouds, the file was converted into ascii format for input into Maxent.

ASTER imagery was prepared in a similar fashion for the most of Central Crete as well;

2.2.6. Pre-Processing of Climate Variables

Climate variables for this study were obtained from two main sources. The first group contains the variables Cloud coverage, Potential and Actual

Evapotranspiration. These data were received as offline content via ITC from USGS (NIEHS, 2007). They form part of the Global Climate Database, which contains interpolated grids of climatic variables based on 16,000 to 20,000 stations worldwide recorded between 1931 and 1960 and published in the November 1991 Research Report RR-91-18 by the International Institute of Applied Systems Analyses (IIASA). The database was last updated in 1996 (NIEHS, 1996).
Cloud coverage is defined as "the actual number of bright sunshine hours over the potential number" (NIEHS, 1996) and therefore a % value, which increases with cloudiness. Potential evapotranspiration is defined as the ability of the atmosphere to remove water from the surface by evaporation or transpiration, while actual evapotranspiration represents the amount of water actually removed, i.e. this variable considers water supply too. Values for both are recorded in mm.
As these data were provided in vector format as a very small-scale interpolated grid (only 4 data points covering the study area!), additional pre-processing was performed: 1) projected the data into target projection, 2) clipped a circular area around Crete which included about 100 data points on neighbouring landmasses (note: for sea areas no data was available) and finally 3) applied a similar interpolation technique as the one used during creation of this database: a regularized spline interpolation with the weight parameter at default 0.1, the number of neighbouring points to include set to 30 and a final resolution of 1 km. This interpolation method is considered suitable for variables changing gradually over large distances, as it passes exactly through the provided points while minimizing curvature of the surface (ESRI Inc., 2005a). The resulting surface was then resampled to 15m prior to extracting land areas only and converting these into ascii format. When interpreting the result / significance of these variables as model input, the mediocre accuracy associated with this interpolation should be taken into account.

The second group of climate variables was downloaded from the Worldclim database (Hijmans et al., 2007). This group contains the selected variables Mean Annual Temperature, Minimum Temperature of Coldest Month, Temperature Seasonality, Isothermality and Annual Precipitation. As the previous climate variables, these data were derived from high-resolution interpolation of terrestrial stations (i.e. not remote sensing data). The selected data were provided in raster format with 30 arc seconds resolution (~1 km grid cells), referenced to Lat/Long degrees and the WGS84 datum. The data is a synthesis of several other climatic databases. All values are based on monthly measurements taken mostly between 1960 and 1990 from ~15,000 to ~48,000 stations worldwide. Hijmans et al. (2007)

used thin plate smoothing splines as well as a DEM derived from the SRTM (U.S.G.S.) to generate the final interpolated climate grids. Temperature and precipitation variables consist of original monthly data. The other bioclimatic variables represent annual trends derived from these original data. Note that temperature values must be multiplied by 0,1 in order to retrieve the correct value in °C. Temperature Seasonality equals the standard deviation * 100 of the annual temperature values. Isothermality relates Mean Diurnal Range to Temperature Annual Range, i.e. expresses the similarity of daily to yearly variations in temperature. The equation applied was: 'Mean of monthly (max temp - min temp) / (Max Temperature of Warmest Month - Min Temperature of Coldest Month) * 100' (Hijmans et al., 2005).

Unfortunately, the provided grids did not cover all land areas completely, which would have resulted in the exclusion of several species observation sites when modelling with these layers. Some pre-processing was therefore necessary for the purpose of this study: 1) clipped out study area 2) if necessary, calculated annual average values from monthly layers, 3) projected data into target projection using NN and accepting default cell size (any smaller cell size proved to be too processing intensive in the next step; "new" cells potentially shifting up to ~500m horizontally was thus accepted), and 4) carried out Inverse Distant Weighted (IDW) interpolation with power set to 2 (default), number of points to be considered set to 5 and a specified vector file acting as barrier during interpolation. The IDW method was chosen because the provided climate grids showed considerable local variation, and the intention was to model the coastal vicinity as close as possible to the local values while preserving the exact values of the provided grids – most other methods would yielded too smooth interpolated surfaces or altered existing values (ESRI Inc., 2005b). To ensure the latter, a vector file was created with line features running parallel to the coastline. When used as a barrier file, interpolation was restricted successfully to the Nodata coastal land areas using only the points located no more than ~ 1 km away from the coast. To avoid consideration of relatively distant coastal point, their number was limited to 5 only. During interpolation, the size of cells with interpolated values was maintained in order to avoid 'false precision'. Finally, all data were turned into integer format and resampled to 15m cell size, before the land areas were extracted and converted into ascii format.

2.3. Modelling Technique

2.3.1. Maxent as Statistical Model

This chapter describes the main principle and characteristics of Maxent and justifies why Maxent – developed by the machine learning community – was the statistical model of choice in this study.

Using the predictor values at the provided species presence locations as well as (default:) 10,000 randomly chosen background pixels from the predictor grids, Maxent computes many probability distributions across all grid cells. The algorithm used follows the maximum entropy principle (closest to uniform), as it includes as many 'options' as possible into the probability distribution while simultaneously excluding all 'options' known to be outside the target distribution – as specified by certain constraints. Constraints are quantified as 'features' in Maxent. They represent the few characteristics known about the target distribution. The process is iterative and starts with assuming a uniform probability distribution. As each feature and its relative weight gets sequentially updated, the 'gain' increases exponentially at suitable locations. A 'gain' value of 1.8 thus indicates that the distribution fits the training sample points ~ 6 times [=exp(1.8)] better than a random distribution. Eventually the distribution is chosen, which has the highest entropy *and* meets the provided constraining criteria. Six different features are available; some express the state of single variables, others express their interaction.

By default only those features are activated for which the provided number of presence records is (automatically) calculated to be empirically sufficient. The effect of these features is that the output probability distribution is forced to remain structurally identical (in terms of statistics) to the distribution of environmental average values measured at the input species observation points. In order to relax this constraint however, 'regularization' is a recommended option. This tool reduces the iterative gain and helps to avoid overfitting of the model especially if few presence records are available. A higher regularization value leads to a wider predicted distribution. To better control predictions outside the range of the training data, a 'clamping' technique is applied. Maxent can handle categorical input data by using background pixel *counts* instead of values to calculate feature averages. It also calculates the relative importance of each predictor using Jacknife (for alternatives see (Verbyla and Litvaitis, 1989). Output is continuous and provided by default as 'cumulative', i.e. cell values indicate the percentage of other cells with equal a equal or lower (raw) probability value. Cells which for which not *all* predictors have a value, are assigned NoData values (Phillips et al., 2004, Dudík et al., 2004, Phillips

et al., 2006).
Maxent was chosen as statistical model primarily because its relative insensitivity to noisy data and small sample size (owing to the generative not discriminative approach), its ability to handle categorical data and to model interactions (Phillips et al., 2006) as well as its promising performance in similar studies (Elith et al., 2006, Hernandez et al., 2006, García Márquez, 2006).

2.3.2. Modelling Procedure

To address research questions (1a,b) the dataset for all of Crete (75 points) was divided into training (75%) and test (25%) presence points. Background points ('pseudo-absences') were randomly sampled from the full study area, because predictor extent covered no other (irrelevant) land masses and (geographic) provided presence records covered the geographic extremities. Given the small sample size however, Maxent was run five times (all values at default of Maxent version 2.3) on both training and test data. Next, using only the top one-third of strongest predictors – defined in terms of strongest added average *training* gain when using built-in Jacknife – another five runs were performed for comparison with previous results.

To address research questions (2a,b), only fieldwork-'enhanced' occurrence sites were selected, partitioned (25% for testing) and run together with all predictors (clipped to Western Crete). Given an even smaller sample size (46 points), seven partitions were created and the five best selected. For visualization purposes, another Maxent run was performed (default settings as before) using a non-partitioned dataset. Note that when using the dataset for Western Crete, random points were only sampled from within a bounding box, which included all 46 'enhanced' presence sites and coincided with available ASTER imagery (extent in target projection: top 3760010m, bottom 3700010m, left 4490000m, right 4575350m; see also Figure 2). Finally, the ASTER-based land cover variable was included as predictor and five (best out of seven) partitions generated using (a) species occurrence sites as before and then (b) occurrence 'plots' instead of 'points' (see chapter 2.24).

Note that predicted distributions for research questions (1a,b) were calculated using not the same number of presence sites for *all* predictors. Unfortunately, NDVI data and all WorldClim variables from Hijmans et al., 2007, failed to cover all of South-Eastern Crete. However, as NDVI was the only predictor not capturing the as the six occurrence sites located there, the inconsistent contribution of this single predictor was accepted. The complete omission of these six sites would have caused a

significant loss in coverage of geographic extremities for *all* other predictors, thus raising the likelihood of extrapolation errors.

To address research questions (3), only a descriptive analysis was carried out, because unfortunately it was not feasible during fieldwork to also collect sufficient ground truth data for a proper evaluation of both classified NDVI and ASTER land cover variables, based on a confusion matrix and e.g. Kappa statistics. These data have instead been subjected to an overlay analysis with the CORINE predictor, thus inferring preliminary labels for both unsupervised classifications. Additional information on apparent habitat preferences of *P. erhardii* were derived from simple frequency statistics of other ground related predictor layers. All variables were analyzed using their total extent across Crete with the exception of the ASTER predictor. All counts were related to the respective 'background' land area (all of Crete, Western Crete respectively), thus taking into account the element of chance.

2.3.3. Evaluation Methods

Each modelled prediction was evaluated by testing (1) if probability values at test (and training) localities were predicted better than random (*significance*) using a simple binomial z / t statistic (Congalton, 1991). In place of Kappa (Cohen, 1960), the equally threshold-dependent but more robust (2) True Skill Statistic (TSS) (Allouche et al., 2006) was employed to assess if predictions were better than random (*significance*). Finally, the (3) threshold-independent Receiver Operator Characteristic (ROC) was employed, which yields the Area Under Curve (AUC) value as single indicator of model performance (Hanley and McNeil, 1982). Kappa and AUC are commonly applied to species distribution models (Fielding and Bell, 1997, Hirzel and Guisan, 2002, Graham and Hijmans, 2006, Phillips et al., 2006), whereas TSS is a recent suggestion (Allouche et al., 2006).

Although research on selecting the optimal threshold for binary predictions of presence-only models is currently very active, no dominant rule has emerged yet for this task (Phillips et al., 2006, Liu et al., 2005, Hirzel et al., 2006). Objective approaches like maximum Kappa (Guisan et al., 1998) are as frequent as subjective ones based on an arbitrary threshold, e.g. 0.5 or 95% specificity. A comprehensive overview of alternatives is provided in Liu et al., 2005, who conclude that the above approaches were inferior to most others. Hence, this study opts for the 'Sensitivity-Specificity-Equality Approach' (Cantor et al., 1999). It determines the optimal threshold by minimising the absolute difference between computed sensitivity and

specificity. The associated 'cumulative gain' value was looked up in the respective background file which was produced for each distribution using the Sample tool in ArcMap. Background points were generated in MiniTab 14.2 (MiniTab Inc., 2005) for all of Crete (n=3486), because presence sites covered the periphery (mean distance of background points: 775 m, SD =510 m) ensuring minimal extrapolation. For Western Crete (n=1112) a subset was clipped using the extent specified in chapter 2.3.2. Binary predictions were based on the *average* cumulative threshold of five runs. Confusion matrices were populated with counts from the *test* (and also training) dataset as presences, and random background points as 'pseudo-absences' (Graham and Hijmans, 2006).

Cohen's Kappa is one measure that can be derived from the confusion matrix. As a validation tool (i.e. when the 'truth' is known), it states the overall *accuracy* of a prediction once the element of chance has been removed (Cohen, 1960, Congalton, 1991, Liu et al., 2005, Guisan and Hofer, 2003). Otherwise, Kappa can serve as a tool to assess reliability of prediction in terms of relative *agreement*. A value near 0 indicates no discrimination (agreement by chance); a value of 1 represents perfect discrimination (agreement); a value of > 0.6 is considered 'good' and >0.8 as 'excellent' (Graham and Hijmans, 2006). Kappa is relative tolerant to zeros in the confusion matrix and considers both omission and commission errors in one parameter.
However, Kappa is unimodally dependent on prevalence, i.e. on the proportion of all presences in the full validation dataset (Fielding and Bell, 1997, Allouche et al., 2006). This would have been a problem in this project, because known species presences were few (training: ~ 50; test: ~15) compared to the high number of random background points (>1000), which were deemed necessary for a reliable confusion matrix given the large study area. Hence, the TSS (Allouche et al., 2006) was computed instead of Kappa. It is defined as TSS = sensitivity + specificity – 1. In contrast to Kappa, TSS values can be used to compare prediction performance independent of both validation dataset size and the prevalence contained therein, while still featuring the same strengths of Kappa: full consideration of sensitivity, specificity and chance (Allouche et al., 2006). The range of TSS values and their 'translation' is identical to Kappa. The features of TSS were also advantageous in addressing research question 2b. Note that if background sampling is adjusted to ensure a prevalence of 50%, Kappa is in fact identical to TSS (Allouche et al., 2006, Hirzel et al., 2006).

As ROC (AUC) analysis (Hanley and McNeil, 1982) is independent of both

threshold setting and prevalence, it is a highly effective method for assessing the performance of ordinal score (i.e. presence-only) distribution models (Allouche et al., 2006). The AUC is derived by using all possible thresholds to plot sensitivity (the probability that a model correctly classifies a presence) versus specificity (the probability that a model correctly classifies an absence). A value of 1 stands for perfect discrimination, if above 0.75 the model is rated 'good', while a value of 0.5 indicates a performance not better than chance (Graham and Hijmans, 2006). The AUC represents thus a single number to denote model performance at all threshold levels (Fielding and Bell, 1997).

However, as presence-only models lack 'true absences', specificity can not be calculated rendering AUC inapplicable – unless a conceptual modification is accepted and 'fractional predicted area' is used instead of true commission (Phillips et al., 2006). The same 3864 background points (1112 in the West) created for TSS evaluation were hence used as 'pseudo-absences'. All calculations were made in the ROC Plotting and AUC Calculation Transferability Test v 1.3 software (Schröder, 2004), which employs bootstrapped confidence intervals calculated with the percentile method described in Buckland et al., 1997. Significance was tested against AUC = 0.5 using the built-in z score functionality at 95% confidence level. The AUC values generated by Maxent (with a larger and different background sample) facilitated a useful comparison. Relative predictor importance was investigated based on Maxent's built-in Jacknife functionality.

3. Results

3.1. Research Question 1a: Prediction Aross Crete

Perhaps surprisingly, a model can be fitted to 'explain' the curious observed presumable distribution of *P. erhardii* across Crete (**Figure 5**). The map was produced using *all* qualified occurrence data (for visualisation purposes only). Note that all *accuracy* assessments in this study are based on modelled predictions using *partitioned* occurrence data.

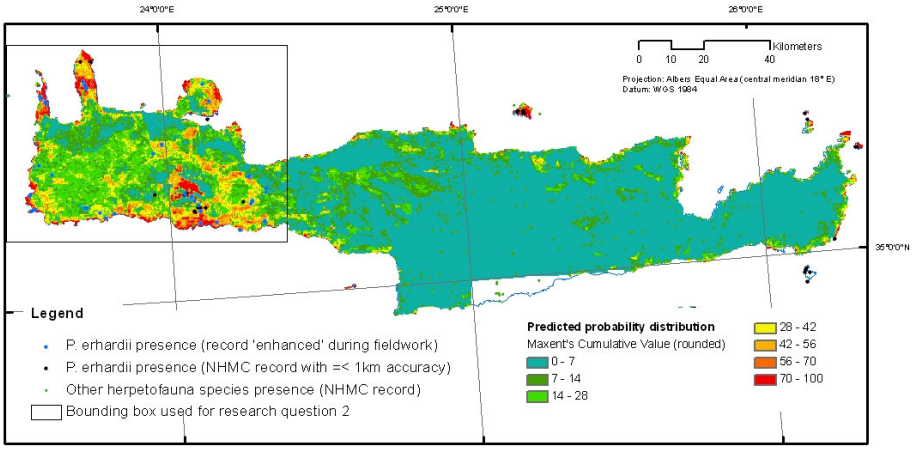


Figure 5 Probability distribution across Crete

Visual analysis suggests very low habitat suitability (more precisely: potential species distribution) for almost the entire Central and Eastern parts of Crete. It is important to note that poor suitability is also predicted for some areas in Western Crete. It is therefore less likely that the prediction suffers from a dominant variable, whose West-East gradient has introduced a drastic spatial autocorrelation bias. All areas coloured yellow and red are considered suitable habitat when applying the binary threshold of 27.33 (cumulative value) found to be optimal in the previous chapter. Although many coastal areas feature highest suitability – including many islets in the East – there are significant stretches of coastline displaying very low suitability. The very strong probability in high altitude areas of Levká Orí may be an artefact of site selection bias, as sampling density was locally high and points coincided strongly with a particular CORINE class. Predicted habitat suitability is generally highest in 'contiguous patches' in the West and isolated coastal areas often only a few kilometres in size.

All five partitioned distributions predict values at *test* locations significantly better than random (one-tailed t-test for one proportion, averages: $z = 4.8380$, $p = 0.0030$). A t-test was required as $n < 30$ (with normal distribution). Background points were kept identical for each run (mean background cumulative value 11.2; proportion classified as presence 14.3 %).

Model performance							Excel						ROC-PLOT					Maxent	
	binomial test						threshold$_{opt}$ = 27.33						threshold-independent (bootstrapped)						
	presence/absence		Ø cum val		st.dev		TSS		st.dev	95% CI		Kappa	AUC	95% CI				AUC	
	train	***test***	*train*	***test***	*train*	*test*	*train*	***test***		*lower*	*t value*	*test*	***test***	*SE*	*lower*	*z value*	***p value***	*train*	*test*
run-1	(50/7)	(12/3)	64.9	**52.5**	27.4	31.8	0.80	**0.72**	0.10	0.67	27.9	0.02	**0.88**	0.05	0.77	7.103	**< 0.0001**	0.948	0.889
run-2	(47/10)	(12/5)	68.1	**42.9**	29.0	26.4	0.75	**0.63**	0.11	0.58	23.6	0.02	**0.87**	0.04	0.79	9.097	**< 0.0001**	0.953	0.877
run-3	(51/6)	(12/5)	66.7	**53.7**	25.3	31.9	0.82	**0.63**	0.11	0.58	23.6	0.03	**0.88**	0.04	0.80	8.877	**< 0.0001**	0.958	0.885
run-4	(48/9)	(8/8)	68.8	**31.3**	29.1	33.4	0.76	**0.42**	0.13	0.37	13.5	0.02	**0.72**	0.07	0.58	3.020	**0.0025**	0.957	0.733
run-5	(50/7)	(12/2)	68.1	**50.2**	28.4	23.1	0.80	**0.78**	0.09	0.74	32.4	0.02	**0.94**	0.02	0.91	27.873	**< 0.0001**	0.947	0.923
mean	(49/8)	(11/5)	67.3	**46.1**	27.8	29.3	0.79	**0.64**				0.02	**0.86**		0.77			0.953	0.861
b'ground	(501 / 3486)		11.2		16.8		prevalence = (0.013 / 0.005)											different b'g	

Details of significance tests above

Test	one-tailed z/t-test	one-tailed t-test at 95% CI	AUC tested with one-tailed z-test at 95% CI
H_0	Ø cum val <= 11.2	$TSS_{average}$ <= 0	AUC <= 0.5
H_A	Ø cum val > 11.2	$TSS_{average}$ > 0	AUC > 0.5

results for **binomial** test of **mean** Ø cum. value:
mean SE = 3.68 (train), 7.37 (test); lower 95% CI = 61.24 (train), 39.12 (test); t value = 4.8380 (test), z value = 15.24 (train), p value = 0.000 (train), 0.0030 (test)

Table 3 Evaluation of distribution across Crete

As expected, Kappa values are heavily affected by the low prevalence of < 1%, ranking barely above 0 even for the training dataset. The non-affected TSS rates the models performance (only) 'moderate' (TSS = 0.5701), based on the average of five partitions and the given threshold. This result however still differs significantly from chance with TSS = 0 (one-tailed t-test at 95% CI, averages: t =4.838, p = 0.003). The model performed very well on this test data given an average AUC of 0.8574. Note that AUC values produced with ROC-PLOT were nearly identical to the ones generated by Maxent. As expected, model performance in all respects is generally better when evaluated against *training* data.

In answer to hypothesis (1a) a one-tailed t-test was performed on the five *training* partitions which together hold an average regularized training gain of 1.858with H_0: cumulative gain <= 1.5 and H_A: cumulative gain > 1.5. The result allows to reject hypothesis H_0 and to accept H_A (t = 5.34 and p = 0.003 at 95% CI).

3.2. Research Question 1b: Strongest Predictors Aross Crete

By far the most important predictor in 4 out of 5 random partitions is the NDVI variable, which in itself accounts for half of the overall training gain and would reduce the gain the most if omitted (Figure 6). While land cover (CORINE) and altitude rank next in terms of unique information carried, cloud cover and actual evapotranspiration contribute each slightly more to the overall gain. Note that this identification of the six most important variables is based on un-partitioned occurrence data, the idea being to utilize the maximum information available.

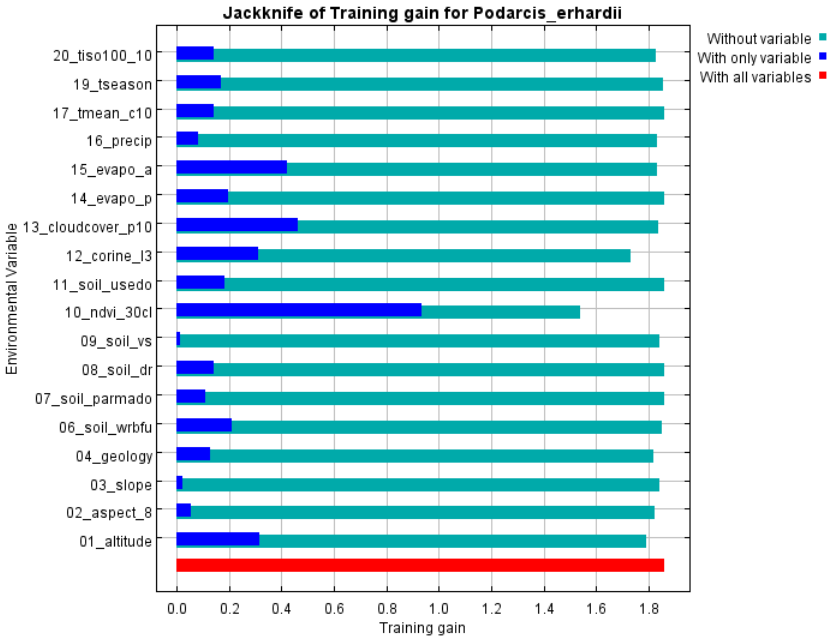


Figure 6 Jacknife results on variable importance across Crete

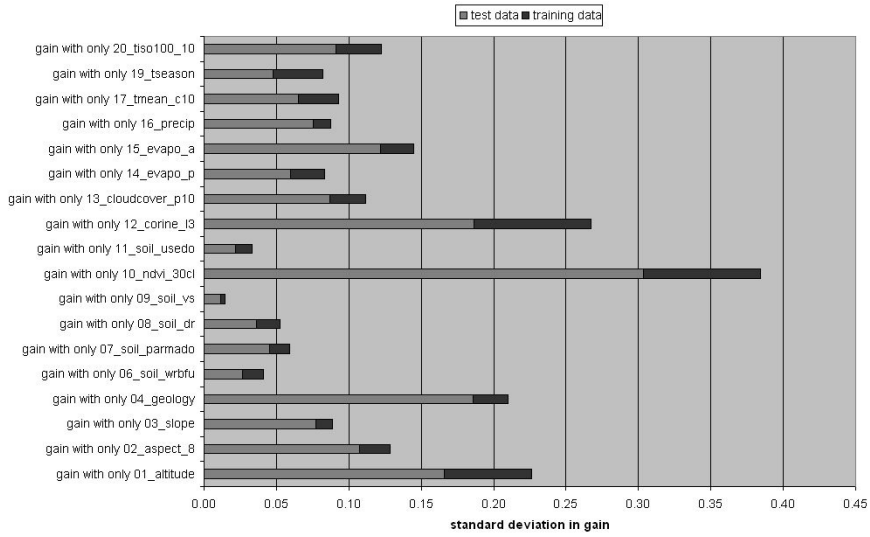


Figure 7 Variability in 'cumulative gain' of predictors across Crete

In addition to absolute gain it is essential to consider predictor variability (Figure 7). As perhaps expected, topographic variables like NDVI, CORINE, altitude and

geology display a much higher standard deviation (in the five test partitioned runs) as climate related predictors. In fact, the individual gain values for geology, CORINE and altitude predictors turn out to be statistically not significant when validated on the test dataset (two-tailed t-test; with H_O: gain = 0 and H_A: gain ≠ 0; 95% CI, t = .03, .06, -.26 respectively; p = .979, .956, .807 respectively). All other predictor gains however (and *all* predictors when using training data) *are* statistically significant – including NDVI. The predicted range preferences of *P. erhardii* within the ecological space spanned by these variables is shown in the response curves in Appendix A.

Five new partitions using only the six strongest predictors (as mentioned above plus 'soil type WRB') yield a regularized training gain of 1.502. This value is significantly lower than 1.858, i.e. the one obtained when using all predictors (H_0: cumulative gain >= 1.858 and H_A: cumulative gain > 1.858; upper bound 99% CI = 1.563, t = 92.12 and p = <0.001).

Model performance							Excel						ROC-PLOT					Maxent	
			binomial test				threshold$_{opt}$ = 27.33						threshold-independent (bootstrapped)						
	presence/absence		Ø cum val		st.dev		TSS		st.dev	95% CI		Kappa	AUC		95% CI			AUC	
	train	***test***	*train*	***test***	*train*	*test*	*train*	***test***		*lower*	*t value*	*test*	***test***	*SE*	*lower*	*z value*	***p value***	*train*	*test*
run-1	(51/6)	(13/3)	61.8	**49.1**	25.9	24.9	0.75	**0.67**	0.10	0.63	26.8	0.04	**0.90**	0.03	0.84	12.156	**< 0.0001**	0.936	0.890
run-2	(51/6)	(6/10)	63.8	**31.6**	24.8	32.1	0.75	**0.23**	0.12	0.18	7.7	0.01	**0.73**	0.06	0.61	3.707	**0.0002**	0.952	0.748
run-3	(51/6)	(9/5)	61.6	**40.8**	26.5	28.2	0.75	**0.50**	0.13	0.44	14.4	0.03	**0.82**	0.05	0.73	6.799	**< 0.0001**	0.932	0.838
run-4	(54/3)	(10/5)	61.2	**40.1**	23.3	26.8	0.80	**0.48**	0.12	0.43	16.0	0.03	**0.82**	0.07	0.69	4.912	**< 0.0001**	0.952	0.814
run-5	(51/6)	(15/2)	60.9	**51.6**	25.4	23.1	0.75	**0.74**	0.08	0.71	38.1	0.05	**0.91**	0.03	0.85	13.235	**< 0.0001**	0.936	0.912
mean	(49/8)	(11/5)	61.9	**42.6**	25.2	27.0	0.76	**0.52**				0.03	**0.83**		0.74			0.942	0.840
b'ground	(501 / 3486)		13.0		17.6		prevalence = (0.013 / 0.005)											different b'g	

Details of significance tests above

Test	one-tailed z/t-test	one-tailed t-test at 95% CI	AUC tested with one-tailed z-test at 95% CI
H_0	Ø cum val <= 13.0	$TSS_{average}$ <= 0	AUC <= 0.5
H_A	Ø cum val > 13.0	$TSS_{average}$ > 0	AUC > 0.5

results for **binomial** test of **mean** Ø cum. value:
mean SE = 3.34 (train), 6.75 (test); lower 95% CI = 56.41 (train), 30.77 (test); z value = 14.65 (train), t value = 4.39 (test); p value = 0.000 (train), 0.000 (test)

Table 4 Evaluation of distribution across Crete using top six predictors only

As NDVI (individual training grain 0,9366) characterizes primarily ground vegetation, hypothesis (1b) H_O must be rejected and H_A accepted. With reference to the extension of research question (1b), it can be concluded that the top one-third of predictors tested here capture about 80% of the response variable in terms of regularized training gain. The resulting probability distribution is shown in Appendix B.

3.3. Research Question 2a: Potential of ASTER Imagery

The effect of adding the ASTER-based land cover variable to the (full) suite of predictors can visually be assessed by comparing Figure 8 and Figure 9. Note that

both distributions were generated using all 46 fieldwork-'enhanced' occurrence sites, i.e. without considering occurrence sites from outside this region (black and grey dots on map). The predicted distribution differs therefore from the previous ones. For an answer to research question 2a and 2b however, this is not relevant.

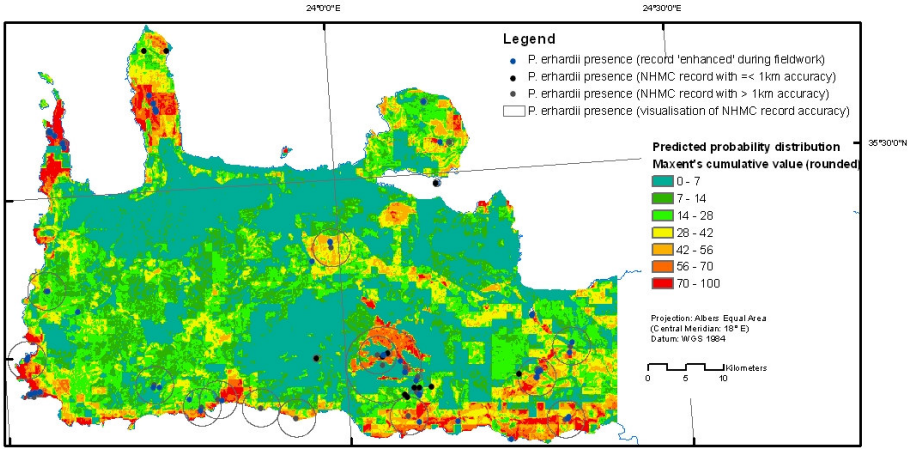


Figure 8 Probability distribution in Western Crete *without* ASTER predictor

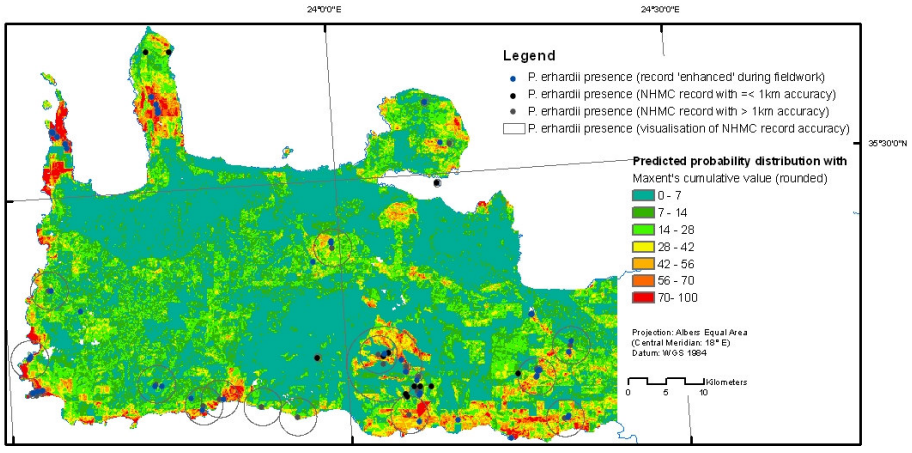


Figure 9 Probability distribution in W. Crete *with* ASTER predictor

The distributions differ in both map output and statistical performance (TSS and AUC). Visual analysis suggests that the inclusion of an ASTER-based land cover variable results in a more concise delineation of areas featuring high probability values. This is logical in the sense that each occurrence point can only relate to a 15m pixel in this layer, whereas for all other layers, an occurrence point relates de

facto to a 90m or 1km pixel (or a rasterized, fairly large polygon). There is thus also a collateral loss in probability for a high number of neighbouring pixels, causing an overall expansion of low probability areas. Note however that many 'red' core areas with no occurrence point inside, remain at their location. This indicates a remarkably strong environmental similarity to known occurrence sites.

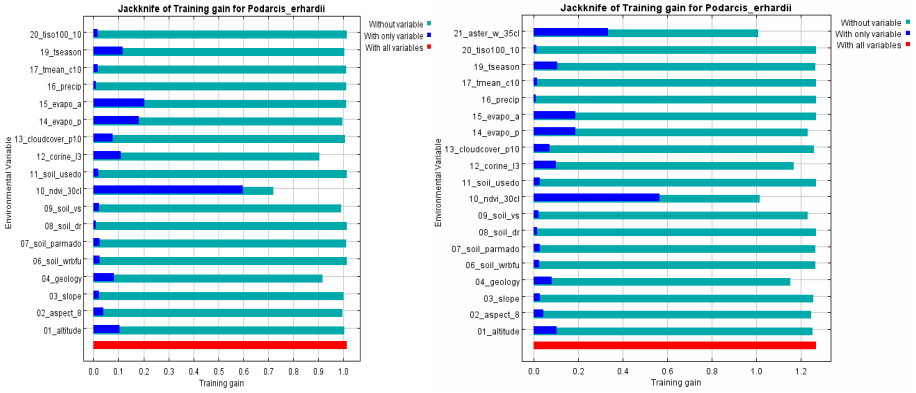


Figure 10 Effects of ASTER inclusion on relative variable importance

As shown in Figure 10, the ASTER-derived land cover variable turns out to carry as much unique information as the NDVI variable. When used as the only predictor, it ranks second after NDVI and is able to generate about 30% of the 'cumulative gain' achievable when all variables are used.

For the binomial tests of significance, the threshold was derived again using the 'Sensitivity-Specificity-Equality Approach' (Cantor et al., 1999) and found to be 28.51 and 31.26, respectively.

Binomial tests on both distributions indicate that average values at *test* sites are significantly higher than the random distribution (one-tailed t-test with H_O: gain <= 18.6 and H_A: gain > 18.6; 95% CI, averages: t = 2.080; p = 0.0320). TSS analysis reveals that models predict the *test* sites only half as good as the training sites with both models ranking 'poor' on test sites (TSS = 0.30; 0.25), but still significantly above chance (one-tailed z-test, average z / t > 4.5, p < 0.001). The good performance of both distributions according to ROC analysis (average AUC = 0.70; 0.76) however, most likely suffers from the low number of partitions, in particular the distribution *without* the ASTER-based variable, as indicated by the higher Maxent AUCs (which use a larger background) and the extremely high standard error of AUC on test data, which causes the significance level to touch the border to pure chance (lower bound 95% CI = 0.50 AUC). Interestingly, the inclusion of the ASTER predictor increases the significance of the AUC, but leads to a lower

significance of the TSS. This may constitute evidence that the latter distribution is overfitted rather than suffering from a low number of partitions (however, TSS is based on two different background data ensuring an optimal threshold).
Note the excellent average AUC value for *training* samples (AUC = 0.893; 0.928).

Model performance							Excel						ROC-PLOT					Maxent	
			binomial test				threshold$_{opt}$ = 28.51						threshold-independent (bootstrapped)						
	(presence/total)		Ø cum val		*st.dev*		**TSS**		*st.dev*	95% CI		Kappa	**AUC**		95% CI			AUC	
	train	***test***	*train*	***test***	*train*	*test*	*train*	***test***	*lower*	*t value*		*test*	***test***	*SE*	*lower*	*z value*	***p value***	*train*	*test*
run-1	(28/34)	(6/11)	61.7	**34.2**	28.0	29.2	0.58	**0.30**	0.15	0.29	67.0	0.02	**0.72**	0.11	0.50	1.976	**0.0482**	0.897	0.689
run-2	(27/34)	(5/11)	61.3	**31.3**	31.3	31.4	0.55	**0.21**	0.15	0.20	46.9	0.02	**0.60**	0.12	0.36	0.842	**0.3999**	0.885	0.666
run-3	(29/34)	(7/11)	61.1	**41.7**	28.0	28.0	0.61	**0.39**	0.15	0.38	87.1	0.03	**0.72**	0.08	0.55	2.610	**0.0091**	0.892	0.763
run-4	(32/34)	(6/11)	63.7	**39.5**	25.7	34.3	0.70	**0.30**	0.15	0.29	67.0	0.02	**0.67**	0.11	0.46	1.561	**0.1185**	0.912	0.733
run-5	(27/34)	(6/11)	60.9	**44.5**	30.5	33.7	0.55	**0.30**	0.15	0.29	67.0	0.02	**0.77**	0.08	0.61	3.364	**0.0008**	0.880	0.778
mean	(29/34)	(6/11)	61.7	**38.2**	28.7	31.3	0.60	**0.30**				0.02	**0.70**		0.50			0.893	0.726
b'ground	(269 / 1112)		18.6		20.8		prevalence = (0.03 / 0.01)											different b'g	

Details of significance tests above

Test	one-tailed z/t-test	one-tailed t-test at 95% CI	AUC tested with one-tailed z-test at 95% CI
H_0	Ø cum val <= 18.6	$TSS_{average}$ <= 0	AUC <= 0.5
H_A	Ø cum val > 18.6	$TSS_{average}$ > 0	AUC > 0.5

results for **binomial** test of **mean** Ø cum. value:
mean SE = 5.33 (train), 9.44 (test); lower 95% CI =52.93 (train), 21.1 (test); t value = 2.08 (test), z value =8.09 (train), p value = < 0.001 (train), 0.032 (test)

Table 5 Evaluation of distribution for Western Crete *without* ASTER

Model performance							Excel						ROC-PLOT					Maxent	
			binomial test				threshold$_{opt}$ = 31.26						threshold-independent (bootstrapped)						
	(presence/total)		Ø cum val		*st.dev*		**TSS**		*st.dev*	95% CI		Kappa	**AUC**		95% CI			AUC	
	train	***test***	*train*	***test***	*train*	*test*	*train*	***test***	*lower*	*t value*		*test*	***test***	*SE*	*lower*	*z value*	***p value***	*train*	*test*
run-1	(31/35)	(7/11)	62.7	**40.8**	26.9	28.7	0.64	**0.30**	0.15	0.22	6.6	0.03	**0.83**	0.07	0.69	4.818	**< 0.0001**	0.925	0.814
run-2	(30/35)	(5/11)	66.7	**32.4**	29.2	26.0	0.62	**0.21**	0.15	0.13	4.6	0.02	**0.75**	0.06	0.64	4.362	**< 0.0001**	0.927	0.751
run-3	(30/35)	(6/11)	65.423	**41.6**	26.9	36.7	0.62	**0.30**	0.15	0.22	6.6	0.02	**0.75**	0.07	0.61	3.425	**0.0006**	0.924	0.774
run-4	(29/35)	(5/11)	64.489	**33.2**	28.3	26.4	0.62	**0.21**	0.15	0.13	4.6	0.02	**0.75**	0.07	0.62	3.638	**0.0003**	0.923	0.761
run-5	(32/35)	(5/11)	66.7	**36.3**	24.2	32.5	0.67	**0.21**	0.15	0.13	4.6	0.02	**0.72**	0.10	0.53	2.236	**0.0253**	0.940	0.756
mean	(30/35)	(6/11)	65.2	**36.9**	27.1	30.1	0.63	**0.25**				0.02	**0.76**		0.62			0.928	0.771
b'ground	(269 / 1112)		13.5		19.0		prevalence = (0.03 / 0.01)											different b'g	

Details of significance tests above

Test	one-tailed z/t-test	one-tailed t-test at 95% CI	AUC tested with one-tailed z-test at 95% CI
H_0	Ø cum val <= 13.5	$TSS_{average}$ <= 0	AUC <= 0.5
H_A	Ø cum val > 13.5	$TSS_{average}$ > 0	AUC > 0.5

results for **binomial** test of **mean** Ø cum. value:
mean SE = 4.55 (train), 8.65 (test); lower 95% CI =55.22 (train), 25.12 (test); t value = 3.15 (test), z value =10.82 (train), p value = 0.000 (train), 0.005 (test)

Table 6 Evaluation of distribution for Western Crete *with* ASTER (points)

From these results it can be concluded that adding the ASTER-based land cover variable significantly increases both model fit (unpartitioned distribution: regularized training gain = 1.27 up from 1.01) and performance (AUC = 0.9242 up from 0.8942). The Null hypothesis of research question 2a) must thus be rejected and H_A accepted.

3.4. Research Question 2b: Replacing Occurrence Points with 'Plots'

Based on visual comparison of Figure 8 Probability distribution in Western Crete *without* ASTER predictor (above) with Figure 11 (below), a spatial consolidation of probability areas can be observed at all levels: while scattered low probability patches disappear, core areas grow more contiguous in shape. This effect was expected because the 'plot' approach aims at capturing apparently suitable habitat locations in a more representative way than by a simple XY point location. Qualified surfaces in the immediate surroundings are thus included in the model input. See Appendix E for a direct comparison clip of all modelled distributions. Note that the calculated optimal binomial threshold for this distribution is 46.76; predicted presences are thus marked by 'orange and warmer' colours (not yellow).

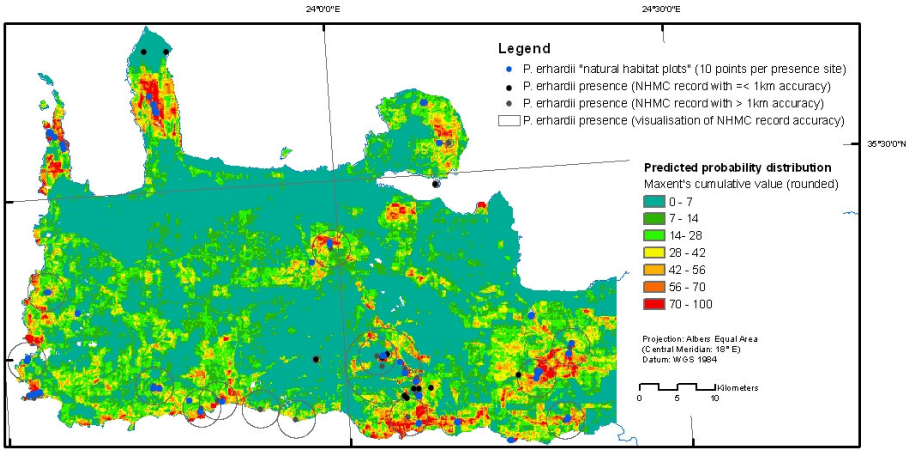
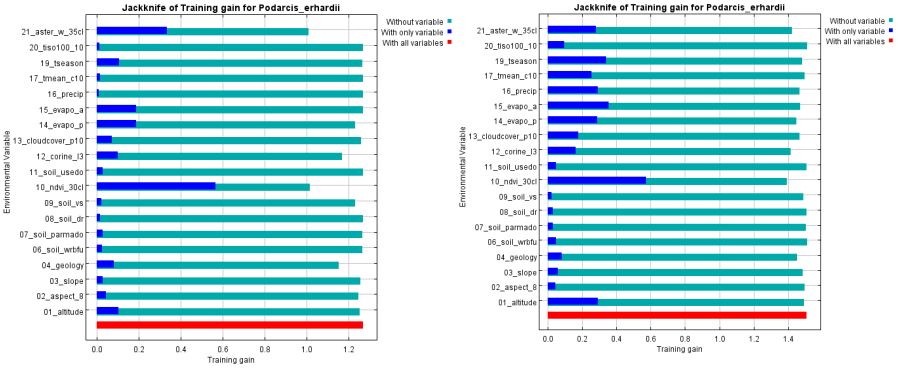


Figure 11 Probability distribution using ASTER predictor and 'plots'

Replacing single point occurrence data with a 10 point occurrence 'plot' situated in 'natural habitat' as identified in-situ, increases the regularized 'cumulative gain' for this distribution model to 1.51 up from 1.27. While NDVI remains the most important predictor, it does not contribute to the additional gain and also stands no longer apart in terms of much unique information carried (just like all other predictors). The ASTER-derived land cover predictor performs in much the same way. While continuous predictors including altitude, seasonality and isothermoality triple their individual contribution and precipitation and temperature even increase it about tenfold(!), most categorical predictors remain constant (Figure 12). This result is somewhat unexpected and will be discussed in chapter 4.3.



Q2a: using 1 single point as presence Q2b: using 10 points as presence 'plot'

Figure 12 'Plots' replacing 'points': effect on variable importance

All results obtained from the binomial test and TSS are found to differ very significantly from random (all z values > 30) and extremely well fitted (AUC_{test} = 0.97). See Table 6 for more statistic details. Although the Null hypothesis 2b) must be rejected and H_A be accepted, this result requires thorough interpretation (see chapter 4.3).

3.5. Research Question 3: Surface Cover Preferences

As pointed out in chapter 2.3.2, analysis of *P. erhardii's* surface cover preferences is constrained to simple descriptive statistics in this study, primarily because insufficient ground truth could be obtained during fieldwork for a supervised classification of the most important layers NDVI and ASTER. Prior to investigating these two predictors further, a general overview of observed trends is provided.

In Table 7 only those classes of each predictor are listed which have received a remarkably high (or very low) number of presence records; the right column shows how much of the total land mass of Crete is covered by the respective class and therefore allows to consider the element of chance. See Appendix G for the complete frequency table.

Predictor	Class	Presences counts (in	% of total)	Presences counts (in	Crete area % of total)
Geology	'Plattenkalk', bedded limestones	17	26	**32**	16
Soil Type WRB	Calcaric Leptosol	57	45	**65**	39
	Chromic Luvisol	*0*	*1*	*0*	***10***
Soil Parent Material	Limestone	57	45	**65**	45
Soil Depth to Rock	shallow (<40cm)	83	76	**83**	63
Soil Land Use	wasteland, shrub	83	75	**83**	62
CORINE Layer 3	Natural grasslands	30	21	**36**	20
	Sclerophyllous vegetation	33	26	**28**	24
	Bare rocks	7	2	**10**	1
	Olive groves	*13*	*16*	*9*	***23***
NDVI 30 unsup. classes	class 18	26	6	**17**	4
	class 15	9	1	**14**	1
	class 6	7	4	**7**	1
	class 27	*2*	*8*	*1*	***11***
	class 29	*0*	*0*	*0*	***11***
ASTER 35 unsup. classes	class 34	**17**	5		
	class 30	**16**	11		
	class 33	**12**	9		
	class 12	**9**	4		
	class 31	**7**	2		
	class 35	**5**	2		
	class 26	*0*	***5***		

Table 7 Selected count statistics of ground variables

Note that for a site-specific analysis aiming at the most frequent combination(s) of preferred predictor classes, a canonical correspondence analysis could be carried out. Given the relatively strong concentration of presence counts in a few NDVI and ASTER classes – and in absence of sufficient ground truth –, both layers were overlaid with the CORINE predictor in order to assign preliminary class description.

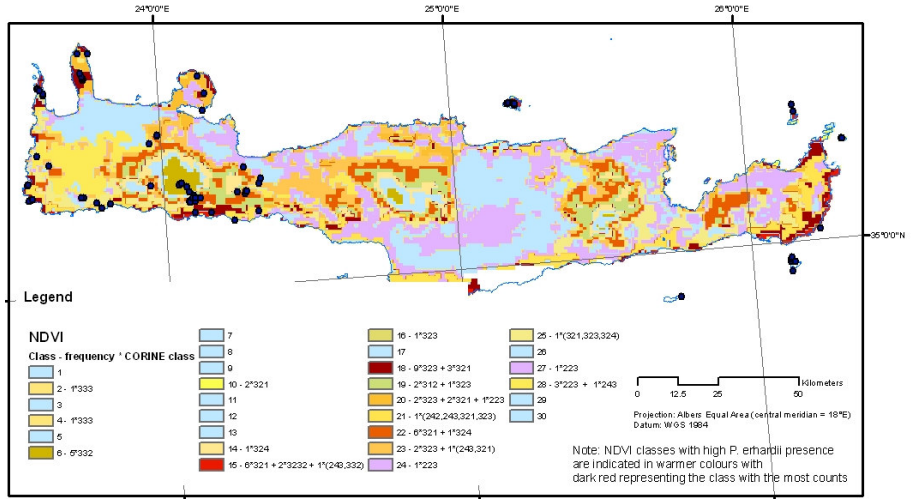


Figure 13 Overlay analysis of NDVI with CORINE predictor

As shown above, NDVI class 18 (covering only 4% of total Crete but containing 17% of all presence records) matches the CORINE class 323 ("sclerophyllous vegetation") in 9 out of 12 cases. As this class is present in both the West and far Eastern parts of Crete but hardly in the Centre, it should receive special attention in any further analysis. Note however, that this overlay is based (only) on a comparison at presence sites; a verification of this match using other points sampled from the background as well as the full consideration of the fieldwork-derived 'natural habitat polygons' instead of just the central point is recommended, but beyond the scope of this study.

It is an important result that about two-thirds of all known presence sites fall into only two CORINE classes (323 and 321), and that the same classes also receive the most counts when overlaid with the NDVI and ASTER variables. For a Pivot table showing the strength of the overlay of both NDVI and ASTER with CORINE in numeric detail see Appendix H.

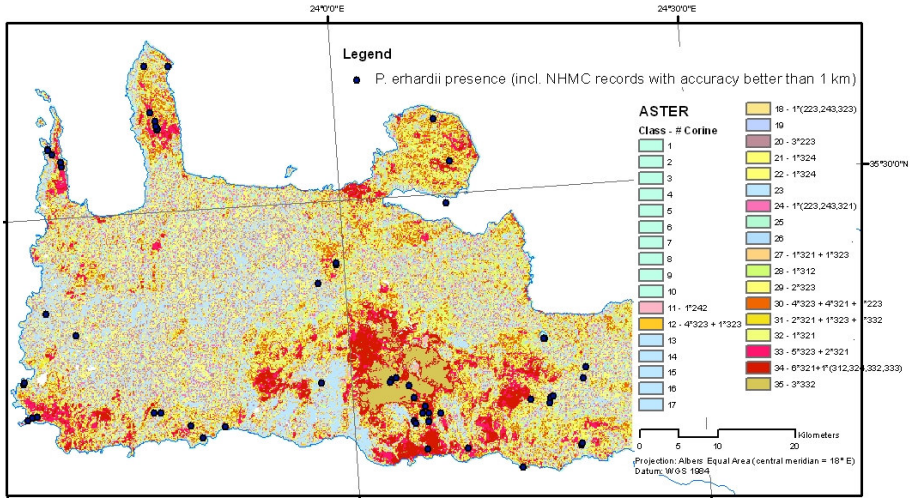


Figure 14 Overlay analysis of ASTER with CORINE predictor

Note that the ASTER-based land-cover predictor – despite it being a temporal snapshot only – is likely to more accurately show the outline of most classes at least for those with low vegetation cover. Some mismatch may thus be a consequence of the limited resolution of CORINE data rather than a true disagreement regarding surface cover. ASTER class 34 hosts the most presence sites in Western Crete and corresponds 6 out of 10 times with CORINE class 321 "natural grassland", followed by ASTER class 30 (and 33) which relate 4 out of 9 (5 out of 7) times to CORINE class 323 "sclerophyllous vegetation".

P. erhardii presence records coincide significantly with 'open areas', i.e. the most frequent NDVI class among presence sites belongs to the lowest third of all NDVI classes generated for Crete (using an ISODATA classification)..
→ refer to xls im anhang, do minitab z test (but it will fail anyway). Accept Ha.

As NDVI class numbers increase with corresponding NDVI value, the null hypothesis of research question 3 can thus technically not be rejected, as 18 > 10 (see Appendix G). The first ten classes however represent mostly coastal water and shores. In principle, the findings suggest indeed a preference for cover types with low vegetation levels.

4. Discussion

Prior to an ecological interpretation of results, the applied evaluation methods and the input data used merit a critical review in order to point out potential technical shortcomings of this research, which may have introduced distortions into the results.

4.1. Critique of Evaluation Methods

As pointed out in Chapter 2.3.3, various evaluation methods for presence-only distribution models exist (Hirzel et al., 2006), but none appears to have matured yet sufficiently to be used as a single measure or statistic summarizing model performance (Pearce and Boyce, 2006). Similarly, sources of uncertainty in presence-only distribution models and associated modelling responses are still an active topic of research (Stoms et al., 1992, Barry and Elith, 2006). The recently proposed 'continuous Boyce index' (Hirzel et al., 2006) constitutes just one out of many examples, which could potentially improve the threshold-based evaluation methods used in this study. Note however that the debate centres on threshold identification and that both TSS and the binomial test carried out in this study are known to perform if applied correctly (see chapter 2.3.3). Using threshold-independent ROC analysis on presence-only models has its limits as well, again mostly imposed by the context in which ROC is calculated. In fact, the excellent AUC results (and low p values) obtained for research question 2b) are not very meaningful at all, because the 'plot' approach caused a relative concentration of

presence points in the core ecological niche (Anderson et al., 2003). The aim however was not to identify a narrow core zone but to create a model that indicates the distribution limits. Therefore, this distribution model is most likely overfitted and suffers from high errors of omission. The rejection of hypothesis 2b) should therefore not be understood as a statement of superiority of the 'plot' approach over the 'point' approach.

The AUC achieved for research question 2a (AUC = 0.70; 0.76 with ASTER) and question 1a (AUC = 0.86) are thus the more appropriate ones to interpret. These values can be considered very good, given that presence-only data can never produce an AUC of 1 because suitable but non-tested habitat is treated as 'absence'. How close the AUC is to its potential maximum, can ultimately only be assessed if it is known how specialized the environmental niche is that *P. erhardii* occupies, because a wider niche corresponds generally with a lower AUC value (Phillips et al., 2004).

Another problem is that most evaluation methods focus on testing the performance rather than the validity of a model. For this, truly independent data – ideally from a different study area – should be used (Hirzel et al., 2006, Elith et al., 2006). When this is not possible – as in this study –, it is all the more important to validate not only modelling results but the appropriateness and accuracy of all input data as well (Corsi et al., 2000).

4.2. Critique of Species Presence Data

For a realistic predictive distribution model, species presence records must cover the full geographic and ecological extent of its known distribution. Geographically this requirement was probably met by the data used for the *Cretan* populations of *P. erhardii* (see Figure 1). If *P. erhardii*'s full tolerance range along each environmental gradient however is not captured in the occurrence data provided by NHMC, then all results of this study are likely to contain errors of omission (Kadmon et al., 2004). Here the judgement of an expert in the species ecology is required; based on the general description given in Poulakakis et al., 2005, this requirement may perhaps been also met. Finally, quality assurance of species presence data must address spatial accuracy and estimate biases. Both measures were carried out in chapter 2.1.3, the former yielding positive the latter insufficient results. All modelling results contain therefore considerable site selection bias. While acknowledging the biases and limitations discussed above, it can be concluded that the distribution modelled to answer research question 1a) reflects in essence a *reduced* version of the *fundamental* niche of *P. erhardii* on Crete. The predicted distribution exceeds the realized niche because biotic interactions, human influence

and physical barriers to migration etc. were not considered (Peterson and Holt, 2003). On the other hand, the model does not show the full fundamental niche, because presence sampling can only take place in the realized niche (Guisan and Tuiller 2005).

4.3. Critique of Environmental Predictors

Several predictor related aspects deserve attention when evaluating distribution model outcomes, because these settings are determined prior to any model run and evaluation test. Model outputs and test results are therefore *influenced* by these aspects but not directly *sensitive* to them.

Firstly, undetected collinearity may cause two correlated variables to balance each other out, displaying a low 'cumulative gain', while in reality each predictor would generate a higher 'cumulative gain' if considered individually. In this study, while Maxent's regularization function has compensated for some of this potentially harmful collinearity, an unknown amount of variable interaction – if it was present – is likely to have remained in the model. As described in chapter 2.2.1, a multi-collinearity test using the VIF method carried out before selecting the final set of predictors, could have reduced this uncertainty.

Secondly, varying resolution (cell size) among selected predictors influences model results. Therefore, care was taken that all predictors featured a cell size appropriate to the *regional scale* of the study, in order to maximize their respective potential distribution (see Table 1 for numeric details). Given the large mapping units of ESDB soil variables, the modelled distribution could be improved if data at a finer resolution were available to replace these predictors; currently they might be more useful for studies at a smaller scale. Similarly, it was observed during fieldwork and in comparison with ASTER imagery, that CORINE data were at times of poor spatial accuracy (deviations of up to 1 km) at known species occurrence sites, especially near (former) polygon borders. The observed significance of the land cover variables CORINE, ASTER and NDIV (a 'pseudo-landcover') does not contradict the findings of Thuiller et al., 2004 who report insignificance of land-cover predictors at a resolution of 50x50km. Results also conform with those of García Márquez, 2006 who found that indirect (e.g. land cover) predictors are of significance at local and regional scale, provided this predictor is of ecological importance to the species.

Thirdly, the number of classes within each predictor has an undetermined influence on model results. The distribution modelled for research question 2b illustrates this effect very well, as the 'plot' approach leads to an approximately tenfold increase in individual 'cumulative gain' of temperature and precipitation variables. As some of

the new points fall into a neighbouring cell and this cell's value tends to be significantly different given the coarse resolution of 1x1km, drastic changes in model output are logical. Any variable with smaller cell size (e.g. altitude) is less sensitive to this effect because of the finer resolution (ensuring less abrupt changes). Even less sensitive are categorical predictors, because of less frequent changes in cell values (e.g. soil) or their finer "thematic resolution" (e.g. ASTER, NDVI), which also lowers the likelihood of *drastic* changes to neighbouring cells (at the regional scale). Although distribution 2b may be overfitted, the observed change in relative variable importance is extremely valuable: if instead of 10 points a single point (or two) with the *average* of these 10 points was taken, overfitting would have been reduced while – most importantly – a more realistic individual contribution to the 'cumulative gain' by temperature and precipitation predictors would have been achieved and the overall model gain increased! Either way, an appropriate number of classes per predictor (fine enough to identify the core zone but large enough to allow the model to reach out to the distribution limits) is essential and an issue that has not been receiving much attention in the species distribution literature.

4.4. Interpretation of Results

All hypotheses have been answered already in the chapter Results and critically evaluated in the previous paragraphs. This chapter attempts an ecological interpretation and discusses the results in light of the current knowledge about the species biogeography.

So what can the environmental predictors found to strongly determine the current distribution of *P. erhardii* according to model (1a), tell us about the preferred habitat and surface (vegetation) cover? The overlay of NDVI and ASTER with CORINE data shows a clear association for the most important classes. All three layers suggest that natural grassland, sclerophyllous vegetation and bare rocks are the surface type of choice in general. Together they captured 74% of all occurrence records. Interestingly, a comparison of the annual profile of NDVI class 18 "sclerophyllous vegetation" reveals a strong similarity with NDVI class 15 (which received the second most presence counts) and matches the CORINE layer 323 "natural grassland" 6 out of 10 times. This points towards a specific vegetation type peaking in mid-June in terms of NDVI (Figure 4).

Also, with NDVI found to be the most important predictor, the vast non-preferred area in the island's centre composed of class 29 (unspecified) and 24 (CORINE 223: "olive groves") may indicate a surface cover type that constrains P. erhardii's

dispersal from the West to the East or vice versa (Figure 13). Ultimately, in-situ observations are required to relate the NDVI classes to specific vegetation communities (Box et al., 1989). Yet, this indirect predictor holds promise as it accounts for outbalancing effects of direct predictors. The heights of the Levká Óri mountain for instance, receive a maximum of precipitation, but little of this can be used by the flora, as the calcareous, rocky ground is so porous that only plants with low water requirements can survive (Fielding and Turland, 2005).

Other predictors are in line with previous knowledge of *P. erhardii's* habitat preferences (Poulakakis et al., 2005). Ideal locations are characterized by bedded limestone or carbonate rocks, shallow soils (< 40cm) with limestone as parent material and preferably calcaric leptosoils. Although the influence of climate variables, in particular cloud cover and actual evapotranspiration seems to be fairly strong on the regional scale (see chapter 3.2 for a short discussion), they shall not be discussed here, as the focus is on preferred ground parameters.

The degree of stoniness did not seem to be of significance, however this could be due to both presence of alternative hiding options (shrub) or the very general classification of input data. Nor was a preference for a specific exposure (aspect) was obtained from the model, but this is likely due to inadequate input data, i.e. no appropriate sampling design.

Although this research has yielded a distribution model with a good fit to the sample data, a good performance on test data and a significant predictor holding the potential for a reasonable ecological explanation, there may be a variety of other – competing or complementary – explanatory factors beyond mere environmental conditions for the observed distribution of *P.erhardii* on Crete. These include direct human impacts like pesticide use and biotic interactions (e.g. competing species, trophic chain positions). In fact, most island populations of P. erhardii are the only lizard species on their island (Engelmann, 1986), which may constitute evidence for the species vulnerability by competitors rather than by predators. On the other hand, as many populations occur in isolated localities rather than throughout large continuous areas (Poulakakis et al., 2005), there is also a possibility that the species' biogeography has been influenced by rivals outcompeting *P. erhardii* for scarce resources. Recent advances in species distribution models aim at incorporating more of these factors, e.g. species migration, population dynamics, biotic interactions and community ecology and at multiple scales (Guisan and Thuiller, 2005).

Another potentially very important explanatory factor are tectonic shifts and especially historic sea-level fluctuations. Significant local fragmentation of *P. erhardii* populations may have occurred during Pliocene flooding of much of the island (Creutzburg, 1963). It was beyond the scope of this project however to

consider these factors.

5. Synthesis

5.1. Conclusions

The main objective of this study was to create a species distribution model able to 'explain' the observed curious geographic distribution of *P. erhardii* on Crete, while using only a range of *environmental* predictors together with provided species occurrence data.

To a certain extent this objective has been achieved. Using Maxent as the statistical model, and by 'enhancing' existing presence records by replacing them with 'representative natural habitat polygons' in the immediate vicinity, the selected predictors allowed a distribution to be modelled that fits both the dominant occurrence in Western Crete and the marginal occurrence on Eastern islets. The regularized training gain was 1.86 with a bootstrapped AUC of > 0.95.

The most useful predictors were found to be primarily NDVI at a 1x1 km scale, followed by cloud cover and actual evapotranspiration as climate variables, and CORINE land cover and altitude as ground related predictors. Soil variables however might have suffered from their high degree of generalization.

Special emphasis was placed on evaluating the usefulness of ASTER imagery for modelling species distribution at this scale. The result is encouraging as the (unsupervised) classified land cover predictor turned out to be second in significance only to NDVI. Although it represents only a temporal snapshot, the ASTER variable showed a reliable association with temporally more stable land cover predictors such as multi-annual NDVI and CORINE. The extensive pre-processing requirements associated with using several ASTER granules, may inhibit its widespread use in studies at this scale.

Replacing single occurrence points with a set of 10 points in order to capture the local environmental conditions more fully, increased the training gain by almost 20%. This test also illustrates the significance of an appropriate 'thematic resolution'

of categorical variables as well as the considerable dependency of continuous predictors on spatial resolution. Although these 10 points covered only ~ 3000m^2 on average, the individual 'cumulative gain' of temperature and precipitation predictors grew almost tenfold. As this approach appears to allow continuous predictors to fully exploit their discriminative power, it merits further testing in species distribution modelling.

Finally, a preliminary descriptive analysis on which surface cover and vegetation types P. erhardii seems to favour the most was carried out. It was found that three specific NDVI classes contain 38% of all occurrence sites while covering only 6% of Crete. Using overlap analysis, these classes were found to associate mostly with the CORINE classes natural grassland, sclerophyllous vegetation and bare rocks.

Pending an evaluation of findings using a truly independent dataset, the results of this research suggest that current environmental conditions are the primary explanatory factor for the observed geographic distribution of P. erhardii on Crete.

5.2. Recommendations

An array of both additional data preparation and analysis steps could be undertaken to improve the results obtained in this study.

An alternative single-species modelling technique could be employed and results compared, e.g. GARP (Stockwell and Peters, 1999) or BRT (Friedman et al., 2000).

A canonical ordination and correspondence analysis (Jongman et al., 2005) on fieldwork-derived macrohabitat data could be employed to test these vegetation data as surrogates in a community model such as GDM (Ferrier et al., 2002) to model potential distribution of *P. erhardii.*

Additional ground truth data for a solid supervised ASTER land cover classification should be collected and used to expand the current ASTER mosaic across all of Crete; verify current predictions in situ.

The current interpolated low resolution cloud cover (and evapotranspiration) dataset could be replaced with higher resolution data to test again the apparent importance of cloud cover as main West-East gradient and its derived significance in explaining *P.*

erhardii's dominant occurrence in the West.

Slope and aspect predictors could be replaced with solar radiation data: the removal of these indirect predictors should reduce the likelihood of error propagation (van Neil et al., 2004); irradiation is likely an essential biological determinant for *P. erhardii* as it controls body heat and thus available energy; both irradiation data and promising enhancement methods (Kumar et al., 1997) are available.

The significance of the NDVI predictor could be further investigated (without additional fieldwork) by deriving new variables from it like 'standard deviation of NDVI' or 'co-efficient of variation of NDVI' as done in Omolo, 2006.

Alternative indices to NDVI, such as the Distance Vegetation Index (DVI) and the related Perpendicular Vegetation Index (PVI) could be calculated. Both have been found to perform slightly better than NDVI for low vegetation levels (McCloy, 2006). As P. erhardii appears to have a preference for *phrygana* vegetation cover, these indices hold the potential to further narrow down preferred surface cover using remote sensing data.

Since both distance and ratio indices (including NDVI) are affected by soil reflectance effects in case that NIR reflection for soil and vegetation is similar; the Transformed Soil Adjusted Vegetation Index (TSAVI) (Baret et al., 1989), which incorporates soil line parameters more flexibly might therefore be another promising alternative (additionally) to NDVI.

6. References

Abrams, M. (2000) The Advanced Space-borne Thermal Emission and Reflection radiometer (ASTER): data products for the high spatial resolution imager on NASA's Terra platform. *International Journal of Remote Sensing,* 21**,** 847-859.

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

Anderson, R. P., Lew, D. & Peterson, A. T. (2003) Evaluating predictive models of species' distributions: criteria for selecting optimal models. *Ecological Modelling,* 162**,** 211-232.

Anderson, R. P. & Martinez-Meyer, E. (2004) Modelling species' geographic distributions for preliminary conservation assessments: an implementation with the spiny pocket mice (Heteromys) of Ecuador. *Biological Conservation,* 116**,** 167-179.

Arnold, E. N. (2002) *A Field Guide to the Reptiles and Amphibians of Britain and Europe,* London, Collins.

Austin, M. (2007) Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modelling,* 200**,** 1-19.

Austin, M. P. (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling,* 157**,** 101-118.

Austin, M. P. & Smith, T. M. (1989) A new model for the continuum concept. *Plant Ecology,* V83**,** 35-47.

Baret, F., Guyot, G. & Major, D. J. (1989) TSAVI: A Vegetation Index Which Minimizes Soil Brightness Effects On LAI And APAR Estimation. Proceedings of the *12th Canadian Symposium on Remote Sensing*, Vancouver, Canada, 1355-1358.

Barry, S. & Elith, J. (2006) Error and uncertainty in habitat models. *Journal of Applied Ecology,* 43**,** 413-423.

Box, E. O., Holben, B. N. & Kalb, V. (1989) Accuracy of the AVHRR vegetation index as a predictor of biomass, primary productivity and net CO2 flux. *Vegetation,* 80**,** 71-89.

Brauner, N. & Shacham, M. (1998) Role of range and precision of the independent variable in regression of data. *AIChE Journal,* 44**,** 603-611.

Brown, J. H. & Lomolino, M. V. (1998) *Biogeography,* Sunderland, Massachusetts.

Buckland, S. T., Burnham, K. P. & Augustin, N. H. (1997) Model Selection: An

Integral Part of Inference. *Biometrics,* 53**,** 603-618.

Buckley, Y., Anderson, S., Catterall, C., Corlett, R., Engel, T., Gosper, C., Nathan, R., Richardson, D., Setter, M., Spiegel, O., Vivian-Smith, G., Voigt, F., Weir, J. & Westcott, D. (2006) Management of plant invasions mediated by frugivore interactions. *Journal of Applied Ecology,* 43**,** 848-857.

Burnam, K. P. & Anderson, D. R. (1998) *Model Selection and Inference. A practical Information-Theoretic Approach,* New York, Springer Verlag.

Busby, J. R. (1991) BIOCLIM - A Bioclimatic Analysis and Prediction System. IN Margules, C. R. & Austin, M. P. (Eds.) *Nature Conservation: Cost Effective Biological Surveys and Data Analysis.* Canberra, CSIRO, 64-68.

Cantor, S. B., Sun, C. C., Tortolero-Luna, G., Richards-Kortum, R. & Follen, M. (1999) A Comparison of C/B Ratios from Studies Using Receiver Operating Characteristic Curve Analysis. *Journal of Clinical Epidemiology,* 52**,** 885-892.

Carpenter, G., Gillison, A. N. & Winter, J. (1993) DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity and Conservation,* V2**,** 667-680.

Catterall, C., Westcott, D. A., Kanowski, J. & Dennis, A. J. (2003) Introduction. Proceedings of the *Animal-Plant Interactions in Rainforest Conservation and Restoration, Workshop 11 November 2003*, Cairns, 1-4.

CNES: *SPOT VEGETATION ten daily synthesis archive*, SPOT IMAGE via Vito Belgium. Online Database Access: http://free.vgt.vito.be/ [retrieved: 2006-09-02].

Cohen, J. (1960) A coefficient of agreement of nominal scales. *Educational Psychology Measurements,* 20**,** 37-46.

Congalton, R. (1991) A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment,* 37**,** 35-46.

Corsi, F., Skidmore, A. & de Leeuw, J. (Eds.) (2000) *Modelling species distribution with GIS,* New York, Columbia University Press.

Creutzburg, N. (1963) Paleogeographic evolution of Crete from Miocene till our days. *Cretan Annals,* 15/16**,** 336-342.

De Leeuw, J., Ottichilo, W. K., Toxopeus, A. G. & Prins, H. H. T. (2002) Application of remote sensing and geographic information systems in wildlife mapping and modelling. IN Skidmore, A. K. (Ed.) *Environmental modelling with GIS and remote sensing.* London, Taylor & Francis, 121-145.

Dudík, M., Phillips, S. J. & Schapire, R. E. (2004) Performance guarantees for regularized maximum entropy density estimation. Proceedings of the.

EC-DGJRC, European Commission - DG Joint Research Centre - Institute for Environment and Sustainability. Online Database Access: http://eusoils.jrc.it/ESDB_Archive/ESDB_Data_1k_raster_distribution/list_of_ESDBv2_1K_rasters.cfm [retrieved: 2006-11-09].

EEA (2000) CORINE Land Cover - Part 1: Methodology. CORINE Land Cover - Part 1: Methodology. Available online: http://reports.eea.europa.eu/COR0-landcover/en [retrieved: 2006-09-20].

EEA: *Corine land cover 2000 seamless vector database (CLC2000)*, European Environment Agency. Online Database Access: http://dataservice.eea.europa.eu/dataservice/metadetails.asp?id=950 [retrieved: 2006-09-15].

Elith, J., Graham, C. H., Anderson, R. P., Dudik, 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., 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.

Engelmann, W.-E. (1986) *Lurche und Kriechtiere Europas: beobachten und bestimmen*, Deutscher Taschenbuch Verlag dtv.

ESBN, European Soil Bureau Network and the European Commission. Online Database Access: http://eusoils.jrc.it/ESDB_Archive/ESDB_Data_Distribution/ESDB_data.html [retrieved: 2006-12-19].

ESBN, European Soil Bureau Network and the European Commission. Online Database Access: http://eusoils.jrc.it/ESDB_Archive/ESDBv2/fr_intro.htm [retrieved: 2007-01-26].

ESRI Inc. (2005a) Arc Toolbox: How Spline (3D Analyst) Works. *ArcGIS Desktop Help. Software Manual.* Redlands, California, U.S.

ESRI Inc. (2005b) How Inverse Distance Weighted (IDW) interpolation works. *ArcGIS Desktop Help. Software Manual.* Redlands, California, U.S.

Ferrier, S., Drielsma, M., Manion, G. & Watson, G. (2002) Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. II. Community-level modelling. *Biodiversity and Conservation,* 11**,** 2309-2338.

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.

Fielding, J. & Turland, N. (2005) *Flowers of Crete,* Richmond, Royal Botanical Gardens.

Friedman, J., Hastie, T. & Tibshirani, R. (2000) Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *Annals of Statistics,* 28**,** 337-407.

Friedman, J. H. (1991) Multivariate Adaptive Regression Splines. *The Annals of Statistics,* 19**,** 1-67.

García Márquez, J. R. (2006) 'Multiscale assessment of the potential distribution of two herpetofaunal species', MSc Thesis, NRM, ITC.

Graham, C. H. & Hijmans, R. J. (2006) A comparison of methods for mapping species ranges and species richness. *Global Ecology and Biogeography,* 15**,** 578-587.

Guisan, A. & Hofer, U. (2003) Predicting reptile distributions at the mesoscale: relation to climate and topography. *Journal of Biogeography,* 30**,** 1233-

1243.
Guisan, A., Theurillat, J.-P. & Kienast, F. (1998) Predicting the Potential Distribution of Plant Species in an Alpine Environment *Journal of Vegetation Science,* 9**,** 65-74.
Guisan, A. & Thuiller, W. (2005) Predictive species distribution: offering more than simple habitat models. *Ecology Letters,* 8**,** 993-1009.
Guisan, A. & Zimmermann, N. E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling,* 135**,** 147-186.
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.
Hastie, T. & Tibshirani, R. (1990) *Generalized Additive Models,* London, Chapman & Hall.
Henle, K., Lindenmayer, D. B., Margules, C. R., Saunders, D. A. & Wissel, C. (2004) Species survival in fragmented landscapes: where are we now? *Biodiversity and Conservation,* 13**,** 1-8.
Hernandez, P. A., Graham, C. H., Master, L. L. & Albert, D. L. (2006) The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography,* 29**,** 773-785.
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., Cameron, S. E., Parra, J. L., Jones, P. G. & Jarvis, A.: *Worldclim version 1.4*. Online Database Access: http://www.worldclim.org/current.htm [retrieved: 2006-12-15].
Hirzel, A. & Guisan, A. (2002) Which is the optimal sampling strategy for habitat suitability modelling. *Ecological Modelling,* 157**,** 331-341.
Hirzel, A. H., Hausser, J., Chessel, D. & Perrin, N. (2002) Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data? *Ecology,* 83**,** 2027-2036.
Hirzel, A. H., Le Lay, G., Helfer, V., Randin, C. & Guisan, A. (2006) Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling,* In Press, Corrected Proof.
Hutchinson, G. E. (1957) Concluding remarks. Proceedings of the *Cold Spring Harbor Symposia on Quantitative Biology*, 415-427.
IPCC (2007) Fourth Assessment Report: Climate Change 2007: The Physical Science Basis. Summary for Policy-Makers. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Approved at the 10th Session of Working Group I of the IPCC, Paris, February 2007. Available online: http://www.ipcc.ch/SPM2feb07.pdf [retrieved: 2007-02-05].
Jaberg, C. & Guisan, A. (2001) Modelling the distribution of bats in relation to landscape structure in a temperate mountain environment. *Journal of Applied Ecology,* 38**,** 1169-1181.
Jensen, J. (2005) *Introductory digital image processing - a remote sensing perspective,* New Jersey, Prentice Hall.

Jongman, R. H. G., Ter Braak, C. J. F. & van Tongeren, O. F. R. (2005) *Data analysis in community and landscape ecology,* Cambridge, University Press.

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

Kéry, M. (2002) Infering the absence of a species: a case study of snakes. *Journal of Wildlife Management,* 66**,** 330-338.

Kumar, L., Skidmore, A. & Knowles, E. (1997) Modelling topographic variation in solar radiation in a GIS environment. *International Journal of Geographical Information Science,* 11**,** 475-497.

Legakis, A. & Krypriotakis, Z. (1994) A biogeographical analysis of the island of Crete, Greece. *Journal of Biogeography,* 21**,** 441-445.

Leica Geosystems (2003) Resampling methods. *ERDAS Field Guide. Software Manual.* 7th ed. Atlanta, Georgia, U.S.

Lévêque, C. & Mounolou, J.-C. (2001) *Biodiversité, dynamique biologique et conservation,* Paris, Dunod.

Lewinski, S. (2005) Land use classification of ASTER image - Legionowo test site. Proceedings of the *25th EARSeL Symposium*, Porto, Portugal.

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.

Lymberakis, P. (2006-10-03) Confirmed occurrence sites of *Podarcis erhardii* on Crete. Personal communication to Herkt, M.; Heraklion.

Mackenzie, D. I. & Royle, J. A. (2005) Designing occupancy studies: general advice and allocating survey effort. *Journal of Applied Ecology,* 42**,** 1105-1114.

Magurran, A. E. (1988) *Ecological Diversity and Its Measurement,* Princeton, Princeton University Press.

Magurran, A. E. (2005) Biological diversity. *Current Biology,* 15**,** 116-118.

Marcal, A. R. S., Borges, J. S., Gomes, J. A. & Pinto do Costa, J. F. (2005) Land cover update by supervised classification of segmented ASTER images. *International Journal of Remote Sensing,* 26**,** 1347-1362.

Margules, C. & Austin, M. P. (1994) Biological models for monitoring species decline: the construction and use of data bases. *Philosophical Transactions of the Royal Society of London,* 344**,** 69-75.

Margules, C. & Pressey, R. L. (2000) Systematic conservation planning. *Nature,* 405**,** 243-253.

McCloy, K. (2006) *Resource Mgmt Info Systems: Remote sensing, GIS and modelling*, Taylor & Francis.

McCullagh, P. & Nelder, J. A. (1989) *Generalized Linear Models,* New York, Chapman & Hall.

Meentemeyer, R., Rizzo, D., Mark, W. & Lotz, E. (2004) Mapping the risk of establishment and spread of sudden oak death in California. *Forest Ecology and Management,* 200**,** 195-214.

MiniTab Inc. (2005) MINITAB (R) Release 14.2 Statistical Software, Available:

http://www.minitab.com [retrieved: 2007-01-15].
Moilanen, A. & Wintle, B. A. (2006) Uncertainty analysis favours selection of spatially aggregated reserve networks. *Biological Conservation,* 129**,** 427-434.
Murwira, A., Skidmore, A. K. p., Prins, H. H. T. p. & Huizing, H. G. J. p. (2003) 'Scale matters ! : a new approach to quantify spatial heterogeneity for predicting the distribution of wildlife', PhD Thesis, ITC and Wageningen University.
NASA: *Earth Observing System Data Gateway*, Land Processes Distributed Active Archive Center. Online Database Access: http://edcimswww.cr.usgs.gov/pub/imswelcome/ [retrieved: 2006-11-10].
NHMC (2006) *Podarcis erhardii observation records in Greece*, National History Museum of Crete, Unpublished Database.
NIEHS (1996) *Quick Reference: Global Climate Database*. Available online: http://webgis.wr.usgs.gov/globalgis/metadata_qr/climate.htm [retrieved: 2007-01-15].
NIEHS: *Global Climate Database*, USGS, National Institute of Public Health and Environmental Protection. Online Database Access: offline content [retrieved: 2006-11-30].
Omolo, D. P. (2006) 'Biodiversity Patterns in Changing Mediterranean Landscape: a modelling perspective', MSc Thesis, GEM programme, ITC.
Pafilis, P., Foufopoulos, J., Poulakakis, N., Lymberakis, P. & Valakos, E. (2007) Digestive performance in five Mediterranean lizard species: effects of temperature and insularity. *Journal of Comparative Physiology,* 177**,** 49-60.
Pearce, J. & Boyce, M. (2006) Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology,* 43**,** 405-412.
Peterson, A. T. & Holt, R. D. (2003) Niche differentiation in Mexican birds: using point occurences to detect ecological innovation. *Ecological Letters,* 6**,** 774-782.
Peterson, A. T. & Robins, C. R. (2003) Using Ecological-Niche Modeling to Predict Barred Owl Invasions with Implications for Spotted Owl Conservation. *Conservation Biology,* 17**,** 1161-1165.
Phillips, S. J., Anderson, R. P. & Schapire, R. E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling,* 190**,** 231-259.
Phillips, S. J., Dudik, M. & Schapire, R. E. (2004) A Maximum Entropy Approach to Species Distribution Modeling. Proceedings of the *21st International Conference on Machine Learning*, Banff, Canada.
Poulakakis, N., Lymberakis, P., Valakos, E., Zouros, E. & Mylonas, M. (2005) Phylogenetic relationships and biogeography of Podarcis species from the Balkan Peninsula, by bayesian and maximum likelihood analyses of mitochondrial DNA sequences. *Molecular Phylogenetics and Evolution,* 37**,** 845-857.
Pressey, R. L., Humphries, C. J., Margules, C. R., Van-Wright, R. I. & Williams, P. H. (1993) Beyond opportunism: Key principles for systematic reserve selection. *Trends in Ecology & Evolution,* 8**,** 124-128.

Raffaelli, D. (2004) Getting to grips with biological diversity measurement. *Journal of Biogeography,* 31**,** 2043-2044.

Rebelo, A. G. (1994) Iterative selection procedures: centres of endemism and optimal placement of reserves. *Strelitzia,* 1**,** 231-257.

Rounsevell, M. D. A., Reginster, I., Araujo, M. B., Carter, T. R., Dendoncker, N., Ewert, F., House, J. I., Kankaanpaa, S., Leemans, R., Metzger, M. J., Schmit, C., Smith, P. & Tuck, G. (2006) A coherent set of future land use change scenarios for Europe. *Agriculture, Ecosystems & Environment,* 114**,** 57-68.

Sarkar, S. & Margules, C. (2002) Operationalizing biodiversity for conservation planning. *Journal of Biosciences,* 27**,** 299-308.

Schröder, B. (2004) ROC Plotting and AUC Calculation Transferability Test v 1.3, Available: http://vg00.met.vgwort.de/na/1d696a44341266aa381a?l=http://brandenburg.geoecology.uni-potsdam.de/users/schroeder/download/roc_auc_dec06.zip [retrieved: 2007-02-02].

Segurado, P., Araujó, M. & Kunin, W. E. (2006) Consequences of spatial autocorrelation on niche-based models. *Journal of Applied Ecology,* 43**,** 433-444.

Stefanov, W. & Netzband, M. (2005) Assessment of ASTER land cover and MODIS NDVI data at multiple scales for ecological characterization of an arid urban center. *Remote Sensing of Environment,* 99**,** 31-43.

Stockwell, D. R. B. & Peters, D. (1999) The GARP modelling system: problems and solution to automated spatial prediction. *International Journal of Geographical Information Science,* 13**,** 143-158.

Stockwell, D. R. B. & Peterson, A. T. (2002) Effects of sample size on accuracy of species distribution models. *Ecological Modelling,* 148**,** 1-13.

Stoms, D. M., Davis, F. W. & Cogan, C. B. (1992) Sensitivity of wildlife habitat models to uncertainties in GIS data *Photogrammetric Engineering and Remote Sensing,* 58**,** 843-850.

Thuiller, W., Araujo, M. B. & Lavorel, S. (2004) Do we need land-cover data to model species distributions in Europe. *Journal of Biogeography,* 31**,** 353-361.

U.S.G.S.: *Shuttle Radar Topography Mission (SRTM) 3 arc second SRTM Format*, Earth Resources Observation & Science (EROS), U.S.G.S. Online Database Access: http://edcsns17.cr.usgs.gov/srtmbil/area6coverage.html [retrieved: 2006-09-07].

U.S.G.S. (2006) *Shuttle Radar Topography Mission (SRTM) "Finished" 3-arc second SRTM Format*. Available online: http://edc.usgs.gov/products/elevation/srtmbil.html#description [retrieved: 2007-01-15].

van Neil, K. P., Laffan, S. W. & Lees, B. G. (2004) Effect of error in the DEM on environmental variables for predictive vegetation modelling. *Journal of Vegetation Science,* 15**,** 747-756.

van Teeffelen, A. J. A., Cabeza, M. & Moilanen, A. (2006) Connectivity,

probabilities and persistence: comparing reserve selection strategies. *Biodiversity and Conservation,* 15**,** 899-919.

Verbyla, D. L. & Litvaitis, J. A. (1989) Resampling methods for evaluating classification accuracy of wildlife habitat models. *Environmental Management,* 13**,** 783-787.

Vermote, E. F., Tanre, D., Deuze, J. L., Herman, M. & Morcette, J.-J. (1997) Second Simulation of the Satellite Signal in the Solar Spectrum,6S: an overview. *Geoscience and Remote Sensing,* 35**,** 675-686.

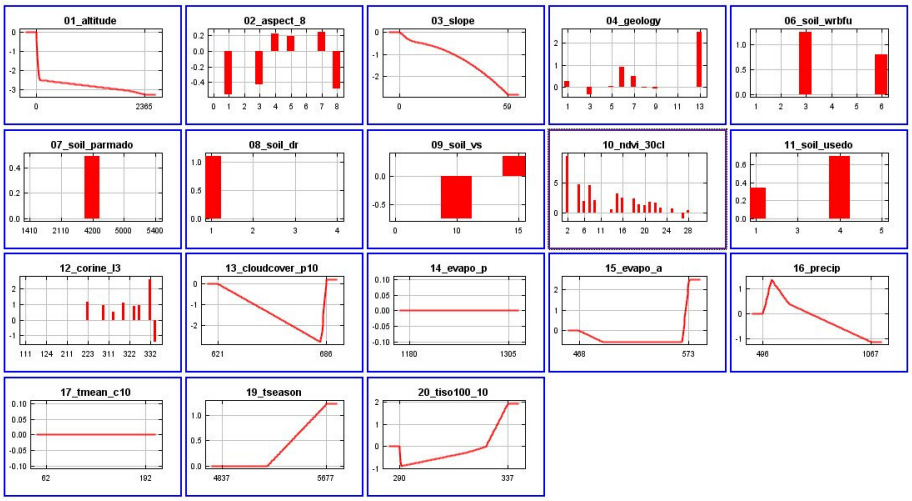
Wageningen University (1986) *Soil Types at Crete (Kriti). Data identical to EEC Soil Map*, Wageningen University, Wageningen.

Wang, X. & Zhang, S. (2006) Evaluation of Land Cover Classification Effectiveness for the Queer Mountains, China, using ASTER satellite data. Proceedings of the *Advanced Technology in the Environmental Field* Lancarote, Canary Islands, Spain.

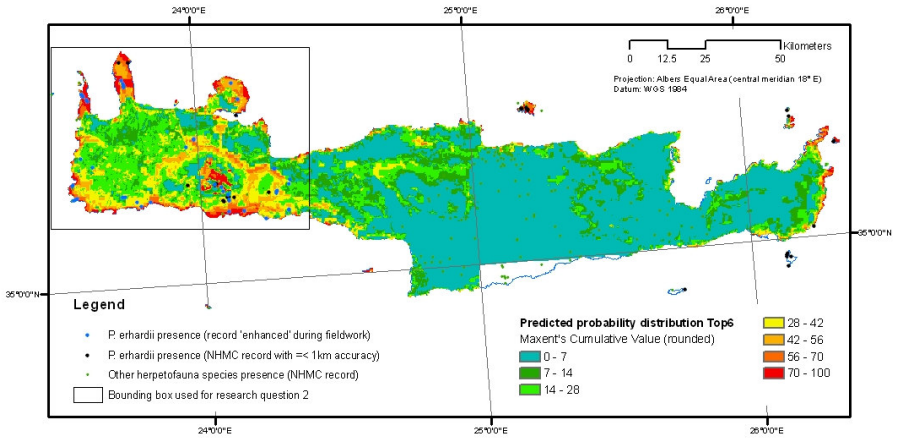
Wikipedia contributors (2007) *Crete*. Available online: http://en.wikipedia.org/w/index.php?title=Crete&oldid=102409437 [retrieved: 2007-01-24].

7. Appendices

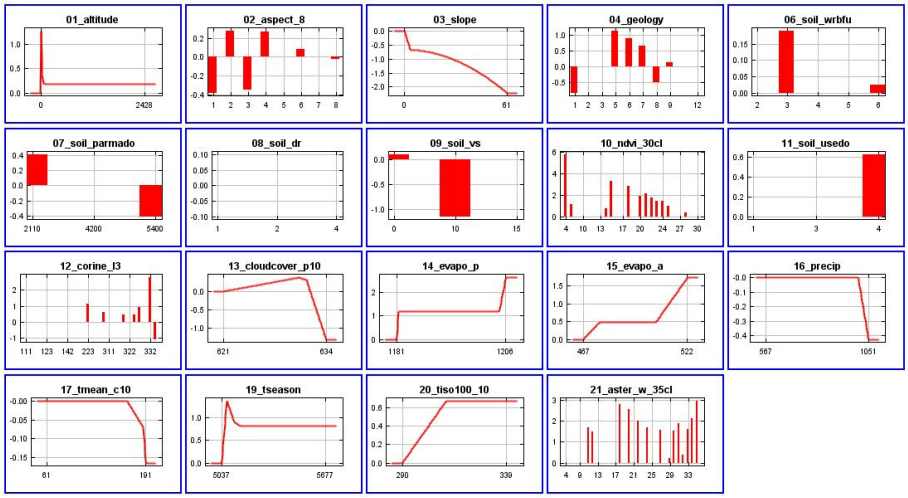
Appendix A Q1A: Response curves of predictors for distribution across Crete reflecting modelled range preferences of *P. erhardii* in ecological space



Appendix B Probability Distribution across Crete using top six predictors only (input: all qualified occurrence sites)



Appendix C Q2A: Response curves of predictors for distribution in Western Crete reflecting modelled range preferences of *P. erhardii* in ecological space (using only fieldwork-'enhanced' presence records)

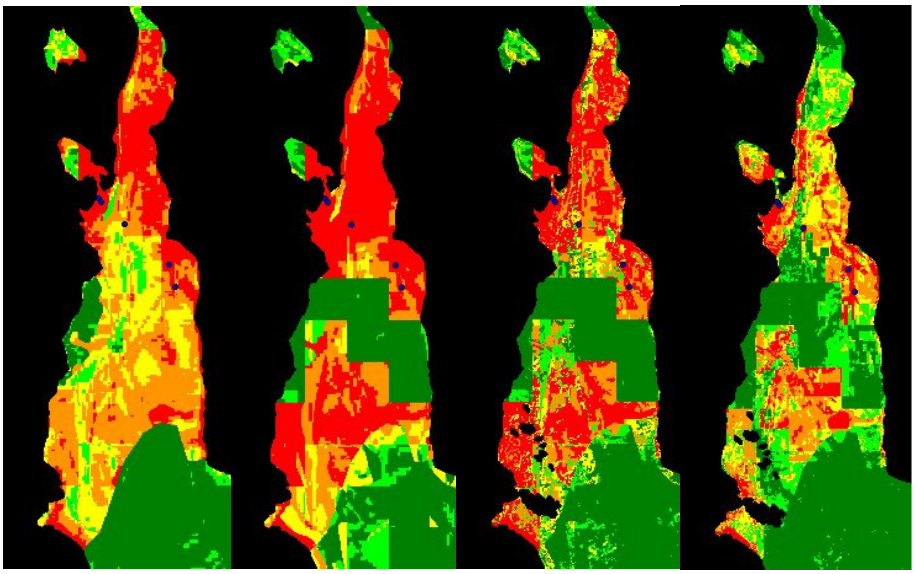


Appendix D Correlation matrix of continuous predictors

	altitude	slope	cloud c.	eva_pot	eva_act	precip	tmean	tmin	tseason	tiso
altitude	1.00									
slope	0.36	1.00								
cloudcover	-0.35	-0.33	1.00							
evapo_pot	-0.42	-0.40	0.92	1.00						
evapo_act	-0.11	-0.18	0.84	0.74	1.00					
precip	0.85	0.46	-0.49	-0.61	-0.08	1.00				
tmean	-0.99	-0.34	0.38	0.42	0.11	-0.84	1.00			
tmin	-0.98	-0.37	0.42	0.47	0.14	-0.87	1.00	1.00		
tseason	0.80	0.48	-0.77	-0.81	-0.53	0.82	-0.81	-0.85	1.00	
tiso	0.61	0.46	-0.56	-0.64	-0.10	0.89	-0.61	-0.66	0.74	1.00

Cell values represent the r^2; values of 0 indicate no collinearity, values of +1 and -1 represent perfect correlation; calculations were performed in ArcMap 9.1 based on full predictor extent across Crete

Appendix E Comparison of previously modelled probability distributions



Q1A: using all predictors except the ASTER-based land-cover variable; both fieldwork-'enhanced' presences & additional NHMC occurrence points with 1km accuracy or better (total Crete)

Q2A: using all predictors except the ASTER-based land-cover variable and only fieldwork-'enhanced' occurrence points (Western Crete)

Q2A: using all predictors and the ASTER-based land-cover variable and only fieldwork-'enhanced' occurrence points (Western Crete)

Q2B: using all predictors and the ASTER-based land-cover variable and only fieldwork-'enhanced' occurrence 'plots' (Western Crete)

Appendix F Evaluation of distribution for Western Crete *with* ASTER (and 'plots' of 10 points replacing former single occurrence point)

Model performance							Excel						ROC-PLOT						Maxent	
	binomial test						threshold$_{opt}$ = 46.76						threshold-independent (bootstrapped)							
	(presence/total)		Ø cum val		*st.dev*		**TSS**		*st.dev*	95% CI		Kappa	**AUC**		95% CI				AUC	
	train	***test***	*train*	***test***	*train*	*test*	*train*	***test***		*lower*	*z value*	*test*	***test***	*SE*	*lower*	*z value*	***p value***		*train*	*test*
run-1	(310/337)	(99/113)	79.5	**74.1**	19.2	21.7	0.84	**0.80**	0.03	0.80	283.5	0.61	**0.97**	0.01	0.96	85.35	**< 0.0001**		0.977	0.968
run-2	(307/342)	(96/108)	82.1	**77.2**	19.6	19.5	0.82	**0.81**	0.03	0.81	280.6	0.61	**0.97**	0.00	0.97	106.86	**< 0.0001**		0.978	0.973
run-3	(299/336)	(107/114)	80.3	**79.5**	19.5	18.8	0.81	**0.86**	0.02	0.86	459.1	0.65	**0.98**	0.00	0.97	113.85	**< 0.0001**		0.978	0.978
run-4	(315/337)	(101/113)	80.8	**76.8**	18.5	20.7	0.85	**0.81**	0.03	0.81	287.0	0.62	**0.97**	0.01	0.96	84.43	**< 0.0001**		0.978	0.970
run-5	(296/337)	(92/113)	79.7	**74.6**	20.1	23.5	0.80	**0.73**	0.04	0.72	194.0	0.58	**0.96**	0.01	0.95	68.80	**< 0.0001**		0.976	0.963
mean	(305/338)	(99/112)	80.5	**76.4**	19.4	20.8	0.82	**0.80**				0.61	**0.97**		0.96				0.977	0.970
b'ground	(90 / 1112)		13.6		18.9		prevalence = (0.23 / 0.09)												different b'g	

Details of significance tests above

Test	one-tailed z/t-test	one-tailed t-test at 95% CI	AUC tested with one-tailed z-test at 95% CI
H_0	Ø cum val <= 13.6	$TSS_{average}$ <= 0	AUC <= 0.5
H_A	Ø cum val > 13.6	$TSS_{average}$ > 0	AUC > 0.5

results for **binomial** test of **mean** Ø cum. value:
mean SE = 1.06 (train), 1.97 (test); lower 95% CI = 78.76 (train), 73.17 (test); z value =63.40 (train), 31.95 (test); p value = 0.000 (train), 0.000 (test)

Appendix G Presence counts per predictor class for a descriptive analysis of surface cover preferences of *P. erhardii*

04_GEOLOGY	Western C.		all Crete	
Class name	% area	% pres	% pres	% area
Plattenkalk', bedded limestones	26	**17**	**32**	16
Carbonate rocks of Tripolis zone nappe	8	17	15	15
Neogene rocks	6	11	14	10
Phyllites - Quartzites	22	22	14	12
Neogene rocks	12	15	11	19
Quaternary rocks	12	7	7	11
Carbonate rocks of Trypali zone nappe	12	11	7	4
Ophiolite complex: of allochthonous tectonic units	0	0	1	1
Carbonate rocks of Pindos zone nappe	1	0	0	3
Carbonate rocks of allochthonous tectonic units	0	0	0	0
Flysch - schists of allochthonous tectonic units	0	0	0	1
Flysch of the Pindos zone nappe	1	0	0	4
Flysch of the Tripolis zone nappe	0	0	0	3
06_SOIL_WRBFU				
Calcaric Leptosol	45	**57**	**65**	39
Calcaric Regosol	22	17	19	25
Eutric Leptosol	29	26	16	17
Eutric Cambisol	0	0	0	1
Calcaric Fluvisol	3	0	0	7
Chromic Luvisol	1	0	0	**10**
07_SOL_PARMADO				
limestone	45	**57**	**65**	45
acid regional metamorphic rocks	29	26	17	17
fluvial clays, silts and loams	25	17	17	32
flysch	0	0	0	1
unconsolidated deposits (alluvium, weathering resi	0	0	0	5
08_SOIL_DR				
shallow (<40cm)	76	**83**	**83**	63
moderate (40 - 80 cm)	22	17	17	25
deep (80 - 120 cm)	0	0	0	5
very deep (> 120 cm)	3	0	0	7
11_SOIL_LUSEPO				
wasteland, shrub	75	**83**	**83**	62
pasture, grassland, grazing land	21	17	17	19
arable land, cereals	5	0	0	13
forest, coppice	0	0	0	5

10_NDVI_30CL	Western C.		all Crete	
Class name	% area	% pres	% pres	% area
class 18	6	**26**	**17**	4
class 15	1	**9**	**14**	1
class 22	8	9	10	7
class 6	4	7	7	1
class 20	7	9	7	5
class 21	4	11	7	7
class 23	8	9	6	7
class 28	16	9	6	6
class 19	5	0	4	5
class 25	7	4	4	7
class 4	0	2	3	0
class 10	0	0	3	0
class 16	1	0	3	0
class 2	0	0	1	0
class 7	0	0	1	0
class 14	4	2	1	3
class 24	1	2	1	**11**
class 27	8	2	1	**11**
class 1	0	0	0	0
class 3	0	0	0	0
class 5	0	0	0	0
class 8	0	0	0	0
class 9	0	0	0	0
class 11	3	0	0	2
class 12	0	0	0	0
class 13	0	0	0	0
class 17	0	0	0	1
class 26	1	0	0	1
class 29	0	0	0	**11**
class 30	16	0	0	6

12_CORINE_L3	Western C.		all Crete	
Class name	% area	% pres	% pres	% area
321 Natural grasslands	21	**30**	**36**	20
323 Sclerophyllous vegetation	26	**33**	**28**	24
332 Bare rocks	2	7	**10**	1
223 Olive groves	16	13	9	**23**
243 Land principally occupied by agricultu	8	7	6	10
324 Transitional woodland-shrub	5	4	4	4
333 Sparsely vegetated areas	3	2	3	3
312 Coniferous forest	5	0	3	3
242 Complex cultivation patterns	5	4	1	6
111 Continuous urban fabric	0	0	0	0
112 Discontinuous urban fabric	1	0	0	1
121 Industrial or commercial units	0	0	0	0
122 Road and rail networks and associate	0	0	0	0
123 Port areas	0	0	0	0
124 Airports	0	0	0	0
131 Mineral extraction sites	0	0	0	0
133 Construction sites	0	0	0	0
142 Sport and leisure facilities	0	0	0	0
211 Non-irrigated arable land	0	0	0	0
212 Permanently irrigated land	0	0	0	0
221 Vineyards	0	0	0	3
222 Fruit trees and berry plantations	2	0	0	1
231 Pastures	0	0	0	0
311 Broad-leaved forest	2	0	0	1
313 Mixed forest	0	0	0	0
322 Moors and heathland	1	0	0	0
331 Beaches, dunes, sands	0	0	0	0
512 Water bodies	0	0	0	0
09_SOIL_VS				
0 % stones	33	43	45	38
5 % stones	52	43	35	42
10 % stones	16	13	20	20

21_ASTER_35CL	Western C.	
Class name	% area	% pres
class 34	5	**17**
class 30	11	**16**
class 33	9	**12**
class 12	4	**9**
class 31	2	7
class 24	5	5
class 20	4	5
class 18	3	5
class 35	2	5
class 27	6	3
class 11	1	3
class 29	6	3
class 32	5	2
class 22	1	2
class 28	**8**	2
class 21	4	2
class 10	0	2
class 26	**5**	0
class 19	4	0
class 25	4	0
class 13	2	0
class 16	2	0
class 14	2	0
class 15	1	0
class 23	1	0
class 17	1	0
class 8	0	0
class 6	0	0
class 9	0	0
class 4	0	0
class 3	0	0
class 7	0	0
class 5	0	0
class 2	0	0
class 1	0	0

Note:
% pres: = proportion of presence sites (out of all presence sites) falling into respective class
% area = proportion of land mass covered by the respective class

Appendix H Preliminary count-based association of NDVI and ASTER with CORINE predictor (table showing NDVI counts above and ASTER counts below)

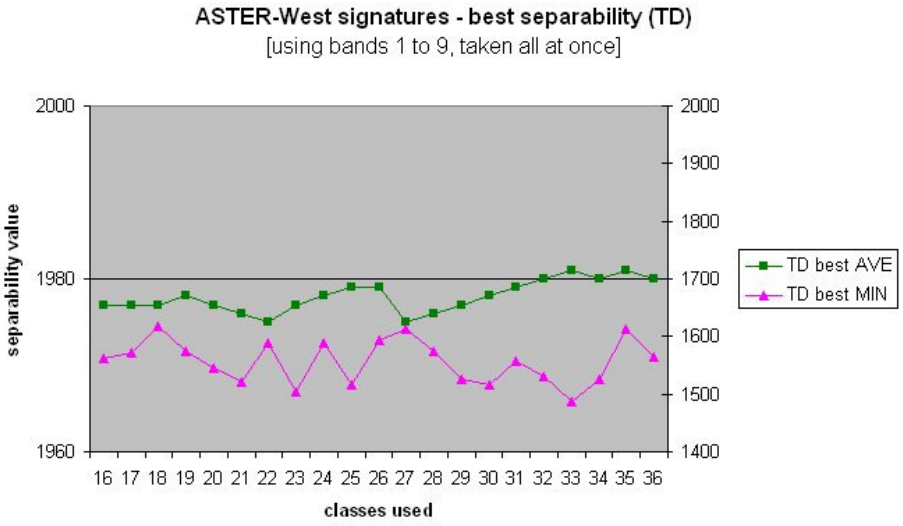
Anzahl von NDVI	COR										
NDVI	223	242	243	312	321	323	324	332	333	NoData	Total
2									1		1
4									1	1	2
6								5			5
7										1	1
10					2						2
14							1				1
15			1		6	2		1			10
16						1				1	2
18					3	9					12
19				2		1					3
20	1				2	2					5
21		1	1		1	1				1	5
22					6		1				7
23			1		1	2					4
24	1										1
25					1	1	1				3
27	1										1
28	3		1								4
NoData											
Total	6	1	4	2	22	19	3	6	2	4	69

Anzahl von ASTER	COR										
ASTER	223	242	243	312	321	323	324	332	333	NoData	Total
10										1	1
11		1								1	2
12			1			4					5
18	1		1			1					3
20	3										3
21							1				1
22							1				1
24	1		1		1						3
27					1	1					2
28				1							1
29						2					2
30	1				4	4					9
31					2	1		1			4
32					1						1
33					2	5					7
34				1	6		1	1	1		10
35								3			3
NoData											
Total	6	1	3	2	17	18	3	5	1	2	58

Appendix I Coefficients for Relative Atmospheric Correction required to mosaic ASTER data from 2002 (far Western Crete) with ASTER data from 2006 (central Western Crete); coefficients derived from a dozen manually placed Pseudo-Invariant Features.

equation applied to match W 2002 to W 2006:							R^2:
new W 2002 band 1 = (	old W 2002 Band 1	-	9.3844	)	/	0.9031	0.9885
	old W 2002 Band 2	-	6.2901	)	/	0.9263	0.9918
	old W 2002 Band 3n	-	6.3934	)	/	0.9362	0.9821
	old W 2002 Band 4	-	1.1448	)	/	0.945	0.9937
	old W 2002 Band 5	-	0.4179	)	/	0.9201	0.996
	old W 2002 Band 6	-	0.3554	)	/	0.9287	0.9953
	old W 2002 Band 7	-	0.3324	)	/	0.9242	0.9937
	old W 2002 Band 8	-	0.175	)	/	0.9168	0.9894
	old W 2002 Band 9	-	0.1345	)	/	0.8893	0.9918

Appendix J Determining the optimum number of classes (35) for ASTER-West ISODATA classification based on TD values



Errata

Page 15: "earthquake-induced collapse in 1450, Crete is now"
Should be replaced by "earthquake-induced collapse in 1450 BC, Crete is now"

Page 22: "Limit projection to Western Crete"
Should be replaced by "Limit extent of input predictors to Western Crete"

Page 29: „only simple multiple regression was performed"
Should be replaced by „Independence of categorical predictors was not tested; for continuous predictors, a simple correlation matrix was compiled showing the correlation coefficients. Values near +1 and -1 indicate strong positive and negative correlation respectively; values near 0 indicate low inter-variable dependency. Predictors showing a high negative correlation indicate potentially harmful collinearity (see also page 59), which in turn may result in a model prediction strength below the potential maximum and an insufficient recognition of the individual contribution of affected predictors."

Page 38:
"ASTER imagery was prepared in a similar fashion for the most of Central Crete as well;"
Should be deleted without substitution.

Page 51: "able to generate about 30% of the 'cumulative gain'"
Should be replaced by "able to generate about 25% of the "cumulative gain'"

Page 51: "inclusion of the ASTER predictor increases the significance of the AUC"
Should be replaced by "inclusion of the ASTER predictor increases the AUC value"

Page 53: "Based on visual comparison of Figure 8 Probability Distribution in Western Crete without ASTER predictor (above) with Figure 11 (below)"
Should be replaced by "Based on visual comparison of Figure 9 (above) with Figure 11 (below)"

Page 57: "*P. erhardii* presence records coincide significantly with 'open areas', i.e. the most frequent NDVI class belongs to the lowest third of all NDVI classes generated for Crete (using an ISODATA classification)..
→ refer to .xls im anhang, do minitab z test (but it will fail anyway). Accept Ha."

Should be deleted without substitution.

Page 58: “the rejection of hypothesis 2b) should therefore not be understood as a statement of superiority”
Should be replaced by “the rejection of the Null hypothesis (2b) should therefore not automatically be understood as a statement of superiority”.

Page 63: “although these 10 points covered only”
Should be replaced by “although these 10 points represent only”