THE APPLICATION OF TREE-BASED METHODS TO SPECIES DISTRIBUTION MODELLING

BY

MARCO GIRARDELLO

A THESIS SUBMITTED IN CANDIDATURE FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

SCHOOL OF BIOLOGY
UNIVERSITY OF NEWCASTLE UPON TYNE

JULY 2009

ABSTRACT

Species distribution models are used increasingly in both applied and theoretical research to predict how species are distributed and to understand attributes of species' environmental requirements. This thesis aims to explore the application of tree-based methods to species distribution modelling. Although these methods have been widely used in other fields of science they have received relatively little exposure in Biogeography and Conservation Biology. The techniques applied include CART, Bagging, Random Forests and Boosted Regression Trees. These were used with four different biodiversity databases to answer different a variety of research questions aimed at: (i) understanding how landscape structure and climate affect species distributions (ii) predicting the potential impacts of climate change on species distributions (iii) to identify areas important for biodiversity conservation. Additionally, the performance of each method was compared with the aim (iv) of making suggestions for the optimal models which should be used by future researchers.

In chapter 2 Boosted Regression Trees were used to quantify the importance of wetland size and weather patterns for waterbird distribution in Britain. As well as revealing the importance of wetland size for waterbirds, , the models proved to be reasonably robust when validated. In chapter 3 this basic form of modelling was expanded, using a database containing amphibian occurrence records for Italy. Random Forests was used to quantify species-climate relationship and to predict amphibian distribution in relation to current and future climate conditions. The results revealed how amphibian distribution is largely controlled by temperature-related variables and highlighted a negative response to future climate changes in most species. In chapter 4 Bagging was used to identify areas important for biodiversity conservation. Specifically, Bagging was used predict the distribution of 232 species of Butterflies in Italy. The predicted surfaces were then used in combination with a species multispecies prioritization tool in order to identify important areas for butterfly conservation. The results

showed that the most areas important for butterfly are located within the Alps, the mountains of central Italy and the island of Sardinia. Finally, in Chapter 5, the predictive accuracy of four modelling techniques based classification trees was compared. This was done using large scale bird distribution data from Italian Common Bird Census. The results showed that Random Forests and Boosted Regression Trees were the best performing techniques and that model performance was highly influenced by species ecological characteristics as well as by the modelling method.

The results of this thesis have shown how tree-based modelling methods can be used for exploring and testing hypotheses about the factors that are important in determining species distribution and making predictions of species distribution for use in conservation contexts. The methods used represent a useful way to visualize and understand relationships between environmental parameters and species distributions and to predict species distributions with high accuracy. Whilst it is true that some tree-based methods can be used instead of statistical modelling techniques others expand the analytical opportunities by enabling analyses that are impossible or very difficult with statistical methods. Hopefully this thesis will serve a source of inspiration for ecologists willing to move away from statistical inference and the P-value dogma and concentrate on understanding the data, and using alternative techniques to predict species distribution with high accuracy.

DECLARATION

No portion of the work presented in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

ACKNOWLEDGEMENTS

I would like to thank my supervisors, Dr. Steve Rushton and Dr. Mark Whittingham for their continuous enthusiasm, support and encouragement.

The opportunity for and completion of this study would have been impossible without the support, encouragement and financial help from family for which I am most grateful.

I would also like to thank everyone at the Life Science Modelling Group for their help and encouragement over the years.

LIST OF ORIGINAL PAPERS

Chapter 3- Models of the climate associations and distributions of amphibians in Italy. *Ecological Research (accepted with major revision)*

Chapter 4- Using niche modelling and landscape zonation to identify important areas for butterfly conservation in Italy. *Animal Conservation (accepted with major revision)*

Chapter 5- A comparison of tree-based methods for modelling species distributions. *Ecological Informatics (in review)*

CONTENTS

LIST OF FIGURES viii

LIST OF TABLES ix

NOMENCLATURE x

CHAPTER 1: INTRODUCTION

1.1. SPECIES DISTRIBUTION MODELS…….. 1

1.2. THE NICHE CONCEPT…… 1

1.3. SPECIES DISTRIBUTION MODELLING METHODS…….. 3

1.4. APPLICATIONS OF SPECIES DISTRIBUTION MODELS…….. 5

1.5. OUTLINE OF THE THESIS……. 7

1.6. SPECIFIC OBJECTIVES…….. 8

CHAPTER 2: DISTRIBUTION PATTERNS OF WATERFOWL WINTERING IN BRITAIN: THE ROLE OF GEOGRAPHY, CLIMATE AND HABITAT

2.1. INTRODUCTION…….. 10

2.2. METHODS……. 12

2.2.1 Waterfowl and environmental data ……. 13

2.2.2 Analyses…… 14

2.2.3 Model evaluation…….. 15

2.4. DISCUSSION……..... 20

CHAPTER 3: MODELS OF THE CLIMATE ASSOCIATIONS AND DISTRIBUTIONS OF AMPHIBIANS IN ITALY

3.1. INTRODUCTION ……..... 23

3.2. 24

METHODS…….. 24

3.2.1. Species and climatic data…….. 27

3.2.2 Analyses…… 29

3.2.3 Model evaluation and predictions……

3.3. RESULTS……..... 30

3.4. DISCUSSION……..... 35

CHAPTER 4: USING BAGGING PREDICTORS AND LANDSCAPE ZONATION TO IDENTIFY IMPORTANT AREAS FOR BUTTERFLY CONSERVATION IN ITALY

4.1. INTRODUCTION …… 39

3.2. METHODS…… 42

4.2.1 Species data and environmental predictors …… 42

4.2.2 Species distribution modelling and model evaluation…… 44

4.2.3 Zonation and management landscapes…… 46

4.3. RESULTS……..... 49

4.4. DISCUSSION…… 55

CHAPTER 5: A COMPARISON OF TREE-BASED METHODS FOR MODELLING SPECIES DISTRIBUTIONS

5.1. INTRODUCTION……..... 61

5.2.METHODS…….. 63

5.2.2 Environmental variables…….. 64

5.2.3 Analyses…… 65

5.2.4 Model evaluation……... 69

5.2.5 Analysis of model performance…… 70

5.3 RESULTS…… 71

5.4 DISCUSSION…… 76

CHAPTER 6: GENERAL DISCUSSION

6.1 SPECIES DISTRIBUTION MODELS- A CHALLENGE PLAGUED BY UNCERTAINTY……. 83

6.2 PUTTING TREE-BASED METHODS WITHIN A WIDER CONTEXT……... 87

APPENDIX 1 …… 90

APPENDIX 2 …… 93

REFERENCES ……... 95

LIST OF FIGURES

FIGURE 2.1	The distribution of the 438 sites surveyed as part of the Wetland Bird Survey from 1993 to 2001	13
FIGURE 3.1	Examples of the projected potential ranges for three amphibian species (*Bombina variegata*, *Rana latastei*, *Salamandrina terdigitata*): (a) current, (b) projected future with unlimited dispersal, and (c) projected future without dispersal.	34
FIGURE 3.2	Projected amphibian species losses (with no dispersal) and gains (with unlimited dispersal)	35
FIGURE 4.1	Relationship between model performance (AUC) and the number of species records records (R2 0.10). Each point represents a species	49
FIGURE 4.2	Results of the basic Zonation, without species weightings. Map shows the results of the ranking for each cell, which varies from 0 to 1. Values closer to 1 are shown in red and represent the cells with a higher biological value, whereas cells with a value closer to 0 are shown in green and have a lower biological value. Arrows indicate: A (Western Alps) B(Eastern Alps) C (Central Appennines) D (Apulia) E (Sardinia).	51
FIGURE 4.3	Results of the Zonation when species weighting was applied. Arrows indicate: Alps (A), Calabria (B) and Sicily (C).	52
FIGURE 4.4	Comparison of the top 10% fraction of cells selected by the two, non-weighted and weighted solutions. The green squares indicate the cells which were selected in both solutions, the orange ones those selected only in the non-weighted solution, and the red ones the ones selected only in the weighted solution.	53
FIGURE 4.5	Average proportion of the original distribution retained for the species of conservation concern as a function of proportion of landscape remaining as lower priority zones. The two different lines show the average of the proportions for the non-weighted (solid line) and weighted solutions (dashed line).	54

FIGURE 4.6 Priority landscapes derived from both the non-weighted (A) and weighted (B) solutions. The landscapes were selected using the top 10% fraction of cells selected by the Zonation. Each landscape (shown by a colour) contains blocks of land that are close together, similar in species composition, and contain a core area present late in the cell removal process 55

FIGURE 5.1 Mean performance (AUC) for each of the four modelling methods (Abbreviations: BAG, Bagging predictors; BRT, Boosted regression trees; CART, Classification tree; RF, Random Forests). Error bars represent mean ± 1 standard error. 72

FIGURE 5.2 Histograms showing the P -value of difference (as obtained from Wilcoxon tests) from the best performing vector. Higher P-values indicate a better performance. Techniques are ranked in order of performance. 73

FIGURE 5.3 Mean AUC (over all methods) vs. prevalence (A) and marginality (B) components. The prevalence component describes a gradient from abundant to scarce species (i.e. from left to right). Marginality is an index which describes how far the species optimum is from the mean environmental profile in the study area. Higher values of the marginality component indicate more marginal species 75

LIST OF TABLES

TABLE 2.1	Area Under the ROC curve (AUC) for each of the thirteen species of waterfowl.	17
TABLE 2.2	Importance of the environmental and geographic variables in determining the distribution of each of the thirteen waterbird species, as obtained from the Boosted Regression Tree models.	19
TABLE .3.1	Characterization of the climatic variables used for the analyses.	26
TABLE 3.2	Results of model validation and predictions for present climate and future climate scenarios. The results show the AUC value for each species and the number occupied grid cells by each species for present and future climate conditions under the assumptions of both unlimited and no dispersal.	31
TABLE 3.3	Five most important predictor variables for each species. The names of the variables are listed in decreasing order of importance (for acronyms see Table 1) . The number between the brackets indicates (eg. 1.17) the mean decrease in accuracy, which represents the overall percentage mean decrease in the prediction error of the model when one particular variable is permuted while all the other variables are held constant. Higher values of mean decrease in accuracy indicate variables that are more important in determining a species' distribution.	32
TABLE 4.1	Characterization of environmental and climatic variables used in the analyses.	44
TABLE 5.1	Predictor variables used for the modelling of bird species using the 4 predictive techniques	65
TABLE 5.2	Results of the linear mixed effect model investigating the determinants of area under the curve (AUC) scores. AUC score were modelled as a function of the main fixed effects and model and their two way interactions. Each unique species was treated as a random effect.	74

1 INTRODUCTION

1.1 Species distribution models

One of the main goals of ecology is the study of species distributional patterns, both spatially and temporally, and the determination of the underlying environmental factors related to these distributional patterns (Brown & Lomolino, 1998). In most cases, however, the distribution of a species is not completely known. Because of the high costs associated with sampling over large regions, species distribution models have been frequently used in ecology and conservation since the 70s (Guisan & Thuiller 2005). Species distribution models are empirical models relating field observations to environmental predictor variables, based on statistically or theoretically derived response surfaces (Guisan and Thuiller 2005). These models serve two important purposes: (1) they are used to formulate and test hypotheses about the factors and processes that are important to organisms, and (2) they can be used to make predictions of species distributions and abundances for use in management decisions.

1.2 The niche concept

The fundamental ecological principle on which species distribution models are based is the concept of niche (Guisan and Zimmerman 2000, Pullian 2002), and here I will give an overview of the concept. The first definition of niche was given by Grinnell (1917), who defined the species niche as the 'environmental requirements of the species' and considered it as an 'ultimate distributional unit of the species'. Later, Elton (1927) defined the niche as 'the role of the species in the

community', which integrates its interactions with other species. Both of Grinnell's and Elton's definitions of niche are conceptually vague (Whittaker et al. 1973, Heglund 2002) but were later rigorously combined by Hutchinson (1957), who describe the niche by 'the coordinates of the species with *n*-dimensional resource axes'. Hutchinson defined the niche as a 'hypervolume' situated in n-dimensional 'hyperspace'; this hypervolume encloses 'conditions that allow the species to exist indefinitely' ('fundamental' niche). However, because of interspecific interactions, the species may be excluded from some parts of its fundamental niche, reducing the hypervolume ('realized' niche). In the real world, the fundamental niche is unlikely to be observed and researchers normally focus on describing the realized niche (Scott et al. 2002, Guisan and Thuiller 2005). All the definitions of niche are highly conceptual and rely on assumptions that could be violated, and several authors have recently criticised or revised the niche concept (Chase and Leibold 2003). However, Hutchinson was the first to provide a formalization of the niche concept and the Hutchsonian niche definition has since become the foundation of much ecological theory and reasoning (Morrison et al. 1998, Pulliam 2000, Scott et al. 2002).

Species distribution models are based on the Hutchsonian definition of the realized niche (Guisan and Zimmermann 2000, Guisan and and Thuiller 2005, Araujo and Guisan 2006). This is because existing species distribution represents the realized rather than the fundamental niche. Species distribution models cannot predict patterns of species distributions based on the fundamental niche, because the fundamental niche is a theoretical abstract concept that cannot be observed in

real world phenomena. Being empirical models, species distribution models can only describe phenomena that are observable in the real world.

1.3 Species distribution modelling methods

A huge number of techniques and tools have become available for modelling species distributions (Guisan and Zimmerman 2000, Guisan et al. 2003, Elith et al. 2006). Classical methods include statistical techniques like Generalized Linear Models (GLM) (McCullagh and Nelder 1989) or Generalized Additive models (GAM) (Hastie and Tibshirani 1990). These methods rely on the use of presence and absence data, but can also incorporate abundance data. More recently, methods of modelling that use presence-only data have been developed, such as ecological niche factor analysis-E.N.F.A. (Hirzel et al. 2002), BIOCLIM (Busby, 1991) DOMAIN (Carpenter et al., 1993). Such methods rely on the definition of environmental envelopes around locations where species occur, which are then compared to the environmental conditions of background areas (Hirzel et al. 2002).

While the benefits of using regression and envelope methods are numerous, including predicting changes in species' distribution from climate change (eg. Hilbert et al. 2004, Raimo et al. 2008, Meynecke 2004) and identifying areas important for biodiversity conservation (eg. Milne et al. 2006, Lehmann et al. 2002), species distribution modelling is complicated by technical difficulties and by data limitations (Guisan & Thuiller 2005).

In recent years, the introduction of machine-learning techniques has opened new avenues to the analyses of ecological data (Fielding 1999). These

methods can be used to solve problems, which derive from assumptions about the statistical distribution of data or restrictive assumptions of parametric modelling methods (Fielding 1999) . Machine learning methods are often more powerful, flexible, and efficient for exploratory analysis than are statistical techniques. Machine learning methods (e.g., Hastie et al. 2001) can be characterized as an analysis of data 1) which automatically makes accurate predictions from data, 2) with the ability to screen a large number of predictor variables and identify the most important predictors, 3) very often they do not require the user to make many assumptions about the forms of relationships between predictor variables and the response variable. There are a variety of modelling methods that have been have been developed within the machine learning community. These include neural networks, rule based classifiers, genetic algorithms and maximum entropy modelling approaches.

Among the most common machine learning methods, classification trees represent an efficient tool, that has been applied in several studies in ecology and conservation biology (e.g. Moisen et al. 2006, Edwards et al. 2006, Thuiller et al. 2003). This type of models are built using brute-force computer algorithms. Classification trees, often known as CART (Classification and Regression Tree), aim to explain the variation in a single response variable with respect to one or more explanatory variables. They work by partitioning the data recursively into smaller homogenous groups with respect to the response variable. Conceptually, a classification tree treats species as if they were constrained to live within certain variable ranges. Every split within the tree marks either a lower or upper bound of

the range for a particular habitat variable. Usually, only one end of the range is recorded into the tree. For example, a classification tree model identifies a series of habitat ranges, the pieces of the tree are conceptually very similar to a quantitative version of Hutchinson's (1957) n-dimensional niche- the habitat space in which a species is able to maintain a population. In addition the baseline methodology of classification trees includes several novel methods which have also been developed. These methods use iterative or bootstrapping procedures to combine several hundreds or thousands of trees together with the aim of improving model accuracy. Although these so far these techniques have been applied in a few studies they have shown a great promise by outperforming most of the traditional modelling methods (Garzon et al. 2006, Elith et al. 2006, Guisan et al. 2007a, Graham et al. 2008, Wisz et al. 2008) .

1.4 Applications of species distribution models

Species distribution models are currently recognised as helpful tools for providing valuable and quantitative information by revealing the most important resources required by a species (Guisan and Thuiller 2005). Models can efficiently guide decision makers and wildlife managers in processes of protection, management or conservation planning (e.g. . Ortega-Huerta and Peterson, 2004; Pawar et al. 2007; Moilanen et al. 2007) or can be used to develop environmental impact assessment programs (e.g. Seiler 2005, Tuck et al. 2001). Furthermore, if coupled with geographic information systems (GIS) technology, species distribution models can be used for producing maps that display the

spatial configuration of the suitable habitats, which enables protection, management and restoration strategies to be implemented within a spatial context (e.g. Rayner et al. 2007, Martinez et al. 2006, Santos et al. 2006). Besides revealing habitat selection patterns, the application of species distribution models to areas where environmental conditions are known but where species distributions are unknown yields habitat suitability maps (Buermann et al. 2008). Predicting to new areas where species distributions are unknown can be very important for the identification of biodiversity hotsposts (Thuiller et. al. 2006) or locations of species of conservation concern (Thomaes et al. 2008). Newly identified hotspots or important areas for species of conservation concern can be the subject of more intensive study. Species distribution models are also useful for predicting areas of suitable habitat that may not be currently used by wildlife species, serving as an aid to species re- introduction or prediction of the spread of an introduced species (Klar et al. 2008, Metzger et al. 2007). Another very important application of species distribution modelling is the prediction of the potential impacts of environmental change on species distributions. In particular models can be used to investigate how different species respond differently to environmental influences and to explore the unique reactions of different species to different environmental change scenarios. In conjunction with the recent surge of interest in the potential effects of climate change, species distribution models have been widely used in both understanding climate-species relationships and predicting how species distribution patterns will change in response to climate change. Numerous studies have extrapolated the likely impacts of global change

on species distribution(e.g. Araujo et al. 2006, Luoto & Heikkinen 2008, Thuiller et. al. 2005, Thomas et al. 2004).

1.5 Outline of the thesis

The thesis follows two major aims, a technical and an ecological one. The former one focuses on the application of various tree-based modelling methods to species distribution modelling. While regression based methods have been widely applied to species distribution modelling, this thesis deals with the application of some novel modelling techniques based on classification trees (Chapter 2, Chapter 3, Chapter 4). The methods applied include Random Forests, Bagging and Boosted Regression Trees, three techniques which were introduced in ecology very recently (Leathwick et al. 2006, Prasad et al. 2006, Cutler et al. 2007) . Additionally the thesis compares the predictive performance of the four tree-based methods for modelling species distributions (Chapter 5). The ecological aim of the thesis is concerned with solving three different large scale ecological problems: (i) understand the major determinants of species distribution and develop some highly predictive models (Chapter 2); (ii) analyse and predict species distribution in relation to present and future climate (Chapter 3); (iii) identify areas important for biodiversity conservation (Chapter 4). Chapter 2 focuses on the application of Boosted Regression Trees to the identification of what factors determine waterbird occurrence in Britain. In Chapter 3 I use another tree based technique named Random Forests to analyses species-climate relationships in amphibians occurring in Italy and to predict their response to climate change. In Chapter 4 I use Bagging predictors in combination with a multispecies conservation planning

tool to predict important areas for butterfly conservation in Italy. In chapter 5 I compare the performance of four tree-based modelling methods for predicting species distributions using large scale breeding bird distribution data. The final chapter 6 of this thesis is a general discussion and brings together the findings of the thesis into an overall synthesis of state-of-the-art species distribution modelling in ecology.

1.6 Specific objectives

The thesis addresses the following specific objectives:

Chapter 2: (1) To develop robust models capable of predicting waterbird distribution in Britain and (2) to assess the importance of habitat and landscape structure, climatic and geographic variables in determining waterfowl occurrence.

Chapter 3: To (1) to assess the importance of current climate in determining amphibian occurrence in Italy; (2) to examine the potential changes in the distribution of amphibians under a $2xCO^2$ future climate scenario.

Chapter 4 (1) To identify areas important areas for butterflies across the Italian peninsula; (2) to identify important areas for species of conservation concern; (3) to identify potential 'management landscapes' based on the similarity in species composition among sites.

Chapter 5 (1) To compare the predictive performance of four different modelling techniques based on decision trees (2) to establish whether model performance is affected by the species' environmental and geographical distributions.

2 DISTRIBUTION PATTERNS OF WATERFOWL WINTERING IN BRITAIN: THE ROLE OF GEOGRAPHY, CLIMATE AND HABITAT

2.1 Introduction

One of the main goals of Ecology is the study of species distributional patterns, both spatially and temporally, and the determination of the underlying environmental factors related to these distributional patterns (Brown & Lomolino, 1998). Understanding how the distribution of species and communities are affected by these factors is needed in many contexts, for example, for applied conservation purposes or for assessing the potential impacts of large scale ecosystem changes such as those brought about by global climatic change. Modelling species-environment relationships at large scale has become very popular during the last decade and examples in ecology and conservation include many types of taxa (e.g. Virkkala et al., 2005; Luoto et al., 2006; Suarez-Seoane et al., 2002; Coops & Catling, 2002; Milne et al., 2006).

Waterfowl are probably one of the best surveyed groups of vertebrates in Europe, and several long-term monitoring schemes have been underway since the 1960s. The high gregariousness and relative ease with which most species can be monitored during the nonbreeding season has enabled ecologists to gather a considerable amount of data about their distribution. Despite the great availability of waterfowl count data there have been very few attempts to understand the

factors affecting their distributional patterns on their wintering grounds (Tuite et al. 1984 Paracuellos and Telleria 2004; Santoul 2004).

In this chapter I examine the relationships affecting the distribution of thirteeen waterfowl species regularly wintering in Britain. The importance of Britain for wintering waterfowl is well known and every year millions of individuals migrate from high artic areas to spend their winter in British inland wetlands and estuaries (Kershaw and Cranswick 2003; Ravenscroft et al. 2003). In Britain, waterfowl have been the subject of monitoring since the 1960s as part of the Wetland Bird Survey, the UK's national monitoring scheme for nonbreeding waterfowl. Up to date, the main output of this monitoring scheme has been aimed at establishing population trends (eg. Kirby et al. 1995; Atkinson et al. 2006), with no attempt to quantify distributional patterns of waterfowl at the national scale.

I use a machine learning method, boosted regression trees (Friedman 2001), to model species-environment relationships. This technique is a development of decision trees and has shown to be very promising for both analysing ecological datasets and predicting species' occurrence (Elith et al. 2006, De'ath 2007). As well as having a superior predictive performance, Boosted regression trees have other advantages over traditional modelling methods, which make it very desirable. These include being insensitive to extreme outliers and the ability to handle interactions between predictors. Up to date boosted regression trees have found many applications in a variety of fields like medicine, genetics,

remote sensing and epidemiology, but there have been very few applications in ecology (Guisan et al. 2007, Elith et al. 2006, Thuiller et al. 2006).

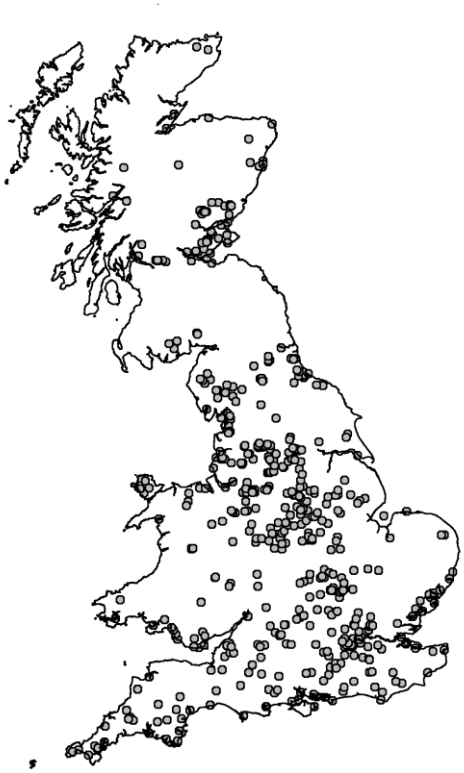
The specific aims of this research were twofold: 1) to develop robust models capable of predicting waterfowl distribution in Britain 2) to assess the importance of habitat, weather patterns and geography in determining in determining waterfowl occurrence.

2.2 Methods

2.2.1 Waterfowl and environmental data

The data were derived from the results of the Wetland Bird Survey (WeBs). The Wetland Bird Survey is the UK's national monitoring scheme for nonbreeding waterbirds. The scheme has been underway since 1965 and is administered and funded by the British Trust for Ornithology (BTO), the Wildfowl and Wetlands Trust (WWT), the Royal Society for the Protection of Birds (RSPB) and the Joint Nature Conservation Committee (JNCC). In brief it consists of synchronised counts which are carried out every month in collaboration with experienced volunteer ornithologists. A detailed description of the sampling procedure has been described by Kirby (1995) and Cranswick et al. (1997). For the present analysis I used a set of 438 sites surveyed from 1996 to 2000 during the month of January. I decided to convert the counts for each into a binary matrix of presences and absences as my intent was to study species' distribution and not to examine species spatio-temporal patterns of abundance. Therefore if a species was recorded at a site it was considered to be present. The distribution of the sites is shown in Fig. 2.1.

Figure 2.1. The distribution of the 438 sites surveyed as part of the Wetland Bird Survey from 1993 to 2001



Two geographic and six environmental variables were used for the analysis: location of each site identified as Latitude and Longitude, surface area of each water body (variable area) calculated for each site calculated from a file provided by the British Trust for Ornithology, mean temperature (ºC) and total amount of

precipitation (mm) of January, and area (m^2) of 3 different land-cover types (variables Grassland, Swamp, Urban) within a 500m buffer surrounding each site. These latter variables were derived from a 25m resolution land-cover map of the UK (Fuller et al. 2005).

2.2.2 Analyses

Boosted regression trees are a development of regression trees which combine the regression tree methodology together with the boosting algorithm (Friedman 2001; Friedman and Meulman 2003). A regression tree differs from conventional regression methods by using a rule partitioning method to classify the data. The tree is built by repeatedly splitting the calibration data, according to a simple rule based on a single explanatory variable (Breiman 1984). At each split, the data are partitioned into two exclusive groups, each of which is as homogeneous as possible in terms of the response variable. The boosting algorithm is a very general method that attempts to "boost" the accuracy of any given learning algorithm by fitting a series of models each having a poor error rate and then combining them to give an ensemble that may perform very well. In a boosted regression tree a series of very simple regression trees is fit and combined to obtain a final model. The model is developed by progressively adding trees in a forward stagewise fashion. At each stage of the fitting sequence, each case of the response variable is classified from the current sequence of trees. These classifications are used as weights (i.e. pseudo-residuals) for fitting the next sequence of trees. The fitting procedure is then continued until all the data have been explained (Friedman 2001).

Although boosted regression trees can be a powerful tool to analyse complex data sets they are also prone to over-fitting (i.e. trees can be added until eventually all the data will be explained). As a consequence of this, the performance of the final model will degrade when applied to new data. In order to overcome this problem procedures like k-fold cross-validation have to be used to identify the optimal number of trees. In this case I used a 10-fold cross validation procedure to identify the optimal number of trees, following the procedure described in Elith et al. (2008). I also assessed the relative influence of each variable in determining species' occurrence. In a single tree the importance of each variable is determined by the reduction of the impurity (i.e. deviance) when that variable is split on. In a boosted regression tree the importance of a variable is simply determined by averaging the importance of that variable across all trees used to construct the final model. I also selected a simple interaction term for the trees, allowing for a two-way interaction. This is done by specifying the size of each individual tree. In The case I specified a tree size of two allowing for two way simple interaction . All the analyses were carried out using the gbm package (http://www.i-pensieri.com/gregr/gbm.shtml) for R (Ihaca and Gentleman 1996) developed by Greg Ridgeway.

2.2.3 Model evaluation

In order to evaluate The models, the original data set were randomly divided into model training (70%) and model evaluation data sets (30%). The discrimination ability of the models was assessed using receiver operating

characteristic (ROC) plot analysis. This technique measures the association between the presence and absence records by using and calculating the area under the curve (AUC) (Fielding & Bell, 1997). AUC relates relative proportions of correctly classified (true positive proportion) and incorrectly classified (false positive proportion) cells over a wide and continuous range of threshold levels, which makes it a threshold-independent measure (Fielding & Bell, 1997). The AUC values range from <0.5 for models with no discrimination ability to 1 for models with perfect discrimination.

2.3 Results

The results of the models are summarized in Tables 2.1 and Table 2.2. Table 2.1 shows the AUC values obtained from model evaluation using the data from the evaluation sites. The AUC values ranged from 0.84 to 0.63, with the Goldeneye and the Mute Swan models performing the best and the Mallard model having the lowest performance. With exception of the Mallard and the Goosander the AUC values were all higher than 0.70. There was not any taxonomic ecological similarity in species having similar AUC values.

The importance of each variable varied amongst the species (Table 2.2), but the most important variables were generally related to climate, geography and size of the water body. Temperature and rain ranked as the most important variables for eight species and were also the second and third most important variable for a further five species. For two species, the Goldeneye, and Goosander temperature and rain were unimportant important ranking seventh and eighth, respectively.

Table 2.1. Area Under the ROC curve (AUC) for each of the thirteen species of waterfowl.

Species	AUC	Species	AUC
Canada Goose	0.83	Pochard	0.79
Gadwall	0.80	Ruddy Duck	0.83
Goldeneye	0.84	Shoveler	0.78
Goosander	0.66	Teal	0.76
Mallard	0.61	Tufted Duck	0.77
Mute Swan	0.84	Wigeon	0.76
Pintail	0.73		

Both Latitude and Longitude were very important for the Canada Goose, the Goldeneye. Latitude was also important for species like the Gadwall and the Mallard, whereas Longitude turned out to be the third most important variable for the Tufted Duck. The grassland variable was particularly important for species

like the Pintail, the Teal, the Wigeon and the Goosander. For the other species, this variable ranked from medium importance (Goldeneye, Shoveler) to unimportant (Canada Goose, Mute Swan, Pochard, Ruddy Duck, Gadwall, Tufted Duck). The urban variable was generally unimportant, ranking very low for more than half of the species, but it was the third most important variable for the Shoveler, the Mute Swan and the Ruddy Duck. Lake area was the most important variable for three species (Goldeneye, Goosander, Ruddy Duck) and was the second most important variable for the Wigeon, the Teal and the Pochard. For one species, the Mallard lake area was the most unimportant variable. Lastly, the area of swamp surrounding each site was unimportant for the majority of the species. For one species, the Mallard, this variable turned out to be the third most important variable and for another species, the Pintail it had a medium importance, ranking as the fourth most important variable.

Table 2.2. Importance of the environmental and geographic variables in determining the distribution of each of the thirteen waterbird species, as obtained from the Boosted Regression Tree models.

	Longitude	**Latitude**	**Rain**	**Swamp**	**Temp**	**Urban**	**Grass**	**Area**
Canada Goose	12.18	56.24	12.54	0.01	7.66	9.78	1.19	0.39
Gadwall	10.57	22.48	31.08	0.01	15.01	5.13	0.86	14.88
Goldeneye	12.76	19.55	1.00	0.06	0.89	3.01	8.20	54.53
Goosander	6.69	3.25	2.64	1.13	21.35	8.18	18.60	38.16
Mallard	7.03	24.52	25.35	13.76	11.07	8.31	5.67	4.29
Mute Swan	4.62	6.86	30.74	0.01	48.17	7.63	1.18	0.78
Pochard	8.87	11.99	16.35	0.03	31.40	6.07	1.22	24.07
Ruddy Duck	0.67	4.66	35.85	0.02	8.45	8.97	3.20	38.18
Shoveler	0.10	6.35	32.97	0.01	36.97	16.71	5.31	1.58
Teal	1.58	7.01	27.47	0.08	17.37	2.42	20.25	23.82
Tufted Duck	28.48	1.95	8.10	0.19	51.48	5.55	2.00	2.25
Wigeon	1.61	3.32	35.11	0.20	12.54	0.67	13.36	33.19
Pintail	3.62	2.79	25.68	14.87	16.15	2.93	27.05	6.91
Mean	7.60	13.15	21.91	2.34	21.42	6.57	8.31	18.69

2.4 Discussion

This is the first study that has examined distribution patterns of wintering waterfowl across the whole of Great Britain. The results demonstrate the particular importance of weather patterns in determining waterfowl distribution. Previous studies have also found weather to have profound effects on waterfowl distribution (Ridgill and Fox 1990; Mallory et al. 2003). Temperature and rainfall can affect animal ecology in many ways; they can either influence a species' distribution directly through physiological effects or indirectly through its influence on resource distribution (Mallory et al. 2003). Indeed variables like temperature and rainfall are important for waterfowl, which are very sensitive to changes in weather conditions. Changes in weather conditions can trigger large scale movements in many species (Ridgill and Fox 1990) and can cause drops in the population size of many species (Newton 2007).

Not surprisingly larger water bodies were associated with occurrence of a number of species. This is in agreement with earlier studies which have shown that water body surface area is an important factor in determining waterfowl abundance and distribution (Tuite et al. 1984; Elmberg et al. 1994; Paracuellos & Telleria 2004, McKinney 2006). It is likely that bigger areas provide birds with more food supply and also tend to reduce edge effects which can increase competition and predation rates (Crozier & Niemi 2003).

The two geographic variables varied in their importance according to each species. The variation in species' occurrence explained by geography is normally considered to be a reflection of spatial dynamics and/or historical patterns of

dispersal in producing distributional patterns. This probably is probably the case for most The study species although it is also likely that the two geographic variables might have reflected the variation in habitat factors which were not considered here.

The interpretation of the effects of the landscape variables is less clear. These were generally unimportant in comparison to geography and climate and size of the water body suggesting landscape composition play a minor role in affecting waterfowl distribution. This finding is differs from the results of McKinney et al. (2006), who found that landscape composition play an important role in determining waterfowl distribution in winter. However McKinney's study was conducted at local scale, where landscape composition is more likely play an important role determining species distribution. The results demonstrate that at a larger scale climatic factors have an overriding influence on waterfowl distribution.

The majority of the models had an AUC value above 0.70 for most species, indicating that they were reasonably robust. The approach used here has been shown to be amongst the most accurate classifiers (Elith et al. 2006; Leathwick et al. 2006; Guisan et al. 2007 Baker et al. 2006; Bricklemyer et al. 2007; Cappo et al. 2005; Moisen et al. 2006). Despite this, only five out of The thirteen species models showed an above-average performance, so I may ask: why was this the case? I believe that this might be due to the lack of important predictors rather than model inadequacy. Rushton (2004) stated that successful species' distribution modelling depends on selecting of a suitable set of environmental variables.

Whilst there is no doubt that the variables selected for this study are relevant to waterfowl, there are also other factors which could have affected their distribution, including hunting pressure, water depth and wetland trophic status (Suter 1994; Holme and Clausen 2006). Unfortunately, none of these variables were available in a digital format for the whole of Great Britain at the time of this study.

The decline of many waterfowl species across the world, has led to a call for these animals to be used to monitor changes in wetland biodiversity (Green 1995; Gibbs 2000). Many studies have demonstrated how wetland loss and habitat modifications have caused declines in waterfowl populations (Fox et al. 1994; Duncan et al. 1999; Long et al. 2007). Despite the coarse nature of The analysis I believe that with some refinement (e.g. including other ecological factors that affect the distribution of these species) The models could be used to predict future changes in the distribution of waterfowl populations in Britain. Moreover The results highlight the importance of rainfall and temperature which suggests that long-term waterfowl monitoring data could be used to assess responses to projected climatic changes.

3 MODELS OF THE CLIMATE ASSOCIATIONS AND DISTRIBUTIONS OF AMPHIBIANS IN ITALY

3.1 Introduction

Analyzing the relationships between the distribution of animal species and climatic variables is not only important for understanding what factors govern species distribution, but also for improving The ability to predict future ecological responses to climate change (Donnelly, 1998; Teixeira & Arntzen, 2002; Thuiller et al., 2004; Parra-Olea et al., 2005; Araujo et al., 2006; Piha et al., 2007). Species distribution models have become an increasingly common method for describing the influence of current and future climate on the distribution of all vertebrate species (e.g. Guisan & Hofer, 2003; Peterson et al., 2002; Araujo et al., 2006; Peterson, 2003; Harrison et al., 2006; Levinsky et al., 2007). By parameterizing a model on current species distributions and climatic variables, it is possible to use the model to make predictions of future changes in distributions under various climatic scenarios (Hannah et al., 2002; Thuiller, 2003). These models can reveal species-specific responses to changes in climatic factors and increase The understanding of the processes controlling current and future species distributions.

In the context of global climate change, amphibians are of particular interest because of their extreme sensitivity to environmental stressors (Alford & Richards, 1999; Carey & Alexander 2003; Baillie et al., 2004; Stuart et al., 2004; Wake, 2007). Several studies have documented the impacts of climate change on

amphibians, including the impacts on their breeding phenology (Blaustein et al., 2001), disease induced mortality (Pounds et al., 2006), and long term population declines (Whitfield et al., 2007). However, whilst most of these studies have focussed on quantifying the impacts of climate change amphibian life history processes, there is a need to consider where the impacts will be greatest and where changes are likely occur.

Italy is a highly diverse country, with several climatic zones, geological substrates and vegetation regions. Few countries in Europe have as rich a herpetofauna as that of Italy, both in terms of the overall number of species and endemic species and in terms of biogeographical composition (Sindaco et al., 2006). However, up to date no empirical modelling analyses have been carried out on the distribution of amphibians in Italy. The only study where amphibian distribution has been modelled at the national scale was that of Maiorano et. al. (2006), using deductive models. Here I use a relatively novel modelling technique, Random forests (Breiman 2001), to model the distribution of amphibian occurrence in Italy in relation to climate. The specific aim of The study were: 1) to assess the importance of current climate in determining amphibian occurrence 2) to examine the potential changes in the distribution of amphibians under a $2xCO^2$ future climate scenario.

3.2 Methods

3.2.1. Species and climatic data

The CKmap (Check-list and distribution of the Italian Fauna) database provided the basis for the analysis presented in this paper (Ruffo & Stoch, 2005). The

CKmap project has been developed through an agreement between the Italian Ministry of Environment and the Natural History Museum of Verona and is aimed at bringing together and computerizing distributional data on all the species of the Italian fauna. The data regarding the distribution of amphibians contained within the database were originally provided to the Ministry of Environment by the Italian Herpetological Society (SHI). These consist of presence records for all thirty five amphibian species from 1705 10 x 10 km UTM squares. The data originate from a variety of sources including regional mapping projects, museum records, and a national survey conducted by the Italian Herpetological Society from 1994 to 2004 (Sindaco et al., 2006). For the purpose of the present analysis I used data for seventeen species for which there were at least 30 records. Because the species' records were not coming from a systematic survey, I assumed pseudo-absences i.e. if a species was not recorded in a square it was considered to be absent. I only used occurrence data collected from the 1980s onwards and I selected squares where a minimum collection effort was made (i.e. squares which had at least 3 species records). I deliberately excluded species that are cave-dwelling, exotic, or confined to small islands.

Climatic data were obtained from WORLDCLIM (version 1.3, http://www.worldclim.org) which is explained in detail in Hijmans et al., 2005. WORLCLIM contains climate data at a spatial resolution of 30 arc seconds (~1x1 km resolution) obtained by interpolation of climate station records from 1950–2000. A future climate scenario was also obtain from WORLDCLIM (http://www.worldclim.org/future.htm). This dataset, which comprise the same set

of variables available for the present climate data layers, is a downscaled version of the predictions of the CCM3 global climate model run under a 2 x CO^2 scenario (Govindasamy et al., 2003). For each of the variables, I used the mean value of all the pixels contained within each 10x10km square. A full list of the climatic variables used in the analyses is provided in Table 3.1.

Table 3.1. Characterization of the climatic variables used for the analyses.

Acronym	Variable
Clim1	Annual Mean Temperature
Clim2	Mean Diurnal Range
Clim3	Temperature Seasonality (standard deviation *100)
Clim4	Max Temperature of Warmest Month
Clim5	Min Temperature of Coldest Month
Clim6	Temperature Annual Range (P5-P6)
Clim7	Mean Temperature of Wettest Quarter
Clim8	Mean Temperature of Driest Quarter
Clim9	Mean Temperature of Warmest Quarter
Clim10	Mean Temperature of Coldest Quarter
Clim11	Annual Precipitation
Clim12	Precipitation of Wettest Month
Clim13	Precipitation of Driest Month
Clim14	Precipitation Seasonality (Coefficient of Variation)
Clim15	Precipitation of Wettest Quarter
Clim16	Precipitation of Driest Quarter
Clim17	Precipitation of Warmest Quarter
Clim18	Precipitation of Coldest Quarter

3.2.2 Analyses

The random forests algorithm Breiman (2001) is a new entry in the field of data mining and is designed to produce accurate predictions that do not overfit the data. The algorithm is based on the well known methodology of classification trees (Breiman 1984). In brief, a classification tree is a rule partitioning algorithm, which classifies the data by recursively splitting the dataset into subsets which are as homogenous as possible in terms of the response variable (Breiman 1984). The use of such procedure is very desirable, as classification trees are non-parametric, able to handle non linear relationships and can deal easily with complex interactions. Random forests uses a collection (termed ensemble) of classification trees for prediction. This is achieved by constructing the model using a particularly efficient strategy aimed at increasing the diversity between the trees of the forest.

Random Forests is built using randomly selected subsets of the observations and a random subset of the predictor variables. Firstly, many samples of the same size as the original dataset are drawn with replacement from the data. These are called bootstrap samples. In each of these bootstrap samples about 2/3 of the observations in the original dataset occur one or more times. The remaining 1/3 or so of the observations in the original dataset that do not occur in the bootstrap sample are called out-of-bag (OOB) for that bootstrap sample. Classification trees are then fit to each bootstrap sample. At each node in each classification tree, only a small number (the default is the square root of the number of observations) of variables are available to be split on. This random

selection of variables at the different nodes ensures that there is a lot of diversity in the fitted trees, which is needed to obtain high classification accuracy. Each fitted tree is then used to predict for all observations that are out-of-bag for that tree. The final predicted class for an observation is obtained by majority vote of all the predictions from the trees for which the observation is out-of-bag. Several characteristics of random forests make it ideal for data sets that are noisy and highly dimensional datasets. These include its remarkable resistance to overfitting and its immunity to multicollinearity among predictors.

The output of random forests depends primarily on the number of predictors selected randomly for the construction of each tree. After trying several values I decided to use the default number suggested by Breiman for classification problems. I made this choice as did not notice any decrease in the out of bag error estimate after trying several values. In order to measure the importance of each variable in used measure of importance provided by Random Forests, based on the mean decrease in the prediction accuracy Breiman (2001). The mean decrease in the prediction accuracy is calculated as follows: Random Forests determines the importance of a predictor variable by calculating the increase in prediction error when the OOB observations for that variable undergo permutation while all other predictor variables are unchanged (Liaw & Wiener, 2002). The importance of all the variables of the model is obtained when the aforementioned process is carried out for each predictor variable. All the analyses were carried out using the randomForest package in R (Liaw and Wiener 2002).

3.2.3 Model evaluation and predictions

In order to evaluate the models, the original data set were randomly split into model training (70%) and model evaluation data sets (30%). The training dataset was used for the construction of the model whereas the evaluation data set was used to test the predictive abilities of The models. The discrimination ability of the models was assessed using receiver operating characteristic (ROC) plot analysis. This technique measures the association between the presence and pseudo-absence records by using and calculating the area under the curve (AUC) (Fielding & Bell, 1997). AUC relates relative proportions of correctly classified (true positive proportion) and incorrectly classified (false positive proportion) cells over a wide and continuous range of threshold levels, which makes it a threshold-independent measure (Fielding and Bell, 1997). The AUC values range from <0.5 for models with no discrimination ability to 1 for models with perfect discrimination. An approximate for classifying the accuracy of the AUC is that proposed by Swets (1988):0.90-1.00 = excellent; 0.80-0.90= good; 0.70-0.80=fair; 0.60-0.70=poor; 0.50-0.60=fail. In The case, if a model had an AUC value of at least 0.70 it was considered to be validated.

When predicting species future distributions considered two different dispersal scenarios. Firstly, assumed unlimited dispersal, such that the future distribution is the entire area predicted by the model; secondly, assumed no dispersal, whereby the future distribution is the overlap between current and future predicted distributions. quantified changes in the occupancy of a species under present and future climate conditions, by transforming the probability of

occurrence from models into presence-absence. did this using the probability threshold that maximised model performance as measured by Cohen's kappa (Manel et. al 2001). finally calculated the gains in the number of species under unlimited dispersal scenario and losses in the number of species under a no dispersal scenario.

3.3 Results

All the models performed quite well when tested again the validation data (see Table 3.2). I considered three classes of model accuracy based on those proposed by Swets (1988). The fair accuracy class (0.70 < AUC < 0.80) included two species, *Rana dalmatina* (0.76) and *Hyla intermedia* (0.77). The good accuracy (0.70 < AUC < 0.80) high (0.9 < AUC < 1) classes included five included eleven species, respectively. The three best modelled species were *Salamadra atra, Discoglossus pictus* and *Hyla sarda*.

Table 3.3 shows the five most important predictor variables, as obtained from the mean decrease in accuracy permutation procedure. The most influential predictors were related to temperature, which were of primary importance to eleven species (*Bombina variegata*, *Bufo viridis*, *Hyla intermedia*, *Rana esculenta*,*Rana italica*, *Rana latastei*, *Rana temporaria*, *Salamandra atra*, *Salamandra salamandra*, *Salamandrina terdigitata*, *Triturus alpestris*). Precipitation variables, on the other hand, were quite important for seven species (*Discoglossus pictus*, *Discoglossus sardus*, *Hyla sarda*, *Rana dalmatina*, *Pelobates fuscus*, *Triturus italicus*, *Triturus vulgaris*). The temperature related variables that more frequently selected among the top five variables maximum

temperature of the warmest month (Clim4), mean temperature of wettest quarter (Clim7) and mean temperature of driest quarter (Clim8). The precipitation variables which more frequently selected among the top five variables were precipitation of the warmest quarter (Clim17), precipitation of the driest quarter (Clim16), precipitation seasonality (Clim14) and precipitation of the coldest quarter (Clim18).

Table 3.2 Results of model validation and predictions for present climate and future climate scenarios. The results show the AUC value for each species and the number occupied grid cells by each species for present and future climate conditions under the assumptions of both unlimited and no dispersal.

Species	AUC	Predicted occupancy-Present	Occupancy-future no dispersal	Occupancy-future unlimited dispersal
Bombina variegata	0.88	573	102	204
Bufo viridis	0.83	1448	1302	2301
Discoglossus pictus	0.99	216	212	282
Discoglossus sardus	0.98	178	39	60
Pelobates fuscus	0.91	32	0	0
Hyla intermedia	0.77	1374	1252	2420
Hyla sarda	0.99	162	25	32
Rana dalmatina	0.76	1169	650	1181
Rana esculenta	0.84	2481	845	2893
Rana italica	0.91	904	399	507
Rana latastei	0.96	252	104	238
Rana temporaria	0.98	630	440	453
Salamandra atra	0.98	184	123	151
Salamandra salamandra	0.91	890	497	606
Salamandrina terdigitata	0.87	360	27	65
Triturus alpestris	0.95	547	347	480
Triturus italicus	0.94	663	439	529
Triturus vulgaris	0.84	1149	1032	1625

Table 3.3. Five most important predictor variables for each species. The names of the variables are listed in decreasing order of importance (for acronyms see Table 1) . The number between the brackets indicates (eg. 1.17) the mean decrease in accuracy, which represents the overall percentage mean decrease in the prediction error of the model when one particular variable is permuted while all the other variables are held constant. Higher values of mean decrease in accuracy indicate variables that are more important in determining a species' distribution.

Bombina variegata	*Bufo viridis*	*Discoglossus pictus*	*Discoglossus sardus*	*Hyla intermedia*	*Hyla sarda*
Clim7 (1.17)	Clim6 (1.12)	Clim16 (0.99)	Clim13 (0.71)	Clim8 (1.08)	Clim7 (0.71)
Clim17 (1.11)	Clim9 (1.12)	Clim17 (0.81)	Clim7 (0.65)	Clim17 (1.08)	Clim13 (0.63)
Clim16 (1.09)	Clim17 (1.1)	Clim14 (0.74)	Clim18 (0.63)	Clim4 (1.06)	Clim14 (0.59)
Clim9 (1.08)	Clim4 (1.09)	Clim13 (0.74)	Clim16 (0.58)	Clim9 (1.06)	Clim16 (0.54)
Clim4 (1.03)	Clim3 (1.09)	Clim18 (0.71)	Clim14 (0.56)	Clim3 (1.05)	Clim18 (0.52)
Pelobates fuscus	*Rana dalmatina*	*Rana esculenta*	*Rana italica*	*Rana latastei*	*Rana temporaria*
Clim3 (0.78)	Clim12 (1.09)	Clim4 (1.15)	Clim7 (1.26)	Clim8 (0.93)	Clim7 (1.06)
Clim18 (0.78)	Clim17 (1.08)	Clim9 (1.14)	Clim18 (1.07)	Clim13 (0.93)	Clim17 (1.05)
Clim12 (0.77)	Clim11 (1.07)	Clim10 (1.07)	Clim17 (1.05)	Clim7 (0.9)	Clim4 (1.05)
Clim11 (0.76)	Clim15 (1.05)	Clim1 (1.05)	Clim8 (0.98)	Clim3 (0.9)	Clim8 (1.03)
Clim16 (0.75)	Clim16 (1.03)	Clim7 (1.03)	Clim3 (0.97)	Clim6 (0.9)	Clim9 (1.01)
Salamandra atra	*Salamandra salamandra*	*Salamandrina terdigitata*	*Triturus alpestris*	*Triturus italicus*	*Triturus vulgaris*
Clim5 (0.79)	Clim4 (1.15)	Clim7 (0.86)	Clim7 (1.14)	Clim17 (1.01)	Clim6 (1.09)
Clim8 (0.77)	Clim9 (1.12)	Clim17 (0.85)	Clim8 (1.06)	Clim18 (0.99)	Clim14 (1.05)
Clim10 (0.71)	Clim1 (1.08)	Clim3 (0.84)	Clim9 (1.05)	Clim7 (0.97)	Clim16 (1.05)
Clim11 (0.61)	Clim17 (1.06)	Clim9 (0.82)	Clim4 (1.04)	Clim11 (0.95)	Clim17 (1.03)
Clim1 (0.60)	Clim8 (1.03)	Clim8 (0.8)	Clim14 (1.01)	Clim13 (0.94)	Clim4 (1.01)

Projection of niche models onto predicted future climate scenario showed a moderate to extreme spatial decrease in amphibian distribution. Table 2 shows the number of occupied cells, in relation current and future climate, under the two dispersal scenarios (unlimited dispersal or no dispersal). All species showed a decrease in their distribution under a scenario of no dispersal and twelve species out eighteen showed a decrease in their distribution under an unlimited dispersal scenario. Six species showed an increase in their distribution under the unlimited dispersal scenario. One species, *Pelobates fuscus*, was predicted to loose 100% of its range under both scenarios of unlimited and no dispersal.

An example of the present and future bioclimatic profiles for three species species is shown in Fig 3.1. The future projections consider the two dispersal scenarios of unlimited and no dispersal. The maps show the observed and predicted distributions for *Bombina variegata*, a toad distributed in the Alps and Appenines, *Rana latastei*, a frog species occurring the Po Valley in Northern Italy, and *Salamandrina terdigitata* an endemic Salamander species occuring in the Appennines in central and Southern Italy. The future projections show that under both unlimited and no dispersal there could be a contraction in the distributions of the three species. Fig 3.2. shows species losses and under the assumption of no dispersal and species gains under the assumption of unlimited dispersal . Most species losses are predict to occur in Southern Italy and Central Italy. In contrast the Alps show a lower species loss together with Northern Italy. With regard to species gains, most them gains were predicted to occur in the Appenines in central Italy and the Northermost parts of the Alps. Moderate

species gains were also predicted to occur along near coastal areas of Eastern Italy.

Figure 3.1. Examples of the projected potential ranges for three amphibian species (*Bombina variegata*, *Rana latastei*, *Salamandrina terdigitata*): (a) current, (b) projected future with unlimited dispersal, and (c) projected future without dispersal.

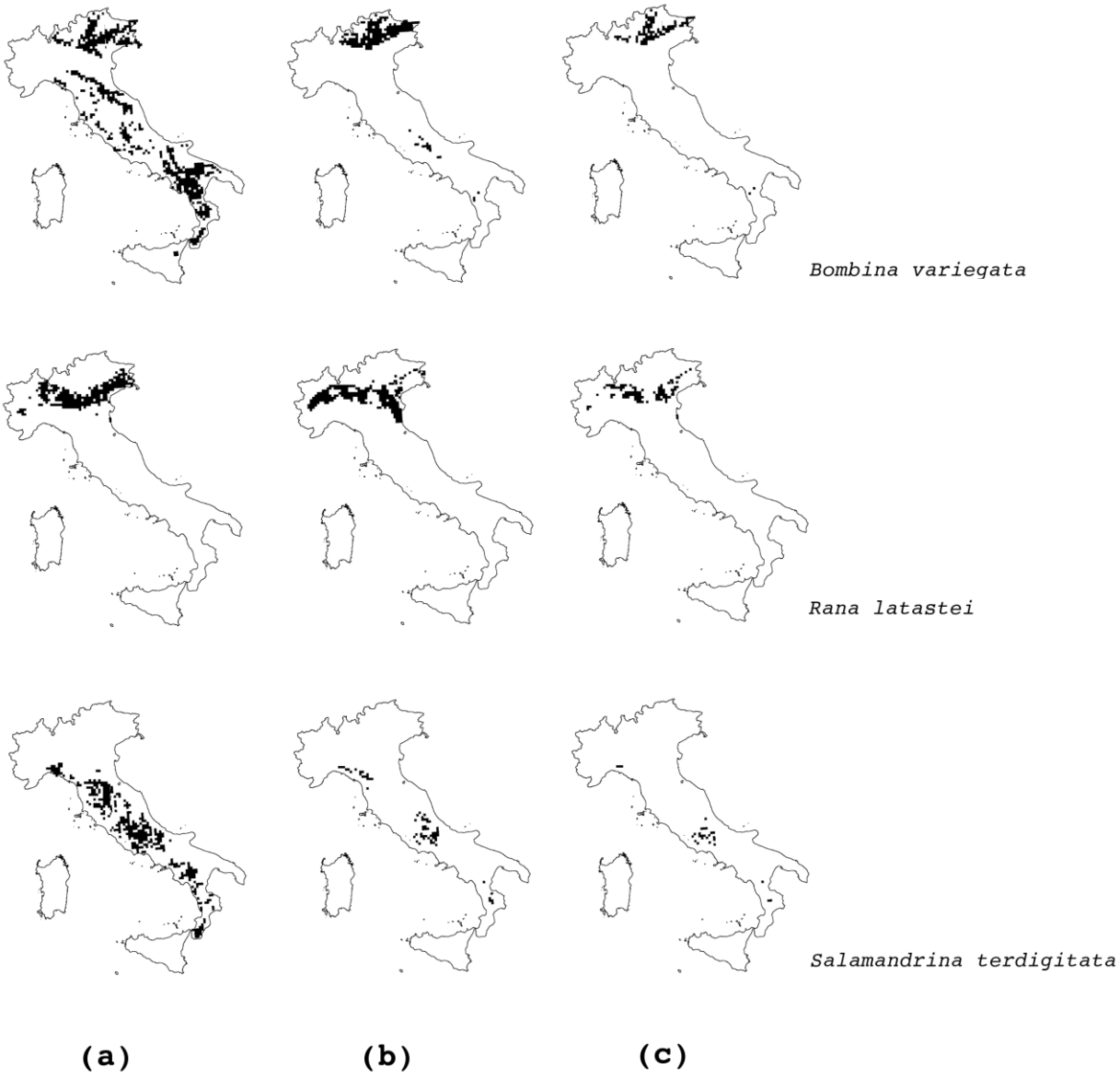
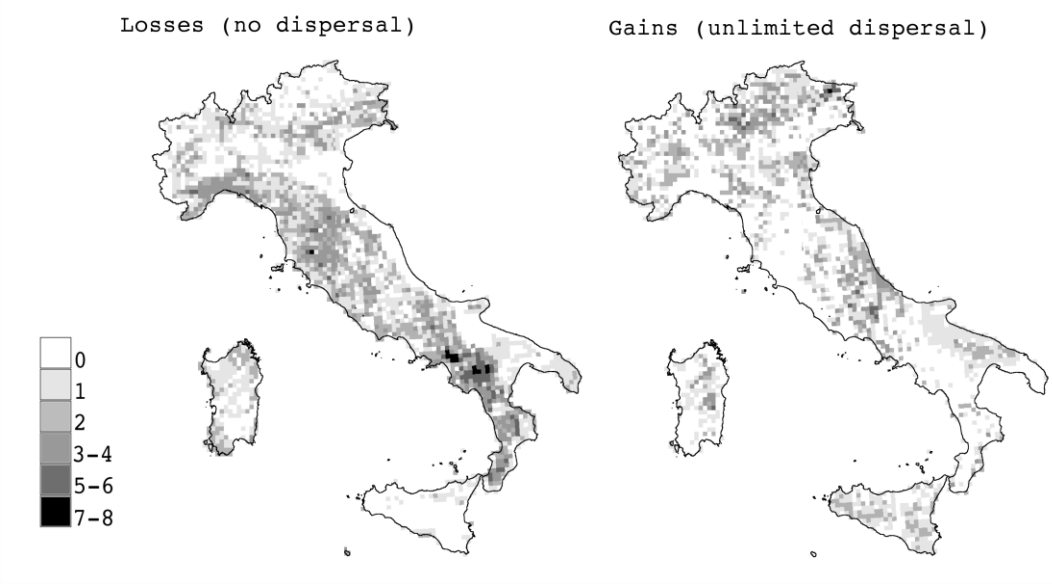


Figure 3.2. Projected amphibian species losses (with no dispersal) and gains (with unlimited dispersal)



3.4 Discussion

Climate has profound effects on species distribution and The results show that this also applies to the amphibian species considered in this study. Their current distribution seems to be largely determined by temperature related factors, which are indicators of physiological requirements for survival and availability of suitable habitat. Precipitation related variables played a major role in determining

the distribution of species living in Mediterrean areas (eg. *Hyla sarda*, *Discloglossus spp*.), indicating that for these latter species "water availability" may be a major of a limiting factor. However, it should be noted that, the climatic data are inherently spatially autocorrelated and it is therefore difficult to interpret the importance of each variable separately. It is instead more likely that groups of variables are acting together to influence the presence or absence of each species.

The projected bioclimatic response under the $2xCO^2$ climate scenario showed considerable effects on the distribution of amphibians. The results showed that the all of the species could potentially undergo large range reductions. In particular predictions for mountain (*Bombina variegata, Rana temporaria*, *Salamandra salamandra, Salamandrina terdigitata, Salamandra atra, Triturus vulgaris, Rana italica)* and Mediterranean species (*Discoglossus sardus, Hyla sarda, Triturus italicus*) showed that distributions of these species could potentially decrease, regardless the possibility for these species to occupy new suitable areas. This finding is not too surprising, as mountainous and Mediterranean areas host a number of highly specialized species with narrow climatic tolerances and any changes in temperatures could have detrimental effects on the distribution of these species. This result consistent with evidence from similar studies on other taxa, which have shown that climate change could have some severe effects on the biodiversity of mountainous and Mediterranean areas (Thuiller et. al, 2005; Levinsky et. al. 2007).

The models predicted noticeable changes in amphibian species richness. As a result of shifts in species ranges, future species-rich areas are predicted to

become relegated to mountain areas of northern and central Italy. contrasted two simple assumptions about changes in species richness of no dispersal or unlimited dispersal. However, the ability of species to occupy new sites will depend on the as well as on the existence of pathways for dispersal. The assumption of no dispersal is likely to be more realistic for amphibians, which are known to be poor dispersers (Smith & Green 2005). This would therefore support the hypothesis that climate change could have some serious negative effects on the amphibian distributions in Italy.

Although the results provide some clear evidence that climate change could have a negative effect on the distribution of amphibians, there are some sources of uncertainty which need considered in order to verify the accuracy of any conclusions generated by this study. This is one example of ecological niche-based modelling that is subject to a well documented number range of assumptions and caveats, as other authors have emphasized (Davis et al., 1998; Thuiller 2004; Araujo et al., 2005; Heikkinen et al., 2006; Hijmans & Graham 2006; Pearson et al., 2006). The results of the future predictions of niche based models cannot be taken as precise forecasts because of the uncertainties present in climate change scenarios and in the modelling techniques used. Model predictions rely not only on the accuracy of the bioclimatic models but also on the ability of the future scenario to depict the accumulation of the cumulative effects of CO^2 on climate. Furthermore the models do not take into account important factors like biotic interactions, which are important in determining the distribution of species under present and future climate conditions(Davis et al., 1998;Araujo & Luoto

2007; Brooker et al., 2007). Despite uncertainties, The findings provide illustration of the potential importance and the likely direction of climate effects on amphibian distributions in Italy.

The decline of many amphibian species throughout the world has led for a call for these animals to be used for monitoring environmental quality and change (Carey et al., 2001;Hopkins, 2007).Many studies have demonstrated the importance of the effects of climate and habitat changes in the decline of amphibian populations (Carey et al., 2001; Whitfield, 2007; Johnson et al., 2007; Nystrom, 2007, Boone et al. 2007). The study, emphasises the potential severity of climate change, which should be taken into account in conservation planning. Amphibian distributions are predicted to change quite dramatically and new areas may need to be protected in order to ensure the persistence of Italian amphibians. The results of this study provide some useful information, indicating if future conservation priorities for some species should be enhanced. Management options include the maintenance of a network of suitable habitats as well as the facilitation of the migration of amphibians to new climatically suitable areas. Detailed guidelines for conservation at the local scale should, however, be based on finer scale analyses. Ideally models for these finer scale analyses should incorporate factors like local population dynamics, landscape characteristics and human interference. However, the level of detail necessary to parameterise such models is so time-consuming and difficult to collect that it may prohibit such approaches.

4 USING BAGGING PREDICTORS AND LANDSCAPE ZONATION TO IDENTIFY IMPORTANT AREAS FOR BUTTERFLY CONSERVATION IN ITALY

4.1 Introduction

One of the most important aspects of any conservation strategy is the identification of high-value sites on the basis of their biodiversity content (Kati et al. 2004; Kelley et al. 2002; Margules and Pressey 2000; Margules et al. 1988). Site prioritization is useful for selecting reserve networks or devising management strategies based on the species composition of a set of sites. However, the limited amount of resources allocated for systematic data collection, makes it very difficult to obtain maximal representation of the overall biodiversity of a region. This is especially true for Mediterranean countries, for which information on the distribution of many species is often incomplete and data are lacking for many regions (Hortal et al. 2004; Ramos et al. 2001). Alternative methods for the identification of important sites for biodiversity conservation, which can have utility in the context of sparse data, are needed. In the last few years the combination of species distribution modelling and complementarity-based site selection algorithms has shown great promise towards achieving this goal (e.g. Ortega-Huerta and Peterson, 2004; Pawar et al. 2007; Moilanen et al. 2007).

Species distribution modelling aims at predicting species geographical distribution from occurrence records and environmental data layers (Guisan and Zimmermann 2000; Guisan and Thuiller 2005; Rushton et al. 2004).

Complementarity-based algorithms (Moilanen et al. 2005; Moilanen 2007) aim to find an optimal set of areas that are jointly as valuable as possible, by taking into account differences, similarities and connectivity between candidate sites.

The present study combines the use of species distribution modelling with a complementarity-based method to identify areas important for butterfly conservation in Italy. More specifically use two relatively novel techniques: bagging predictors and Zonation. Bagging predictors (Breiman 1996) belong to a general class of methods, based on classification trees. This class of methods, termed ensemble, include other techniques like Boosted Regression Trees (Friedman 2001) and Random Forest (Breiman 2001). Ensemble modelling (Araujo and New 2007) differs fundamentally from conventional techniques as it does not seek the single most parsimonious model, but aims to fit large number of models which are then combined together to make accurate predictions. Bagging has been shown to be a very promising technique in many areas of science where predictive modelling is required. (Carreiras et al. 2006; Lawrence et al. 2004; Myles et al. 2004; Radivojac et al. 2004; Rizzoli et al. 2002; Shah et al. 2007) However, up to date, bagging has been used only in one other ecological study (Prasad et al. 2006). Zonation is a site prioritization method introduced by Moilanen et al. (2005). This method produces a hierarchical prioritisation of a landscape based on the biological value of sites. The landscape is thus zoned according to its conservation potential and different degrees of protection can be applied to different zones. Zonation has been used successfully for the identification of conservation priorities in British Butterflies (Moilanen et al.

2005; Early and Thomas 2005) and the identification of important areas for fish in New Zealand (Moilanen et al. 2008)

Butterflies are an important component of biodiversity, and their sensitivity to environmental changes makes most species vulnerable to extinction (Thomas et al. 2006; Thomas et al. 2006; Roy et al. 2001; Maes & Van Dyck 2001; Menendez et al. 2007; Franco et al. 2006; Hill et al. 2002). Additionally several studies have demonstrated how the populations of a number of species have declined in many parts of Europe in relation to direct habitat modifications (e.g. van Swaay et al. 2006; van Swaay and Warren 1999; Thomas 1995; Wenzel el al 2006; Polus et al. 2007; Aviron et al. 2007). Given the importance of butterflies as indicators of the health of the whole ecosystem (Thomas 2005; Thomas and Clarke 2004) and their higher sensitivity to habitat changes in comparison to other popular bioindicator groups, (Thomas et al. 2004) there is an urgent need for the establishment of conservation priorities for butterflies in those areas which have the highest concentration of butterfly diversity.

Italy is of outstanding importance for butterflies as there are few countries in Western Europe which have as rich a butterfly fauna, both in terms of the overall number of species and the number of endemic species (Balletto and Kudrna 1985; Balletto 1992; Balletto et al. 2005a). The importance of Italian butterflies for the conservation of European butterfly fauna is probably best expressed by the fact that one from every two species native to Europe lives in Italy (Balletto and Kudrna 1985). Italy has thus both national and international obligations to conserve these species and the habitats upon which they rely.

Here, I constructed species-environment models using butterfly occurrence data derived from a national biodiversity database and used the results of the models to select areas of conservation importance. The specific objectives of The research were: (i) to identify areas important areas for butterflies across the Italian peninsula, (ii) to identify important areas by taking into account species of conservation concern (iii) to identify some possible management landscapes based on the similarity in species composition among sites.

Methods

4.2.1 Species data and environmental predictors

The Check-list and distribution of the Italian Fauna (CKmap) database provided the basis for the analysis of this research. The CKmap project has been developed through an agreement between the Italian Ministry of Environment and the Natural History Museum of Verona and is aimed at bringing together and computerizing distributional data on all the species of the Italian fauna. The data regarding butterflies used for these analyses consist of 59130 presence records for 272 species of butterflies (in sensu Balletto et al. 2005b). For the purpose of the present analysis only considered those species which were also listed Red Data Book of European Butterflies (Swaay and Warren 1999). only used occurrence data collected from the 1980s onwards. Species with less than 5 records were discarded and selected squares where a minimum collection effort was made (i.e. squares which had at least 20 species records). The final database that was used for the analyses comprised occurrence records for 232 species distributed across 670 10x10km squares. Theses squares make up about 20% of the total number of

squares of the grid used to map the species distributions (3356 squares). A full list of the species names is given in APPENDIX 1. Because the species records were not coming from a systematic survey, assumed pseudo-absences i.e. if a species was not recorded in a square it was considered to be absent.

Nineteen environmental were used for the analyses (Table 4.1). Annual mean temperature and total precipitation data were obtained from Agency for Environmental Protection and Technical Services (APAT, http://www.apat.it/). National climate maps were created using smoothing splines (Hutchinson 1991). Three altitude and one slope variable were derived from a digital elevation model (DEM) (http://srtm.csi.cgiar.org/). Nine land cover types were derived from a digital CORINE data base (EEA 2000). The baseline resolution of all the environmental layers was 100 m (the lowest possible resolution for the CORINE Land Cover map, which was the layer with the coarser spatial resolution). The DEM, which had an original resolution of 90m, was resampled to obtain a pixel size of 100m. All the environmental layers were aggregated to match the resolution of the species data (10x10km). For each 10km square therefore calculated: the mean value of all the pixels for each of the climatic, slope and altitude variables, the minimum and maximum value of all the pixels for the altitude variable, and the area (ha) of each the nine land cover types.

Table 4.1. Characterization of environmental and climatic variables used in the analyses.

Variable description
Yearly total amount of precipitation (mm)
Total amount of precitation for January (mm)
Total amount of precitation for July (mm)
Mean slope of each square (º)
Mean annual temperature (ºC)
Mean temperature of January (ºC)
Mean temperature of July (ºC)
Maximum altitude (m.a.s.l.)
Minimum altitude (m.a.s.l.)
Mean altitude (m.a.s.l.)
Area of urban development (ha)
Area of arable land (ha)
Area of broad leaved forest (ha)
Area of coniferous forest (ha)
Area of sparse vegetation (ha)
Area of grassland and pastures (ha)
Area of moorland (ha)
Area of marshes and bogs (ha)
Agricultural areas with a significant portion of natural vegetation (ha)

4.2.2 Species distribution modelling and model evaluation

I constructed species-environment models using tree-based classification models with bootstrap aggregation or bagging. Classification trees (Breiman et al., 1984) consist of recursive partitions of the dimensional space defined by the

predictors into groups that are as homogeneous as possible in terms of the response. The tree is built by repeatedly splitting the data into two exclusive groups, defined by a simple rule based on a single explanatory variable at each step. Bagging uses an ensemble (i.e. a collection) of classification trees for prediction (Breiman 1996). The idea underlying bagging is the recognition that part of the output error in a single regression tree is due to the specific choice of the training data set. Therefore, if several similar data sets are created by resampling with replacement (i.e. bootstrapping) and classification trees are grown without pruning and averaged, the variance component of the output error is reduced (Breiman 1996). When a bootstrap resample is drawn, about 37% of the data is excluded from the sample, but other data are replicated to bring the sample to full size. The portion of the data drawn into the sample in a replication is known as the ‘‘in-bag’’ data, whereas the portion not drawn is the ‘‘out-of-bag’’ data. The major advantage in using bagging is that the final model will always have an improved predictive performance.

The most important tuning parameter for bagging trees is the number of bootstrap replicates, hence the number of trees. Breiman (1996) suggested that a number of trees higher than 25 tend not to produce a significant test set error reduction. In The case, when constructing The bagging models combined 50 trees. All the analyses were carried out using the ipred package (Peters et al. 2002) for R (Ihaca and Gentleman 1996).

In order to evaluate the performance of models the original data set was randomly divided into model training (70%) and model evaluation data sets

(30%). The discrimination ability of the models was assessed using receiver operating characteristic (ROC) plot analysis (Fielding and Bell 1997). This technique measures the association between the presence and pseudo-absence records by using and calculating the area under the curve (AUC). AUC relates relative proportions of correctly classified (sensitivity) and incorrectly classified (specificity) cases over a wide and continuous range of threshold levels, which makes it a threshold-independent measure. The AUC values range from <0.5 for models with no discrimination ability to 1 for models with perfect discrimination. An approximate for classifying the accuracy of the AUC is that proposed by Swets (1988): 0.90-1.00 = excellent; 0.80-0.90= good; 0.70-0.80=fair; 0.60-0.70=poor; 0.50-0.60=fail. Models with an AUC value of at least 0.70 were considered to be sufficiently accurate to be used in further analyses

4.2.3 Zonation and management landscapes

The results of all the models which showed a positive validation were subsequently used for the Zonation. also ran the Zonation with the results from all the models, irrespective of their accuracy. The aim of this analysis was to explore the effect of eliminating potentially important species from the final solution.

A Zonation analysis produces a hierarchical prioritisation of the landscape based on the biological value of sites (Moilanen et al. 2005; Moilanen 2007). The algorithm proceeds by removing least valuable cells (here 10x10 km) in a landscape while minimizing the loss rate of biodiversity and connectivity. The

order of cell removal gives a landscape zoning with most important areas remaining last. The output of the Zonation analysis is the ranking of each site, allowing for the identification of the most important areas for species persistence when a certain proportion of the land surface remains. Landscape Zonation can be initialized using different removal rules, depending on the conservation planning goals. used the area-core Zonation removal rule which aims to minimize biological loss by trying to retain the core areas of each species until the end of cell removal. regarded this removal rule as more appropriate because The aim was to establish important areas for conservation for all species and emphasising the locations with the highest occurrence levels of species, rather than focussing on species rich areas.

An important aspect that needs to be taken into consideration when planning any conservation strategy is that not all species are of equal value in terms of their conservation importance. Zonation can take this into account by using a species weighting procedure, which stresses the selection of high-value cells towards species of conservation concern (Arponen et al. 2004). used a weighting scheme similar to the one used by Early and Thomas (2007), based on the threat categories listed in the European Red Data Book of Butterflies (van Swaay and Warren 1999). assigned a weight of 1 to species not classified at risk, a weight of 2 to classified as lower risk species, a weight of 3 to species classified as vulnerable, a weight of 4 to species classified as endangered, and a weight of 5 to species classified as critically endangered. After carrying out the baseline analyses compared all the weighted and the unweighted solutions and determined

the degree of overlap in the top 10% fraction of the cells selected by each solution.

used the top 10% fraction for both weighted and non-weighted solutions to classify spatially separate clusters of 10x10km squares into different management landscapes. This method allows for the identification of management landscapes based on the distance and similarity in species composition between sites (Moilanen et al. 2005). The identification of landscapes requires the specification of: 1) the maximum distance allowed between cells that are included in the same landscape 2) the maximum difference in species composition between two cells to be joined in the same landscape, and 3) an inclusion minimum which determines how highly ranked cells must be included in each of the management landscapes. Previous studies for which high resolution data ($1km^2$) were available, used a maximum distance between cells of 10km, based on the assumption the maximum colonization distance of intermediate mobile species is less than 10km (Moilanen et al. 2005; Early and Thomas 2007). Because the coarse resolution of The dataset, decided to set the maximum distance between cells to a value of 30km (i.e. 3 cells).

specified maximum difference in species composition of 0.3. This indicates that at the most three species out of the out of ten will not be included in those cells (Moilanen et al. 2005). Finally defined as an inclusion minimum of 2 indicating that a management landscape could only be retained in the final solution if it contained one or more cells that were in the top-ranked 2% of cells.

All of the analyses were carried out using the Zonation software version 1.0 (http://www.helsinki.fi/bioscience/consplan/).

4.3 Results

Models for 182 species were sufficiently accurate (AUC ≥ 0.70) to be used in the zonation analyses. The graph shown in Fig. 4.1. relates model performance with the log of the total number of records for each species and the AUC of each model. Model performance showed a negative correlation with total the number of records for each species (R^2=0.10 $P < 0.001$). Models with a poor performance were generally related to species with a high number of records.

Figure 4.1. Relationship between model performance (AUC) and the number of species records records (R^2 0.10). Each point represents a species.

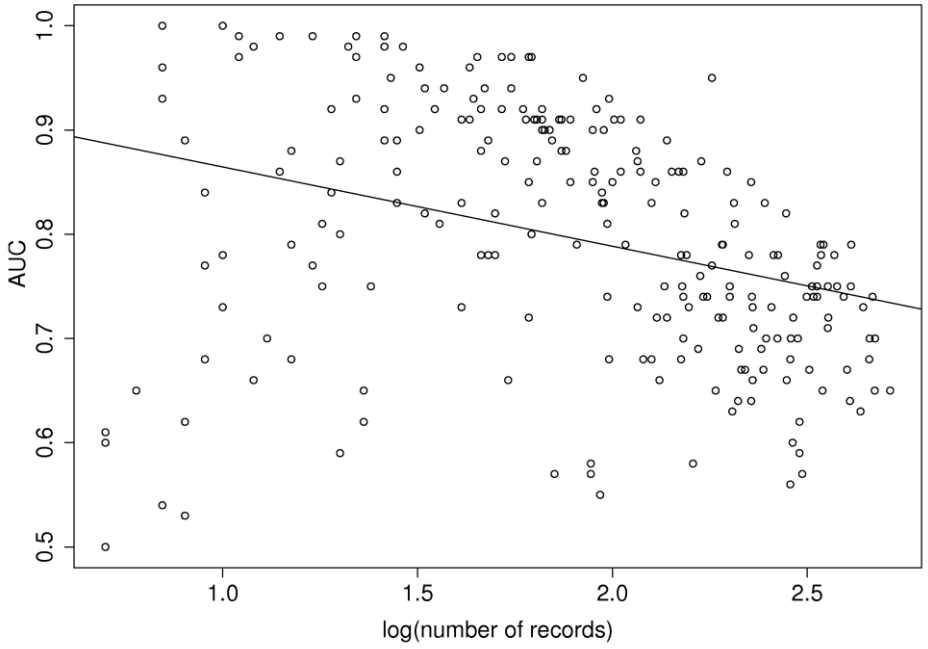


Figure 4.2. shows the results of the basic Zonation, with no species weighting, using the results of the 182 models which were sufficiently accurate (AUC $\geq$ 0.70). This solution highlighted how both Eastern (A) and Western (B) Alps are the two areas with the highest biological value. These were followed by the central Appennines (C) the Apulia region (D) and Sardinia (E). Figure 4.3. shows the results of the weighted Zonation using the species weighting scheme.

The main difference in the weighted and non-weighted solutions was in the selection high-value cells within the Alps (A) in Calabria (B) and in Sicily (C). These three areas were particularly important in the weighted solution when compared with the non-weighted solution. When compared with the top 10% fraction of the non-weighted and weighted solutions (Fig. 4.4) there was an overlap of 60% of the squares selected.

Figure 4.5 shows performance curves for both non-weighted and weighted solutions. The curves clearly show that when weighing was applied the average protection for species of conservation concern was increased. When comparing the top 10% fraction of the solutions using the 182 models which were successfully validated and the solutions using all 232 models, obtained an overlap of 76% of 86% for the non-weighted and weighted solutions, respectively.

Figure 4.2. Results of the basic Zonation, without species weightings. Map shows the results of the ranking for each cell, which varies from 0 to 1. Values closer to 1 are shown in red and represent the cells with a higher biological value, whereas cells with a value closer to 0 are shown in green and have a lower biological value. Arrows indicate: A (Western Alps) B(Eastern Alps) C (Central Appennines) D (Apulia) E (Sardinia).

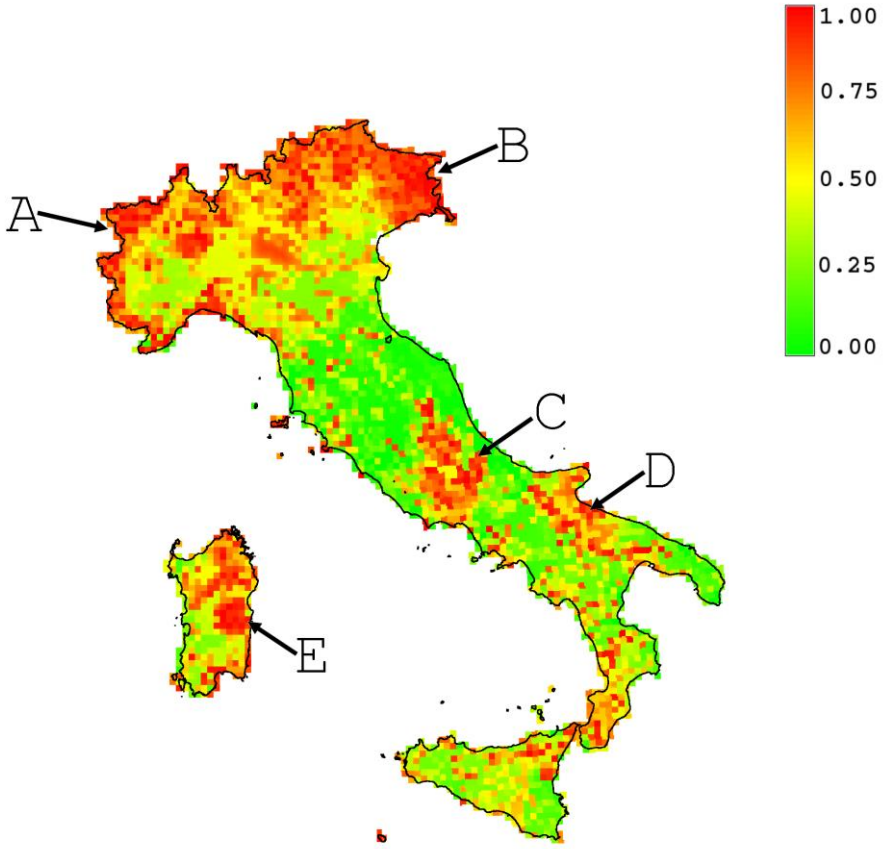


Figure 4.3. Results of the Zonation when species weighting was applied. Arrows indicate: Alps (A), Calabria (B) and Sicily (C).

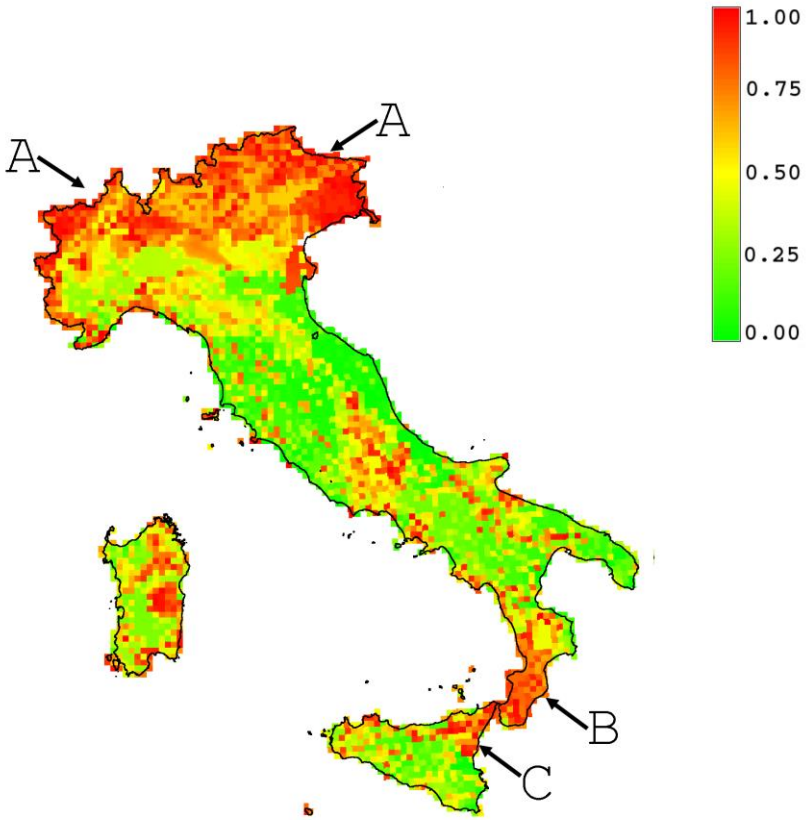


Figure 4.4. Comparison of the top 10% fraction of cells selected by the two, non-weighted and weighted solutions. The green squares indicate the cells which were selected in both solutions, the orange ones those selected only in the non-weighted solution, and the red ones the ones selected only in the weighted solution.

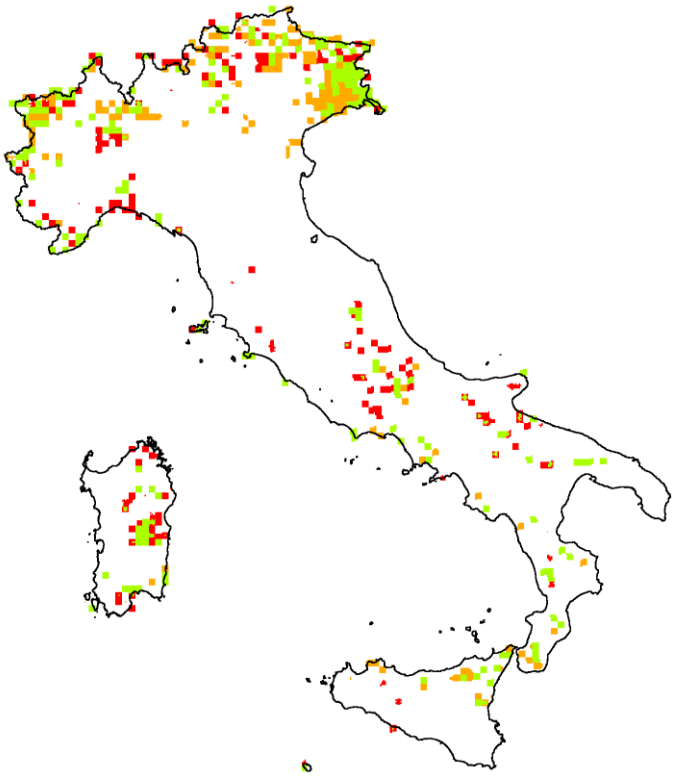


Figure 4.5. Average proportion of the original distribution retained for the species of conservation concern as a function of proportion of landscape remaining as lower priority zones. The two different lines show the average of the proportions for the non-weighted (solid line) and weighted solutions (dashed line).

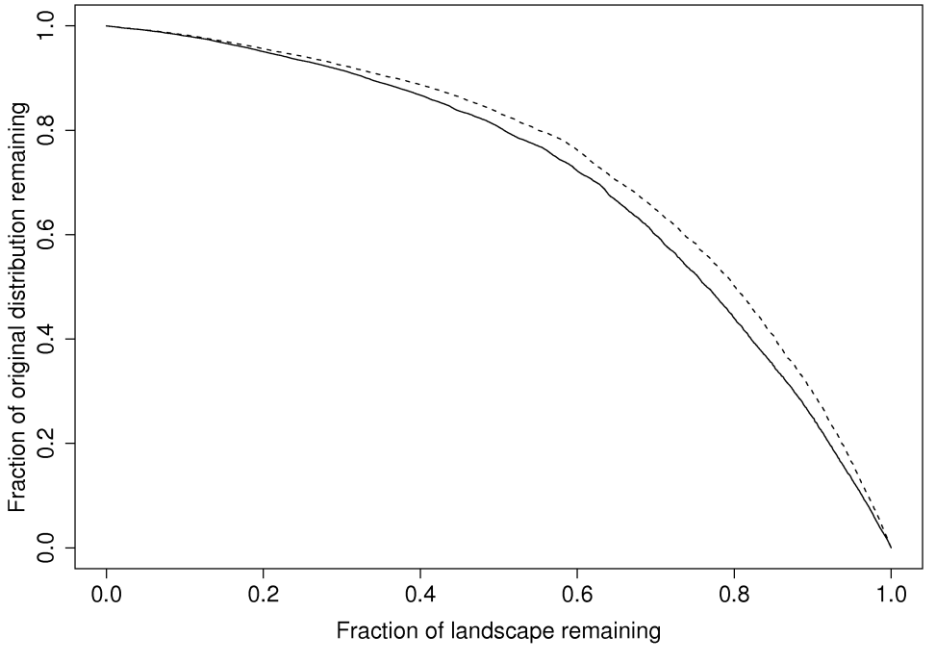
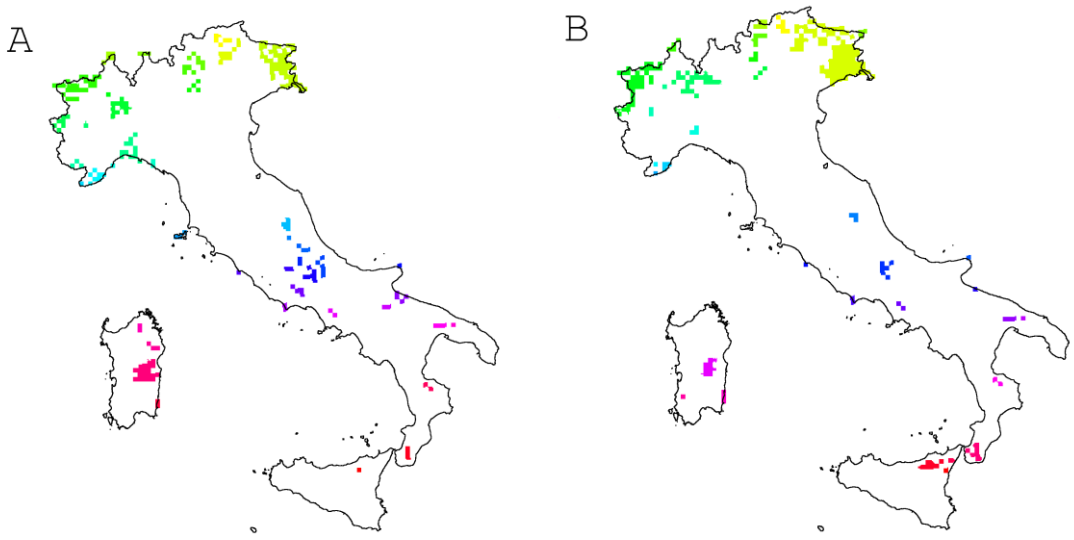


Fig. 4.6 shows the management landscapes selected using the landscape identification procedure for both non-weighted (A) and weighted (B) solutions. Forty four landscapes for the non-weighted solution and thirty four landscapes were identified for the weighted solution. Both of the solutions identified similar landscapes based on species composition. These can be broadly characterized into

Alpine landscapes in the eastern and Western Alps, landscapes occurring in central Italy and the Appennines and Southern Italy .

Figure 4.6 Priority landscapes derived from both the non-weighted (A) and weighted (B) solutions. The landscapes were selected using the top 10% fraction of cells selected by the Zonation. Each landscape (shown by a colour) contains blocks of land that are close together, similar in species composition, and contain a core area present late in the cell removal process.



4.2 Discussion

This study represents the first analysis aimed at identifying important areas for butterfly conservation in Italy. Previous research (Balletto and Kudrna 1985; Balletto 1992) focussed on assessing the conservation status of butterflies either regionally or nationally, but never aimed at setting conservation priorities for this

group in a spatially explicit context. produced a hierarchy of priorities using probabilities of occurrence derived from niche models. Despite the fact that not all species were included in the analysis, The results clearly demonstrate the strength of combining species distribution modelling and Zonation for the exploration of realistic scenarios for biodiversity protection over an extensive geographic area.

The results of the basic Zonation analyses (without species weights) showed that the most important areas for butterflies are the Alps, the central Appennines, the Apulia region, and the island of Sardinia. The high species diversity and the number of species with a restricted distribution occurring in these areas make up for the majority of Italian butterfly species. The Alps are a well known hotspot, hosting 106 species, more than thirty eight percent of the total number of species occurring in Italy (Tontini et al. 2003). The Alps also host the majority of species of conservation concern, with fifteen out of the thirty species of conservation concern included in this study. Amongst these, ten species are vulnerable and one is considered endangered. The Appennines are also a stronghold for many species, hosting a total 64 species (Tontini et al. 2003). Sardinia is also important. Because of its isolation and geological history, this island hosts a significant number of species of conservation concern and species with a narrow distribution (Dapporto and Dennis 2007).

The introduction of species weights into the Zonation analyses altered the spatial distribution of a fair proportion of the high value sites. More specifically, the weighted solution highlighted again the importance the Alps together with the Calabria region and Sicily. This finding, thus confirms that these areas are

important for the majority of species of conservation concern. The Alps, for instance, host eleven out of the twenty six species of conservation concern included in this study. Moreover the inclusion of weights for species typical southern species like *Anthocharis damone,* placed a greater emphasis in the selection of high value cells in Sicily and Calabria (see Fig 3). Species weighting was thus beneficial for identification of important areas for species of conservation concern.

The selection of the management landscapes identified a series of landscapes on the basis of the similarity of species composition. As highlighted by Moilanen et al. (2005) the main purpose of this analysis is not to propose a reserve structure but to identify landscapes that could be subjected to more detailed planning. However The results could be important within the context of butterfly management in different parts of the Italian peninsula. The procedure identified sites that provide protection for a full range of species, including Alpine and Appenninic and Mediterranean species. Unfortunately, the resolution of the species data did not allow us to make any comparison with the existing network of protected areas or introduce any aggregation between the cells. The design of any networks of protected areas would require a detailed analysis using high resolution data which goes beyond the one presented here.

Not all species were included in the Zonation analysis because of the relatively poor performance the models. However, a comparison of the solutions with all the 232 species and the 182 species for which model validation was adequate showed that the top 10% fraction of the cells selected by the two

solutions agreed in 76% and 86% for the non-weighted and weighted solutions, respectively. This result suggests that even with the exclusion of 50 species, The results were still robust, thus supporting the generality of The analyses. Whilst acknowledge that excluding a proportion of the species from the analyses might have been a serious limitation of The approach, it must also be borne in mind that conservation decisions cannot be delayed until biodiversity complete surveys are available, as this will greatly diminish the efficacy of any conservation strategies (Maiorano et al. 2006; Margules and Pressey 2000) The approach used here is far from being the ultimate solution, but believe that the problem of rare species could be taken into account by integrating it with other methods like deductive modelling, (eg. Maiorano et al. 2006; Romero-Calcerrada and Luque 2006; Rondinini et al. 2005; Dayton and Fitzgerald 2006).

Significant changes have occurred to the Italian landscape in the last forty years, as a consequence of the abandonment of agricultural land, changes in forestry policy, urban development processes, causing a profound effect on the territorial mosaic (Falcucci et al. 2007). These changes in landscape heterogeneity have had a significant effect on the biodiversity (Falcucci et al. 2007), including butterflies (Balletto 1992).

Many species of butterflies are highly dependent on man made biotopes such as dry grassland and meadows, which are typically, maintained by traditional forms of farming management such as livestock grazing and hay-making (Tontini et al. 2003; Van Swaay et al. 2006). The abandonment of traditional farming practices and livestock overgrazing are potential threats for many butterfly species

in the Alps and the Appennines. The loss of suitable habitats threatens the populations of many hygrophilous species living in the Padano Venetian plains. Moreover the development of tourist resorts along the Mediterranean coastline could have affected several species in the Mediterranean part of Italy. A wide range of factors threaten important habitats for butterflies, and the future conservation of many populations will depend on the conservation of these habitats (van Swaay and Warren 1999; van Swaay 2006; van Swaay and Warren 2006).

Unfortunately there is currently no monitoring scheme for butterflies in Italy and it is therefore difficult to relate any population changes to the potentially important habitat changes discussed above. Although the decline of some species seems to mirror what is happening to other populations in Europe (van Swaay and Warren 1999), it cannot be established whether any of these population changes reflect changes in habitat quality or composition. Given the tremendous importance of Italy for butterflies, there is an urgent need to assess how current and future habitat changes are likely to influence the Italian butterfly fauna. believe that The results could be considered as an initial assessment upon which to monitoring schemes should be implemented. These schemes should be concentrated in those areas judged as important from The results. Additionally the identification of management landscapes on the basis of species composition will hopefully serve as a basis of devise any regional conservation plans for the establishment of conservation priorities in these areas.

In conclusion have identified important areas for butterfly conservation by combining two approaches of species distribution modelling and Zonation. Italy is of unique importance for butterflies in Europe and an important area for several populations of threatened species. believe that information presented here can potentially allow efforts to be focused on areas of high biological value and those important for species of conservation concern.

5 A COMPARISON OF TREE-BASED METHODS FOR MODELLING SPECIES DISTRIBUTIONS

5.1 Introduction

Accurate knowledge about the distributional patterns of species is an essential prerequisite for biodiversity conservation and management. However, the limited amount of resources allocated for data collection, makes it very difficult to obtain complete information about the distribution of one or more species over large regions. As a consequence of this, both researchers and decision makers have to rely on predictive modelling to estimate patterns of species distribution (Guisan & Thuiller 2005). In niche or species distribution modelling species occurrence data are combined with environmental variables to infer the ecological requirements of a species. The geographic distribution of a species is then predicted by mapping the area where these environmental requirements are met (Guisan and Zimmermann 2000; Guisan and Thuiller 2005).

Classical methods used in species distribution modelling include statistical techniques like Generalized Linear Models (GLM) (McCullagh and Nelder 1989) or Generalized Additive models (GAM) (Hastie and Tibshirani 1990). While the benefits of using these techniques are numerous, including predicting changes in species' distribution from climate change (eg. Hilbert et al. 2004, Raimo et al. 2008) and identifying areas important for biodiversity conservation (eg. Milne et al. 2006, Lehmann et al. 2002), species distribution modelling is complicated by technical difficulties and by data limitations (Guisan & Thuiller 2005). Recent

advances in machine-learning techniques might be used to solve these problems, which generally derive from assumptions about the statistical distribution of data or restrictive assumptions of parametric modelling methods. Machine learning methods make fewer assumptions about the relationships between the variables.

Among the numerous machine learning methods available for species distribution modelling, classification trees represent an efficient tool, that has been applied in several species distribution modelling studies (e.g. Moisen et al. 2006, Edwards et al. 2006, Thuiller et al. 2003). Classification trees (Breiman et al. 1984), often known as CART, explain the variation in a single response variable with respect to one or more explanatory variables. They work by partitioning the data recursively into smaller and more homogenous groups with respect to the response variable. In addition the baseline methodology of classification trees includes several novel methods which have also been developed. These methods use iterative or bootstrapping procedures to combine several hundreds or thousands of trees together with the aim of improving model accuracy. Although these methods have been widely used in bioinformatics they have received relatively little exposure in ecology (Garzon et al. 2006, Leathwick 2006; Prasad et al. 2006). In this study I compare the predictive accuracy of four techniques based on classification trees, using large scale bird distribution data, with the aim of testing the performance of these four techniques to make suggestions for the optimal models which should be used by future researchers.

Although numerous studies have compared the predictive accuracy of different species distribution modelling techniques with respect to the type of

algorithm used (Elith et al. 2006, Prasad et. al. 2007, Garzon et al. 2006), fewer studies have investigated systematically how variation in species geographical and environmental distributions affect model' performance. (Segurado and Araujo 2004, Tsoar et al. 2007, Brotons et al. 2004, Luoto et al. 2007). Because this latter aspect is thought to particularly important as it could provide researchers the means to use theoretical or expert knowledge to predict which species are suitable for modelling, I decided to examine how species ecological characteristics could influence the predictive accuracy of the four techniques compared here. The specific aims of this study were: (i) to compare the predictive performance of four different modelling techniques based on decision trees (ii) to establish whether model performance is affected by the species' environmental and geographical distributions.

5.2 Methods

5.2.1 Species data

The data were derived from the results of the Italian Breeding Bird Monitoring Programme (M.I.T.O.). The M.I.T.O. started in 2000 and is administered by Italian Centre for Ornithology (CISO), the University of Milano Bicocca, the University of Calabria and the Association Faunaviva. Volunteer ornithologists carried out point counts during the breeding season; each observer recorded every species present in and out of a 100m radius circle. For a detailed description of the methodology see Fornasari and de Carli (2002). Two types of survey are carried out within the MITO monitoring programme (Fornasari and de Carli 2002): 1) a survey based on randomly selected point counts distributed

throughout the country and 2) a stratified survey of point counts in Special Protection Areas (SPAs) and Special Areas of Conservation (SACs). For the purpose of the present analysis I used data derived from the random surveys carried out between April and July 2000. These data comprise 6019 point counts inside 448 10x10km UTM squares. Although the original data were collected using a point count methodology, the data available to us consisted of presence/absence breeding records at the spatial resolution of 10x10km. For the analyses I used occurrence data for 104 species of birds. Species selection included a set of bird species with different extent in their distributions and ecological characteristics. A full list of the species is shown in APPENDIX 2.

5.2.2 Environmental variables

Thirteen environmental variables were used for the analyses (Table 5.1). Annual mean temperature and total precipitation data were obtained from the Agency for Environmental Protection and Technical Services (APAT, http://www.apat.it/). National climate maps were created using smoothing splines (Hutchinson 1991). One altitude and one slope variable were derived from a digital elevation model (DEM) (http://srtm.csi.cgiar.org/). Nine land cover types were derived from a digital CORINE data base (EEA 2000). The baseline resolution of all the environmental layers was 100 m (the lowest possible resolution for the CORINE Land Cover map, which was the layer with the coarser spatial resolution). The DEM, which had an original resolution of 90m, was resampled to obtain a pixel size of 100m. All the environmental layers were aggregated to match the resolution of the species data (10x10km). For each 10km

square I therefore calculated: the mean value of all the pixels for each of the climatic, slope and altitude variables, the minimum and maximum value of all the pixels for the altitude variable, and the area (ha) of each the nine land cover types.

Table 5.1: Predictor variables used for the modelling of bird species using the 4 predictive techniques

Variable description
Yearly total amount of precipitation (mm)
Mean slope of each square (º)
Mean annual temperature (ºC)
Mean altitude (m.a.s.l.)
Area of urban development (ha)
Area of arable land (ha)
Area of broad leaved forest (ha)
Area of coniferous forest (ha)
Area of sparse vegetation (ha)
Area of grassland and pastures (ha)
Area of moorland (ha)
Area of marshes and bogs (ha)
Agricultural areas with a significant portion of natural vegetation (ha)

5.2.3 Analyses

The modelling methodologies compared included the following: CART (Classification and Regression Tree), Bagging, Random Forests, and Boosted Regression Trees. The simplest method, CART, uses a single classification tree for prediction. Random Forests, Bagging and Boosting use a combination of trees

for prediction. This group of techniques, known as ensemble methods, differ from CART, as they do not seek the single most parsimonious model, but aim to fit large numbers of simple models which are then combined together to obtain accurate predictions (Araujo and New 2007).

CART

CART or classification tree and regression trees, were originally introduced by Breiman et al. (1984). In brief a classification tree divides the dimensional space defined by the predictors into groups that are as homogeneous as possible in terms of the response. The procedure begins with the entire data set, also called the root node, and formulates split defining conditions for each possible value of the explanatory variables to create candidate splits. Next, the algorithm selects the candidate split that minimises the misclassification rate and uses it to partition the data set into two subgroups. The algorithm continues recursively until all the data will be completely explained. Because of this model fitting procedure classification trees are usually overfitted. Tree pruning is thus required to reduce the model to an optimal size. In order to select the optimal size of each of the tree models I ran a series of 50 10-fold cross-validations and then selected the most frequently occurring tree size using the 1-SE rule (De'ath and Fabricius 2000). This procedure favours the largest tree for which the cross-validated error falls within 1 SE of the minimum relative error determined by cross-validation. Analyses for CART were carried out using the rpart package for R (Ihaka & Gentleman 1996).

Bagging

One major problem with classification trees is their high variance i.e. even a small change in the data can result in large changes in the resulting model. Bagging (Breiman, 1996) is an effective technique for reducing model variance. Instead of calculating the prediction for a single regression tree, bagging averages it over a collection of trees fitted on series bootstrap samples drawn from the original data set, improving thereby the prediction accuracy. When a bootstrap resample is drawn, about 37% of the data is excluded from the sample, but other data are replicated to bring the sample to full size. The portion of the data drawn into the sample in a replication is known as the ‘‘in-bag’’ data, whereas the portion not drawn is the ‘‘out-of-bag’’ data. In bagging all the classification trees are grown without pruning (Breiman 1996). The most important tuning parameter for bagging trees is the number of bootstrap replicates, hence the number of trees. Breiman (1996) suggested that a number of trees higher than 25 tend not to produce a significant test set error reduction. In the present case, when constructing bagging models I combined 50 trees. The analyses for Bagging were carried out using the ipred package (Peters et al. 2002) for R.

Random Forests

The Random Forests algorithm (Breiman 2001, Cutler et al. 2007) is a relatively novel machine learning technique and is designed to produce accurate predictions that do not overfit the data. Similarly to Bagging, Random Forests aims at reducing model variance. However, in Random Forests, two types of randomness are introduced: the first type is similar to Bagging (i.e. randomly sample with replacement n observations); The tree is then constructed using two

thirds of the data from the bootstrap replicate. The second source of randomness is introduced in the model by selecting randomly (without replacement) a number of predictors when constructing each tree. This approach yields a highly diverse ensemble of trees (i.e. the forest) that have both low bias and low variance. The output of random forests depends primarily on the number of predictors selected randomly for the construction of each tree. In order to establish the optimal number of predictors for each species I used a search grid containing values ranging from three to ten. I chose the models which had the lowest out bag error estimates. The analyses for random Forests were carried out using the Random Forests package in R (Liaw and Wiener 2002).

Boosted regression Trees

Boosted regression trees combine the algorithms of classification trees together with the boosting algorithm (Friedman 2001, 2002). The boosting algorithm is a very general method that attempts to "boost" the accuracy of any given learning algorithm by fitting a series of models each having a poor error rate and then combining them to give an ensemble that improves performance (Elith et al. 2008). In a boosted regression tree a series of very simple regression trees are fit by progressively adding trees in a forward stagewise fashion. At each stage of the fitting sequence, each case of the response variable is classified from the current sequence of trees. These classifications are used as weights (i.e. pseudo-residuals) for fitting the next sequence of trees. The fitting procedure is then continued until all the data have been explained (Friedman 2001). Although boosted regression trees can be a powerful tool to analyse complex data sets they

are also prone to over-fitting (i.e. trees can be added until eventually all the data will be explained). As a consequence of this, the performance of the final model will degrade when applied to new data. Optimising boosted regression trees for prediction involves the choice of the optimal number of trees, which maximise prediction accuracy. I used 10-fold cross validation procedure similar to the one adopted by Leathwick et al. (2006) to identify the optimal number of trees. I specified a tree size of 2 for individual trees, allowing for the inclusion of simple two terms interactions between variables. The analyses for Boosted Regression Trees were carried out using the gbm package (http://www.i-pensieri.com/gregr/gbm.shtml) for R.

5.2.4 Model evaluation

In order to evaluate the models, the original data set were randomly split into model training (70%) and model evaluation data sets (30%). The training dataset was used for the construction of the model whereas the evaluation data set was used to test the predictive abilities of the models. Model performance was tested using the area under the curve (AUC) of receiver-operating characteristic plot (ROC). ROC plot analysis measures the association between the presence and absence records by using and calculating the area under the curve (AUC) (Fielding & Bell, 1997). AUC relates relative proportions of correctly classified (true positive proportion) and incorrectly classified (false positive proportion) cells over a wide and continuous range of threshold levels, which makes it a threshold-independent measure (Fielding and Bell, 1997). The AUC values range from <0.5 for models with no discrimination ability to 1 for models with perfect

discrimination. AUC is not an absolute measure of model performance and it was recently criticised (Lobo et. al 2008). As long as a model can predict species absences quite well, it is easy to obtain high AUC scores if the evaluation data contain absence points selected from a very large geographical area (Lobo et al. 2008). Despite this major disadvantage, AUC can still provide a useful measure of relative model performance between models. As the choice of the evaluation metrics may influence the results, I also tested model predictions with a second measure, the maximised kappa. Because max-kappa provided similar results I present only the AUC results here.

5.2.5 Analysis of model performance

I tested for a difference in the performance of the four methods using a repeated measure anova. To assess whether some techniques were better than others, I used an approach similar to the one used by Guisan et al. (2007b), which consisted in finding the best performing technique for each species. For each technique I compared its performance across all the species to the vector of best performance using a Wilcoxon test and plotted the corresponding P values, which provide an index of the deviation form the best performance.

In order to examine how ecological characteristics could explain model performance I calculated three measures to describe the species' geographical and environmental distributions. The species geographical distribution was described using prevalence, that is the ratio of presence squares to the total sample. The environmental distributions were described with two measures: (1) marginality (or niche position), which reflects how far the species optimum is from the mean

environmental conditions in the study area; and (2) tolerance (or niche breath), which describes how variable the species association to environmental factors is with reference to the available range in the study area. 2001). Marginality and tolerance were calculated using E.N.F.A. (Ecological Niche Factor Analysis) (Hirzel et al. 2002).

I examined how model type and species ecological characteristics influence model performance using a linear mixed-effect model. Because of the high intercorrelation between prevalence, marginality and tolerance, I used Principal Component Analysis (PCA) to summarize the major patterns of variations of the four variables (Brotons et al. 2004). I obtained two independent components: 1) a marginality component positively associated to species marginality (r=0.97) and 2) a prevalence component negatively correlated with prevalence (r=0.80). I modelled the AUC scores as a function of the following fixed effects: prevalence marginality components, and their two-way interaction with model type. Species were included as a random effect.

5.3 Results

Ranking of overall model performance was: 1) Random Forests (RF) 2) Boosted Regression Trees (BRT), 3) Bagging (BAG), and CART (Fig 5.1 and Fig. 5.2).

Figure 5.1: Mean performance (AUC) for each of the four modelling methods (Abbreviations: BAG, Bagging predictors; BRT, Boosted regression trees; CART, Classification tree; RF, Random Forests). Error bars represent mean ± 1 standard error.

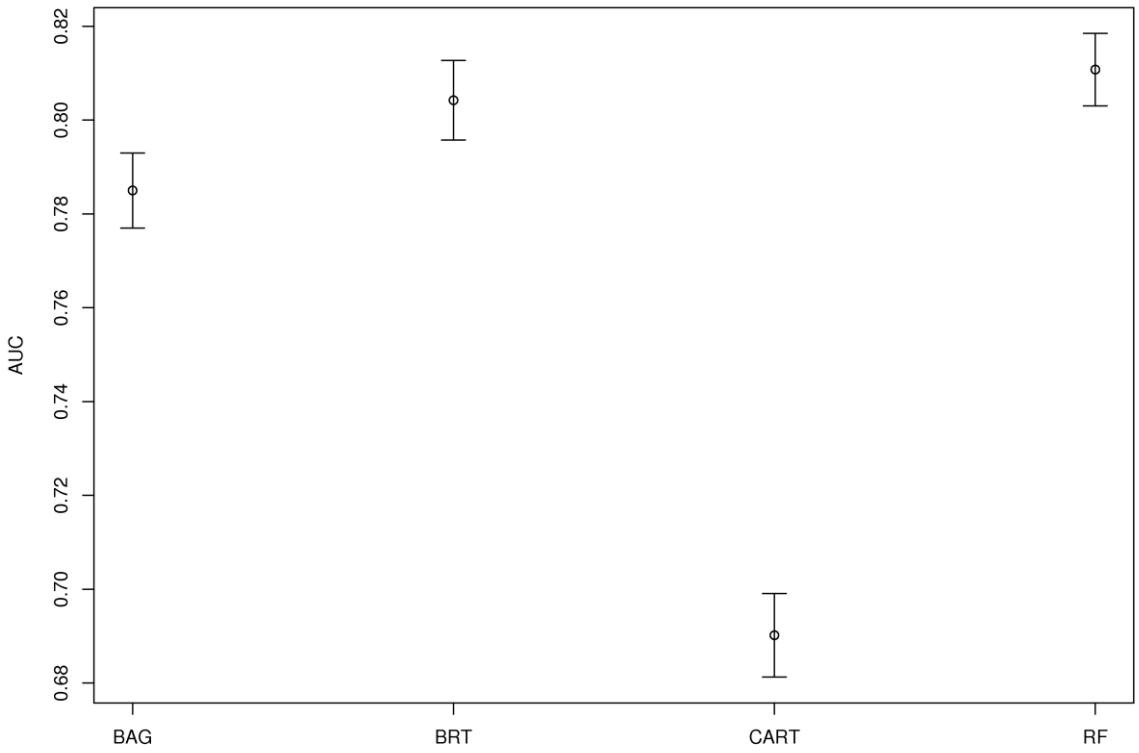
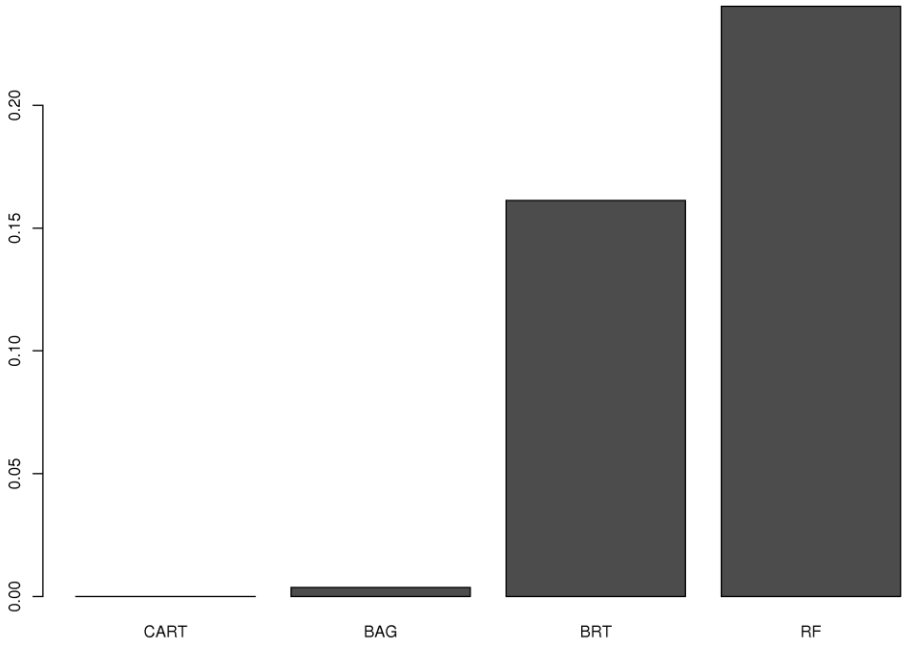


Figure 5.2: Histograms showing the P -value of difference (as obtained from Wilcoxon tests) from the best performing vector. Higher P-values indicate a better performance. Techniques are ranked in order of performance.



The repeated measure ANOVA indicated that differences among model type were highly significant (P <.0001). According to the classification by Swets (1988) no technique had a performance that was consistently above chance for all of the species i.e. not all the AUC values were all above 0.69 (Fig 5.1 and APPENDIX 2). In addition to overall differences in model performance I found that variance in the AUC values for each differed considerably across techniques. Ideally

algorithms should yield predictions with high AUC values and low variability across species. Random Forests was the technique which Random Forests which had the highest mean AUC value (0.81) and the lowest variance (0.006), followed by Boosted Regression Trees which performed was only slightly performed less well (0.80) and had a similar variance (0.007). Bagging was the third technique in order of mean AUC performance (0.78) and variance (0.006). Finally CART was the was which had the lowest mean AUC value (0.69) and the highest variance (0.008). The linear mixed effect model indicated a strong response of model performance to species marginality and prevalence (Table 2).

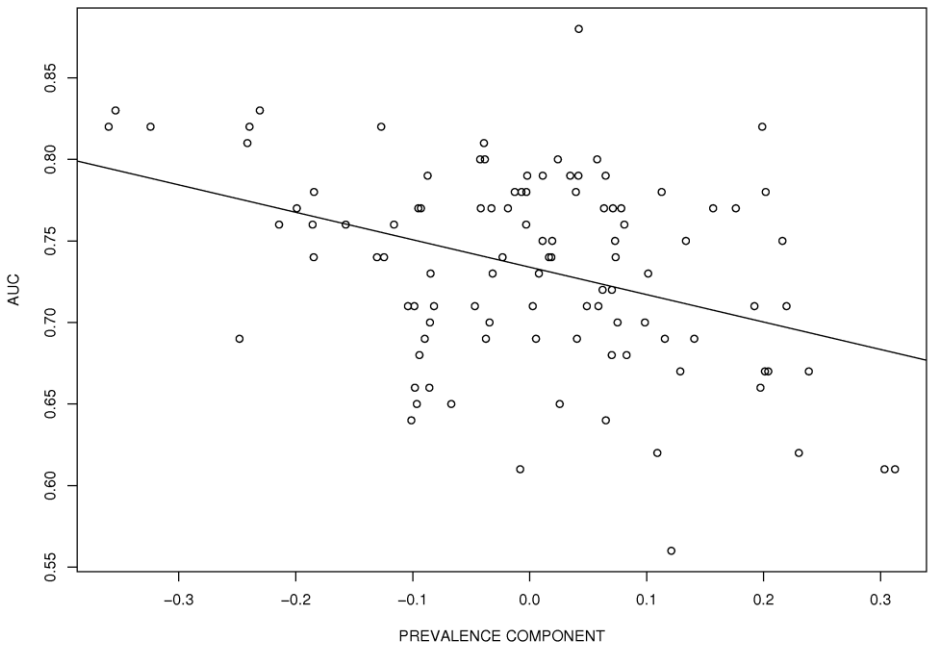
Table 5.2. Results of the linear mixed effect model investigating the determinants of area under the curve (AUC) scores. AUC score were modelled as a function of the main fixed effects and model and their two way interactions. Each unique species was treated as a random effect.

	Degrees of freedom	F-value	P-value
Intercept	1	20952.20	<.0001
Model	3	99.80	<.0001
Marginality Component	1	33.38	<.0001
Prevalence Component	1	16.13	0.0001
Model x Prevalence Component	3	0.39	0.7577
Model x Marginality Component	3	396.66	<.0001

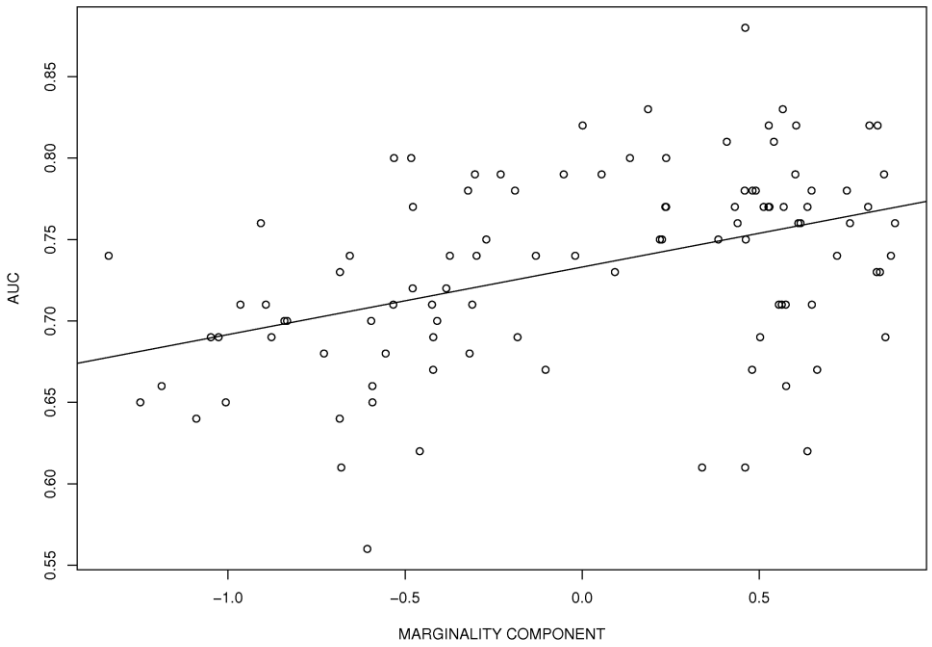
The two way interaction between model type and the marginality component was also significant. Graphical representation of the relationship between the species' marginality and prevalence component and AUC are presented in Fig 5.3. Species with higher prevalence values were generally more difficult to model (Fig. 5.3A), and generally had lower AUC values. On the contrary, as marginality increased model predictive accuracy decreased (Fig. 5.3B).

Figure 5.3: Mean AUC (over all methods) vs. prevalence (A) and marginality (B) components. The prevalence component describes a gradient from abundant to scarce species (i.e. from left to right). Marginality is an index which describes how far the species optimum is from the mean environmental profile in the study area. Higher values of the marginality component indicate more marginal species

A



B



5.4 Discussion

The results demonstrated that Random Forests and Boosted regression trees have a distinctive advantage over CART and Bagging when predicting species distributions. Specifically, Random Forests had slightly superior performance to Boosted regression trees. This latter technique has been the subject of comparison in a number of studies (Elith et al. 2006, Guisan et al. 2007a, Graham et al 2008, Wisz et al. 2008, Leathwick et al. 2006). These have all shown that Boosted Regression Trees are one of the best methods currently available for species distribution modelling. However, this is the first time that the predictive performance of Boosted Regression Trees was compared with Random Forests. The results indicate that Random Forests might be better suited than Boosted regression trees for predicting species distributions.

The third best performing technique was Bagging. Bagging is similar to Random Forests, as it uses bootstrapping to reduce the variance of the model. However, in Random Forests predictors are also selected randomly, which introduces another source of diversity into the ensemble of trees. To my knowledge there is only one other study where Bagging was used within an ecological context (Prasad et. al. 2006). The authors found that that the predictive accuracy of Bagging was slightly lower than Random Forests. As well as being concordant with Prasad's et. al (2006) findings, the results showed that Bagging had lower performance in comparison to Boosted Regression Trees. During the analyses I found that CART performed poorly in comparison to the other three methods. This confirms the lower predictive ability of CART, in comparison to ensemble methods (Prasad et al. 2006, Garzon et. al 2006). Although I employed cross-validation to find an optimal tree structure, this method does not seem to be practical for selecting the best model, which maximise predictive accuracy (Friedman 2001).

Similarly to previous studies I found that species' ecological characteristics affect model predictive accuracy (Poyry et al. 2008 Tsoar et al. 2007, McPherson et al. 2007, Segurado and Araujo 2004, Seoane et al. 2005, Hernandez et al. 2006). Species with higher prevalence values were generally more difficult model. Several studies have examined the effects of prevalence on model performance, and an important role of prevalence on the predictive accuracy of species distribution models (Manel et al. 2001, Luoto et al. 2005, Berg et al. 2004, Stockwell and Peterson 2002). The results support the hypothesis

that species widespread in geographic space are generally more difficult to model than species with a more restricted spatial distribution. On one hand, this finding coincides with the observations made by other studies (e.g. Luoto et al. 2005; Stockwell and Peterson 2002; Segurado and Araujo 2004), who reported that model performance degraded with species which have a wider distribution. On the other hand, it contrasts with two other studies, which showed that accuracy of species distribution models was better at an intermediate level of prevalence (McPherson et al. 2004) or was independent of prevalence (Manel et al. 2001). As stressed Brotons et al. (2004), an evaluation of the effects of prevalence on model performance is difficult, as prevalence is likely to vary both with species ecological characteristics and relative sampling effort. Species with specialized ecological requirements will tend to have a restricted range and therefore fewer records will be available for these than generalist species. Low sampling effort and bias in data collection could also influence species prevalence.

I found that marginality was also very important in determining model predictive accuracy. Specifically, species with a higher marginality were better modelled than those with a lower marginality. This result, suggests that species with a restricted niche are generally well predicted, whereas widespread species are modelled less accurately. Other studies have shown that species marginality can have a profound influence on model predictive accuracy (Tsoar et. al. 2007, Brotons et al. 2002, Stockwell and Peterson 2002). Two main explanations have been proposed to explain the effects of marginality on model predictive accuracy: Brotons et al. (2004) suggested that species inhabiting a wide range of habitats in

a certain area might not be limited by any of the environmental variables measured at the scale at which the models are fitted. However, it also possible that widespread species will show regional differences in their ecological niche, as a result of local adaptations. Modelling all of these populations together would overestimate the ecological breadth of species and thus decrease model accuracy (Stockwell and Peterson 2002).

Overall in spite of the similarity of the four techniques examined, I found a significant difference in their performance. On the basis of the results it is possible to point some of trends in the performance which should be verified or disproved by further research. Among these trends are the following: Random Forests and Boosted Regression Trees are the best methods, and using these two methods will allow one to reach either the best or close the best performance with relatively easy parameter tuning. On the contrary, Bagging and CART do not seem to be particularly competitive in their performance. However, it should be borne in mind that the choice of an appropriate modelling method will depend on the species being modelled and the goals of the modelling exercise (Segurado and Araujo 2004). Machine learning methods should be considered as the analysis tool when little prior knowledge exists of an ecological system or when accurate predictions are the desired product from an analysis. Both of these conditions are often met in reality, and I believe that and ecologists should make more use of these techniques. When used appropriately, machine learning methods are not data dredging and their strengths in exploratory data analysis make them a logical component, or even end product, of a thorough analysis of data.

Conclusion

This study has provide a first evaluation of the performance of four tree based classifiers within an ecological context. The results show that Random Forests and Boosted regression Trees, are the best tree based modelling methods for predicting species distributions. The potential applications of these techniques are many, ranging from the identification of important areas for species of conservation concern to the prediction of species distributions in relation to climate change. Further research should be aimed comparing the techniques applied in this study with a broader range of modelling techniques. I believe that a comparison with a broader range of modelling techniques would prove useful especially for ensemble methods which have been introduced very recently in ecology. As emphasised by the results, future comparisons should not only examine the predictive accuracy of a modelling techniques with the respect to algorithm used, but, should also take into account species ecological characteristics. Taking species ecological characteristics into account will allow researchers to predict what species are more suitable for species distribution modelling.

6 GENERAL DISCUSSION

The effectiveness of conservation actions is strongly dependent on the quality and the amount of ecological knowledge regarding the species or ecosystem under study. Therefore, it is of prime importance that ecologists, conservation biologists and landscape managers use appropriate methodological approaches to tackle specific conservation problems. This thesis has shown how tree-based modelling methods can be used for exploring and testing hypotheses about the factors that are important in determining species distribution and making predictions of species distribution for use in conservation contexts. I have employed four different modelling techniques to: understand how landscape structure and climate affect species distributions (Chapter 2 and 3), develop robust models capable of predicting species distributions under both present and future environmental conditions (Chapter 3), identify areas important for biodiversity conservation (Chapter 4), and to compare the relative performance of each modelling method in order to make recommendations to future researchers (Chapter 5).

In Chapter 1 Boosted Regression Trees were used to quantify the importance of wetland size and weather patterns for waterbirds wintering in Britain. As well as revealing a major role of weather patterns in determining waterbird occurrence, the models proved to be reasonably robust when validated. In chapter 3 this basic form of modelling was expanded, using a database containing amphibian occurrence records. Random Forests was used to quantify

species-climate relationship and to predict amphibian distribution in relation to current and future climate conditions. The results revealed how amphibian distribution is largely controlled by temperature-related variables and highlighted a negative response to future climate changes in most species. In Chapter 4 Bagging were used within an applied conservation context. Specifically, Bagging was used predict the distribution of 232 species of Butterflies in Italy. The predicted surfaces were then used in combination with a species multispecies prioritization tool in order to identify important areas for butterfly conservation. The results showed that the most areas important for butterfly are located within the Alps, the mountains of central Italy and the island of Sardinia. Finally, in Chapter 5, I compared the predictive accuracy of four modelling techniques based classification trees with the aim of making suggestions for the optimal models which should be used by future researchers. This was done using large scale bird distribution data from Italian Common Bird Census. The results showed that Random Forests and Boosted Regression Trees were the best performing techniques and that model performance was highly influenced by species ecological characteristics as well as by the modelling method.

Although the modelling techniques used herein have proved to be a flexible tool which can be used for a variety purposes, they are subject to a number of limitations which should be taken into consideration when interpreting the results. The first part of the discussion will therefore be aimed at discussing possible limitations of the modelling approach used . The second part of the

discussion will focus on the role of tree-based modelling methods in ecology and how researchers could benefit from their use.

6.1 Species distribution models-a challenge plagued by uncertainty

Several issues can contribute to uncertainty in model predictions and can hamper attempts to identify plausible relationships between species distributions and environmental variables. Not surprisingly these issues have received considerable attention in the ecological literature (e.g. Araujo and Guisan 2006, Heikkinen et al. 2006, Pearson et al. 2006) . Here I discuss two some issues which could have affected the models developed in this thesis.

Effects of species ecological characteristics

An important finding, that has emerged in all the chapters, is that a species' geographical distribution has important effect on model predictive accuracy. Models with high performance were those obtained for species with a narrower geographical distribution. Conversely, models with low predictive accuracy were obtained for species with a wider geographical distribution. This pattern was consistent across all the different species groups (British Waterbirds, Italian Amphibians, Butterflies and terrestrial birds). Several studies have stressed the importance of a species geographical distribution on the predictive accuracy of species distribution models (Poyry et al. 2008, McPherson et al. 2007, Segurado and Araujo 2004, Hernandez et al. 2006 , Luoto et al. 2005, Stockwell and Peterson 2002). The results of this thesis do indeed support the hypothesis that species geographical attributes widespread in geographic space are generally

more difficult to model than species with a more restricted spatial distribution. This is particularly evident in Chapter 5 where four different modelling techniques were applied to the same dataset. All the techniques showed a poor performance when applied to species with a wider geographical distribution. Stockwell and Peterson (2002) proposed as a biological explanation, suggesting that widespread species often show local or regional differences in ecological characteristics. Therefore, the more widespread a species is, the more likely it is use different habitats thus increasing the likelihood that more factors determine its distribution. However, it also possible the scale of the analysis could have affected the results. In general the degree to which environmental factors affect patterns of species distribution may be influenced by two aspects of scale: the spatial extent of the area considered and the spatial resolution at which patterns are examined, which is defined the smallest unit of area where any given attribute of interest can no longer be broken down into constituent parts. Species with a wider geographical distribution might not be limited by any environmental factors at the scale which models were fitted. In other words the scale of the analyses did not capture the full species niche, thus leading to a poor model performance. Unfortunately, due to sampling limitations one cannot have species occurrence data at scale as broad as full species niche and at a fine enough resolution that could be useful for applied conservation purposes. Consequently, some ecological processes are necessarily not captured or reflected in the results of the models because the independent variables were collected or represented at too coarse of a resolution to accurately depict all the complex relationships between species and

their environment. Species distribution patterns reflected in the results could therefore be misleading. A possible solution this problem would be the use of hierarchical modelling approach (Pearson et al. 2004). By taking advantage of the hierarchical nature of the drivers of species distributions, a hierarchical modelling approach would allow one to capture the full species niche by including those factors that are relevant at coarse scale and factors that are more relevant at local scale. Hierarchical models could be used to predict species distribution at a local scale while capturing the full species' niche. This approach could, however, be very time-consuming when carrying out comprehensive biodiversity assessments which include hundreds of species.

Effects of species data and model type and evaluation metrics

A possible source of uncertainty in model prediction might have been derived from the quality of the species data. In chapter 2 and 5 data from two monitoring schemes were used, whereas in chapters 3 and 4 presence data only were used. Since the modelling methods applied here require both presence and absence data pseudo absences I had to assume pseudo-absences for the Butterfly and Amphibian datasets in chapter 3 and 4. Intuitively, one would expect the use of pseudo absence bias would lead to inaccurate predictions, as several authors have cautiously said (). Yet, several previous studies have shown that species distribution models may still perform reasonably well, even if pseudo absences were chosen. However in both chapter 3 and 4 contained the validation data contained pseudo absence data, so it is unknown whether the discriminatory power of models as tested using ROC plots.

As demonstrated in Chapter 5 modelling technique can have a considerable impact on the predictive accuracy of the model. On the basis of the results it is possible to point some of trends in the performance which should be verified or disproved by further research. Among these trends are the following. Random Forests and Boosted Regression Trees are better suited for predicting species distributions than other methods. These methods will allow one to reach either the best or close the best performance with relatively easy parameter tuning. However, the choice of the method will be dictated have to be based on the goals of the study and the computational resources available. For example, Random Forest requires less tuning, than in Boosted Regression Trees which require at least the identification of the optimal number of trees used for prediction. On the other hand with on large data sets, Random Forests (i.e. more than 3000 samples) training requires very significant computing resources and in some cases the use of this technique may be prohibitively time consuming. To this end Boosting can provide a valuable alternative, especially when the size of each tree be set to a small value.

Finally another factor which could affect the interpretation of model results is the method used for evaluating model predictive accuracy (i.e. AUC). Although the AUC is regarded as best currently available method for evaluating the performance of species distribution models it was recently criticised (Lobo et al 2008) . According to Lobo et. al (2008) the AUC could be a misleading measure of model performance as the total extent to which models are carried out highly influences the rate of well predicted absences an the AUC scores. This

means that as long as a model can predict species absences quite well, it is easy to obtain high AUC scores if the evaluation data contain absence points selected from a very large geographical area (Lobo et al. 2008). This is a serious drawback of this evaluation metric and there does not seem to be a valid alternative to AUC yet (Lobo et al. 2008).

6.2 Putting tree based modelling methods within a wider context

Analyzing large amounts species distribution data have become an issue of keen interest and elucidating species distributions patterns has become a vital important in many conservation programmes. The results presented this thesis show that tree-based methods can be useful tool in both theoretical and applied research. Whilst the focus of thesis was on modelling species occurrence, the modelling techniques applied herein can be used with types of ecological data such as abundance or density data. Tree-based methods are flexible and a useful way to visualize and understand relationships between environmental parameters and species distributions and future studies could expand upon the analyses carried out in this thesis by using these methods to test more specific ecological and biogeographical hypotheses. Most importantly, however, tree-based methods can be used to predict species distributions with high accuracy. This thesis did not compare their predictive performance with traditional statistical modelling techniques, though a number of studies have shown how some of the ensemble methods based on classification trees (Elith et al. 2006, Wisz et al. 2008, Graham et al. 2008) consistently outperform regression-based techniques like GLM and

GAM. Although statistical modelling techniques can be very useful in many of situations, tree-based modelling methods can efficiently discover the most important information from the complex and noisy ecological data. Hopefully the results of this thesis will also serve a source of inspiration for ecologists willing to move away from the p-value dogma (Fielding 1999) and will allow them to concentrate on understanding the data and using these techniques to predict species distributions with a higher accuracy. Tree-based methods will never be a panacea for all the problems with ecological data. However, they are another set of tools that ecologists should be aware of (Fielding 1999). Whilst it is true that some statistical modelling techniques can be used instead of tree-based methods others expand the analytical opportunities by enabling analyses that are impossible or very difficult with statistical methods. This set of methods constitute an important and flexible tool that should be added to the ecologist's toolbox. It should, however, be borne in mind that the choice of the appropriate modelling method will primarily depend on the goals of the study, and every ecologist should be fully aware of the limitations of the techniques being used. Species distribution models can only depict a picture of this should be understood by the modeller and the end user. The goal of modelling is not to reflect the full reality but to construct models which make biological sense, approximate this reality and constitute useful tools (Burnham and Anderson 1998). It is realistic to believe that a 'true' model will perfectly explain the biological data we observed (Hastie et al. 2001). Ecological systems are the results of many small effects, individual heterogeneity or interactions at multiple spatial and temporal scales and

models are simply low dimensional abstractions of infinite-dimensional forces acting on individuals. The way this abstraction is achieved mainly determines the usefulness of the models. As Box (1976) stated ‘all models are wrong, but some are useful’ and Burnham and Anderson (1998) stated that ‘increased sample size allows to chase full reality, but never to catch it’.

APPENDIX 1

List of butterfly species studied in chapter 4, model performance and threat status (LR(nt)=Lower Risk nearly threatened, VU=Vulnerable,EN=Endangered, CR=Critically Endangered)

Species	AUC	Threat status	Species	AUC	Threat status	Species	AUC	Threat status
Aglais urticae	0.70		*Brintesia circe*	0.83		*Erebia aethiopella*	0.89	
Anthocharis cardamines	0.67		*Cacyreus marshalli*	0.62		*Erebia aethiops*	0.86	LR(nt)
Anthocharis damone	0.86	VU	*Callophrys rubi*	0.72		*Erebia alberganus*	0.83	
Anthocharis euphenoides	0.83		*Carcharodus alceae*	0.75		*Erebia cassioides*	0.93	
Apatura ilia	0.85		*Carcharodus floccifera*	0.58		*Erebia epiphron*	0.91	
Apatura iris	0.80		*Carcharodus lavatherae*	0.57		*Erebia eriphyle*	0.78	
Aphantopus hyperantus	0.90		*Carterocephalus palaemon*	0.89		*Erebia euryale*	0.93	
Aporia crataegi	0.57		*Celastrina argiolus*	0.75		*Erebia gorge*	0.92	
Arethusana arethusa	0.81		*Charaxes jasius*	0.85		*Erebia ligea*	0.86	
Argynnis adippe	0.73		*Chazara briseis*	0.78		*Erebia manto*	0.86	
Argynnis aglaja	0.78		*Coenonympha arcania*	0.74		*Erebia medusa*	0.91	VU
Argynnis elisa	1.00		*Coenonympha corinna*	0.99		*Erebia melampus*	0.90	
Argynnis niobe	0.81		*Coenonympha darwiniana*	0.90		*Erebia meolans*	0.91	
Argynnis pandora	0.68		*Coenonympha dorus*	0.59		*Erebia mnestra*	0.96	
Argynnis paphia	0.72		*Coenonympha elbana*	0.96		*Erebia montana*	0.87	
Aricia agestis	0.77		*Coenonympha gardetta*	0.97		*Erebia neoridas*	0.78	
Aricia artaxerxes	0.91		*Coenonympha glycerion*	0.99		*Erebia oeme*	1.00	
Aricia cramera	0.99		*Coenonympha oedippus*	0.82	CR	*Erebia pandrose*	0.91	
Aricia eumedon	0.83		*Coenonympha pamphilus*	0.70		Erebia pharte	0.94	

Aricia nicias	0.98		*Coenonympha rhodopensis*	0.87		Erebia pluto	0.92	
Boloria dia	0.75		*Colias alfacariensis*	0.7		Erebia pronoe	0.97	
Boloria euphrosyne	0.66		*Colias croceus*	0.75		Erebia stirius	0.96	
Boloria napaea	0.91		*Colias hyale*	0.78		Erebia styx	0.97	
Boloria pales	0.92		*Colias palaeno*	0.89	LR(nt)	Erebia triaria	0.88	
Boloria selene	0.88		*Colias phicomone*	0.95		Erebia tyndarus	0.84	
Boloria thore	0.81	VU	*Cupido alcetas*	0.86		Euchloe ausonia	0.78	
Boloria titania	0.92	VU	*Cupido argiades*	0.79		Euchloe crameri	0.89	
Brenthis daphne	0.64		*Cupido minimus*	0.68		Euchloe insularis	0.97	
Brenthis hecate	0.66		*Cupido osiris*	0.68		Euchloe simplonia	0.98	EN
Brenthis ino	0.94		*Danaus chrysippus*	0.88		Euphydryas aurinia	0.68	VU
Euphydryas cynthia	0.93		*Libythea celtis*	0.66		Melitaea trivia	0.74	
Euphydryas intermedia	0.97	EN	*Limenitis camilla*	0.85		Melitaea varia	0.94	
Gegenes nostrodamus	0.65		*Limenitis populi*	0.83		Minois dryas	0.89	
Gegenes pumilio	0.83	LR(nt)	*Limenitis reducta*	0.78		Muschampia proto	0.7	
Glaucopsyche alexis	0.75	VU	*Lycaena alciphron*	0.73		Neozephyrus quercus	0.65	
Glaucopsyche melanops	0.73		*Lycaena dispar*	0.91		Neptis rivularis	0.91	
Gonepteryx cleopatra	0.82		*Lycaena hippothoe*	0.86	LR(nt)	Neptis sappho	1.00	LR(nt)
Gonepteryx rhamni	0.59		*Lycaena phlaeas*	0.74		Nymphalis antiopa	0.74	
Hamearis lucina	0.7	LR(nt)	*Lycaena thersamon*	0.82		Nymphalis polychloros	0.64	
Hesperia comma	0.69		*Lycaena tityrus*	0.83		Ochlodes venata	0.74	
Heteropterus morpheus	0.88		*Lycaena virgaureae*	0.87	LR(nt)	Oeneis glacialis	0.92	
Hipparchia aristaeus	0.98		*Maculinea alcon*	0.75	VU	Papilio alexanor	0.68	
Hipparchia fagi	0.7		*Maculinea arion*	0.72	EN	Papilio hospiton	0.98	
Hipparchia fidia	0.61		*Maculinea rebeli*	0.72		Papilio machaon	0.79	
Hipparchia neomiris	0.99		*Maculinea teleius*	0.99	VU	Pararge aegeria	0.75	
Hipparchia semele	0.67		*Maniola jurtina*	0.74		Parnassius apollo	0.95	VU
Hipparchia statilinus	0.71		*Maniola nurag*	0.77		Parnassius mnemosyne	0.83	
Hyponephele lupinus	0.78		*Melanargia arge*	0.91		Parnassius phoebus	0.97	VU
Hyponephele lycaon	0.86		*Melanargia galathea*	0.68		Pieris brassicae	0.65	
Inachis io	0.66		*Melanargia occitanica*	0.77		*Pieris bryoniae*	0.90	

Iolana iolas	0.87		*Melanargia russiae*	0.91		*Pieris ergane*	0.84	
Iphiclides podalirius	0.79		*Melitaea athalia*	0.64		*Pieris mannii*	0.69	
Issoria lathonia	0.56		*Melitaea aurelia*	0.79	VU	*Pieris napi*	0.65	
Lampides boeticus	0.72		*Melitaea britomartis*	0.62		*Pieris rapae*	0.70	
Lasiommata maera	0.71		*Melitaea cinxia*	0.70		*Plebeius argus*	0.79	
Lasiommata megera	0.78		*Melitaea deione*	0.75		*Plebeius argyrognomon*	0.72	
Lasiommata paramegera	0.92		*Melitaea diamina*	0.87		*Plebeius glandon*	0.92	
Lasiommata petropolitana	0.90		*Melitaea didyma*	0.73		*Plebeius idas*	0.85	
Leptidea sinapis	0.63		*Melitaea parthenoides*	0.68		*Plebeius optilete*	0.95	
Leptotes pirithous	0.85		*Melitaea phoebe*	0.60		*Plebeius orbitulus*	0.97	
Plebeius trappi	0.65		*Pontia daplidice*	0.74		*Satyrium ilicis*	0.82	
Polygonia c-album	0.75		*Pseudophilotes baton*	0.69		*Satyrium pruni*	0.84	
Polygonia egea	0.76		*Pseudophilotes vicrama*	0.79		*Satyrium spini*	0.74	
Polyommatus amandus	0.75		*Pyrgus alveus*	0.9		*Satyrium w-album*	0.55	
Polyommatus bellargus	0.74		*Pyrgus andromedae*	0.94		*Satyrus actaea*	0.53	
Polyommatus coridon	0.79		*Pyrgus armoricanus*	0.68		*Satyrus ferula*	0.79	
Polyommatus damon	0.89	LR(nt)	*Pyrgus bellieri*	0.66		*Scolitantides orion*	0.86	VU
Polyommatus daphnis	0.73		*Pyrgus cacaliae*	0.94		*Spialia orbifer*	0.6	
Polyommatus dolus	0.73		*Pyrgus carlinae*	0.92		*Spialia sertorius*	0.67	
Polyommatus dorylas	0.74		*Pyrgus carthami*	0.81		*Thecla betulae*	0.57	
Polyommatus eros	0.91	LR(nt)	*Pyrgus malvae*	0.54		*Thymelicus acteon*	0.77	VU
Polyommatus escheri	0.72		*Pyrgus onopordi*	0.58		*Thymelicus lineola*	0.67	
Polyommatus hispana	0.8		*Pyrgus serratulae*	0.85		*Thymelicus sylvestris*	0.63	
Polyommatus icarus	0.65		*Pyrgus sidae*	0.88		*Vanessa atalanta*	0.62	
Polyommatus ripartii	0.5		*Pyronia cecilia*	0.86		*Vanessa cardui*	0.67	
Polyommatus semiargus	0.78		*Pyronia tithonus*	0.76		*Zerynthia polyxena*	0.74	
Polyommatus thersites	0.78		*Satyrium acaciae*	0.73				
Pontia callidice	0.91		*Satyrium esculi*	0.93				

APPENDIX 2

List of bird species studied in chapter 5 and model performance, as measured by the AUC.

Species	BAG	BRT	CART	RF	Species	BAG	BRT	CART	RF
Acrocephalus arundinaceus	0.85	0.88	0.71	0.87	*Emberiza cia*	0.79	0.81	0.60	0.82
Acrocephalus palustris	0.91	0.88	0.54	0.92	*Emberiza cirlus*	0.83	0.86	0.74	0.86
Acrocephalus scirpaceus	0.84	0.85	0.71	0.84	*Emberiza citrinella*	0.79	0.80	0.62	0.82
Aegithalos caudatus	0.65	0.64	0.64	0.64	*Erithacus rubecula*	0.84	0.86	0.82	0.86
Alauda arvensis	0.69	0.68	0.61	0.69	*Falco tinnunculus*	0.63	0.72	0.62	0.69
Alcedo atthis	0.73	0.78	0.60	0.77	*Fringilla coelebs*	0.80	0.82	0.77	0.82
Anas platyrhynchos	0.81	0.83	0.73	0.84	*Fulica atra*	0.79	0.87	0.54	0.84
Anthus trivialis	0.89	0.92	0.82	0.92	*Galerida cristata*	0.85	0.85	0.78	0.87
Apus apus	0.68	0.67	0.62	0.68	*Gallinula chloropus*	0.83	0.85	0.70	0.84
Ardea cinerea	0.84	0.84	0.76	0.86	*Garrulus glandarius*	0.67	0.69	0.61	0.69
Ardea purpurea	0.88	0.91	0.71	0.91	*Himantopus himantopus*	0.79	0.83	0.74	0.84
Buteo buteo	0.73	0.76	0.73	0.76	*Hippolais polyglotta*	0.83	0.78	0.64	0.81
Calandrella brachydactyla	0.87	0.89	0.81	0.88	*Hirundo rustica*	0.72	0.73	0.66	0.73
Carduelis carduelis	0.74	0.75	0.61	0.75	*Ixobrychus minutus*	0.82	0.88	0.71	0.84
Carduelis chloris	0.71	0.70	0.63	0.74	*Jynx torquilla*	0.59	0.56	0.50	0.63
Certhia brachydactyla	0.75	0.77	0.78	0.77	*Lanius collurio*	0.65	0.62	0.59	0.66
Certhia familiaris	0.79	0.89	0.85	0.89	*Larus cachinnans*	0.82	0.81	0.74	0.82
Cettia cetti	0.78	0.80	0.71	0.81	*Larus ridibundus*	0.78	0.82	0.70	0.83
Charadrius dubius	0.69	0.78	0.63	0.77	*Loxia curvirostra*	0.91	0.94	0.60	0.95
Circus aeruginosus	0.82	0.87	0.66	0.84	*Luscinia megarhynchos*	0.81	0.83	0.79	0.82
Cisticola juncidis	0.84	0.86	0.73	0.85	*Melanocorypha calandra*	0.65	0.89	0.89	0.83
Columba palumbus	0.76	0.76	0.62	0.80	*Merops apiaster*	0.76	0.81	0.73	0.81
Corvus corax	0.86	0.89	0.75	0.87	*Miliaria calandra*	0.79	0.78	0.74	0.80

Corvus corone cornix	0.73	0.77	0.67	0.74	*Milvus migrans*	0.60	0.58	0.52	0.67
Corvus corone corone	0.74	0.81	0.72	0.79	*Motacilla alba*	0.65	0.66	0.61	0.68
Corvus monedula	0.65	0.65	0.58	0.66	*Motacilla cinerea*	0.81	0.82	0.61	0.81
Coturnix coturnix	0.68	0.69	0.64	0.74	*Motacilla flava*	0.85	0.86	0.78	0.87
Cuculus canorus	0.75	0.76	0.68	0.77	*Muscicapa striata*	0.62	0.62	0.56	0.64
Delichon urbica	0.72	0.67	0.57	0.71	*Nycticorax nycticorax*	0.86	0.89	0.80	0.93
Dryocopus martius	0.91	0.95	0.79	0.92	*Oenanthe oenanthe*	0.76	0.77	0.62	0.79
Egretta garzetta	0.88	0.91	0.82	0.90	*Oriolus oriolus*	0.75	0.76	0.70	0.76
Parus ater	0.89	0.91	0.83	0.90	*Pyrrhula pyrrhula*	0.88	0.86	0.73	0.90
Parus caeruleus	0.74	0.76	0.74	0.76	*Regulus ignicapillus*	0.77	0.81	0.68	0.80
Parus cristatus	0.87	0.89	0.76	0.91	*Regulus regulus*	0.90	0.92	0.76	0.94
Parus major	0.61	0.66	0.52	0.65	*Remiz pendulinus*	0.85	0.84	0.69	0.86
Parus montanus	0.96	0.97	0.93	0.97	*Saxicola rubetra*	0.86	0.79	0.57	0.83
Parus palustris	0.81	0.83	0.78	0.82	*Serinus serinus*	0.75	0.77	0.69	0.77
Passer italiae	0.76	0.74	0.68	0.76	*Sitta europaea*	0.74	0.77	0.68	0.77
Passer montanus	0.75	0.74	0.70	0.75	*Streptopelia decaocto*	0.75	0.76	0.69	0.77
Pernis apivorus	0.56	0.64	0.55	0.66	*Streptopelia turtur*	0.83	0.83	0.75	0.85
Phalacrocorax carbo	0.80	0.85	0.57	0.86	*Sturnus vulgaris*	0.79	0.81	0.72	0.80
Phasianus colchicus	0.84	0.84	0.70	0.85	*Sylvia atricapilla*	0.81	0.86	0.68	0.88
Phoenicurus ochrurus	0.84	0.85	0.75	0.84	*Sylvia cantillans*	0.87	0.89	0.61	0.92
Phoenicurus phoenicurus	0.77	0.79	0.76	0.80	*Sylvia communis*	0.84	0.84	0.83	0.85
Phylloscopus bonelli	0.84	0.82	0.75	0.85	*Sylvia melanocephala*	0.89	0.93	0.90	0.92
Phylloscopus collybita	0.87	0.87	0.77	0.88	*Tachybaptus ruficollis*	0.90	0.91	0.78	0.90
Pica pica	0.73	0.73	0.70	0.75	*Troglodytes troglodytes*	0.84	0.90	0.64	0.85
Picoides major	0.76	0.74	0.65	0.76	*Turdus merula*	0.79	0.78	0.75	0.82
Picus viridis	0.80	0.78	0.72	0.82	*Turdus philomelos*	0.77	0.77	0.65	0.77
Podiceps cristatus	0.78	0.81	0.52	0.83	*Turdus pilaris*	0.85	0.85	0.79	0.86
Ptyonoprogne rupestris	0.81	0.83	0.59	0.82	*Turdus viscivorus*	0.87	0.97	0.57	0.96
Upupa epops	0.80	0.86	0.61	0.82	*Vanellus vanellus*	0.69	0.69	0.58	0.71

REFERENCES

Abellan, P., Sanchez-Fernandez, D., Velasco, J., Millan, A. (2005). Conservation of freshwater biodiversity: a comparison of different area selection methods. *Biodiversity and Conservation* **14**: 3457-3474.

Alford R. A. & Richards S. J. (1999). Global amphibian declines: A problem in applied ecology. *Annual Review of Ecology and Systematics* **30**: 133-165.

Araujo M. B. & Luoto M. (2007). The importance of biotic interactions for modelling species distributions under climate change. *Global Ecology and Biogeography* **16**: 743-753.

Araujo M. B., Pearson R. G., Thuiller W. & Erhard M. (2005). Validation of species-climate impact models under climate change. *Global Change Biology* **11**: 1504-1513.

Araujo M. B., Thuiller W. & Pearson R. G. (2006). Climate warming and the decline of amphibians and reptiles in Europe. *Journal of Biogeography* **33**: 1712-1728.

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

Araujo, M.B., New, M. (2007). Ensemble forecasting of species distributions. *Trends in Ecology and Evolution* **22**: 42-47.

Arponen, A., Heikkinen, R.K., Thomas, C.D., Moilanen, A. (2005). The value of biodiversity in reserve selection: Representation, species weighting, and benefit functions. *Conservation Biology* **19**: 2009-2014.

Atkinson, P.W., Austin, G.E., Rehfisch, M.M., Baker, H., Cranswick, P., Kershaw, M., Robinson, J., Langston, R.H.W., Stroud, D.A., Van Turnhout, C. & Maclean, I.M.D. (2006). Identifying declines in waterbirds: The effects of missing data, population variability and count period on the interpretation of long-term survey data. *Biological Conservation*, **130**: 549-59.

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

Aviron, S., Jeanneret, P., Schupbach, B., Herzog, F. (2007). Effects of agri-environmental measures, site and landscape conditions on butterfly diversity of Swiss grassland. *Agriculture Ecosystems and Environment* 122: 295-304.

Baillie, J. E. M., Hilton-Taylor, C. and Stuart, S.N. (2004). IUCN Red List of Threatened Species. A Global Species Assessment. , IUCN, Gland, Switzerland and Cambridge,UK. xxiv+ 191 pp.

Baker, C., Lawrence, R., Montague, C. & Patten, D. (2006) Mapping wetlands and riparian areas using Landsat ETM+ imagery and decision-tree-based models. *Wetlands*, **26**: 465-474.

Balletto L., Bonelli S., Cassulo L. (2005a). Mapping the Italian butterfly diversity for conservation. In: Studies on the Ecology and Conservation of butterflies in Europe 71-76. Kühn, R. Feldmann, J. Thomas, J. Settele (Eds). Sofia & Moscow: Pensoft.

Balletto, E. (1992). Butterflies in Italy: status, problems and prospect. In: Future of Butterflies in Europe: strategies for survival. T. Pavlicek-van Beek, A.H. Ovaa and J.G. van der Made (Eds) Wageningen: Proceedings of the International Congress.

Balletto, E., Kudrna, O. (1985). Some aspects of the conservation of butterflies in Italy, with recommendations for a future strategy. *Bolletino della Societa entomologica italiana* **117**: 39–59.

Balletto, E., Bonelli, S., Cassulo, L. (2005b). Checklist e distribuzione della fauna italiana. 10.000 specie terrestri e delle acque interne: Insecta Lepidoptera Papilionoidea (Rhopalocera). Memorie del Museo civico di Storia Naturale di Verona, 16, 259–263.

Berg, A., Gardenfors, U. and von Proschwitz, T. (2004). Logistic regression models for predicting occurrence of terrestrial molluscs in southern Sweden - importance of environmental data quality and model complexity. *Ecography*, **27**:83-93.

Blaustein A. R., Belden L. K., Olson D. H., Green D. M., Root T. L. & Kiesecker J. M. (2001). Amphibian breeding and climate change. *Conservation Biology* **15**: 1804-1809.

Boone M. D., Semlitsch R. D., Little E. E. & Doyle M. C. (2007). Multiple stressors in amphibian communities: Effects of chemical contamination, bullfrogs, and fish. *Ecological Applications* **17**: 291-301.

Box, G.E.P. (1976). Science and statistics. *Journal of the American Statistical Association* **71**: 791-799.

Breiman L, Freidman J, Olshen R, Stone C. (1984). Classification and regression trees. Belmont (CA): Wadsworth, 358 p.

Breiman, L. (1996). Bagging predictors. *Machine Learning* **24**:123-140.

Breiman, L. (2001). Random forests. *Machine Learning* **45**:5-32.
Bricklemyer, R. S., Lawrence, R. L., Miller, P. R. & Battogtokh, N. (2007). Monitoring and verifying agricultural practices related to soil carbon sequestration with satellite imagery. *Agriculture Ecosystems & Environment* **118**: 201-210.

Brooker R. W., Travis J. M. J., Clark E. J. & Dytham C. (2007). Modelling species' range shifts in a changing climate: The impacts of biotic interactions, dispersal distance and the rate of climate change. *Journal of Theoretical Biology* **245**: 59-65.

Brotons, L., Thuiller, W., Araujo, M.B. and Hirzel, A.H. (2004). Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography* **27**:437-448.

Brown, J.H. & Lomolino, M.V. (1998). Biogeography, 2nd edn. Sinauer Associates, Sunderland, Mass.

Buermann, W., Saatchi, S., Smith, T. B., Zutta, B. R., Chaves, J. A., Mila, B. & Graham, C. H. (2008). Predicting species distributions across the Amazonian and Andean regions using remote sensing data. *Journal of Biogeography* **35**: 1160-1176.

Burnham, K.P. and Anderson, D.R. (1998). Model Selection and Multi-model inference: A Practical Information-Theoretic Approach. Springer-Verlag, New York.

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*. CSIRO, Australia.

Cabeza, M., Araujo, M.B., Wilson, R.J., Thomas, C.D., Cowley, M.J.R., Moilanen, A. (2004). Combining probabilities of occurrence with spatial reserve design. *Journal of Applied Ecology* **41**: 252-262.

Cappo, M., De'ath, G., Boyle, S., Aumend, J., Olbrich, R., Hoedt, F., Perna, C. & Brunskill, G. (2005). Development of a robust classifier of freshwater residence in barramundi (*Lates calcarifer*) life histories using elemental ratios in scales and boosted regression trees. *Marine and Freshwater Research* **56**: 713-723.

Carey C. & Alexander M. A. (2003). Climate change and amphibian declines: is there a link? *Diversity and Distributions* **9**: 111-121.

Carey C., Heyer W. R., Wilkinson J., Alford R. A., Arntzen J. W., Halliday T., Hungerford L., Lips K. R., Middleton E. M., Orchard S. A. & Rand A. S. (2001). Amphibian declines and environmental change: Use of remote-sensing data to identify environmental correlates. *Conservation Biology* **15**: 903-913.

Carpenter, G., Gillison, A. N. & Winter, J. (1993). Domain - a Flexible modeling procedure for mapping potential distributions of plants and animals. *Biodiversity and Conservation* **2**: 667-680.

Carreiras, J.M.B., Pereira, J.M.C., Campagnolo, M.L., Shimabukuro, Y.E. (2006). Assessing the extent of agriculture/pasture and secondary succession forest in the Brazilian Legal Amazon using SPOT VEGETATION data. *Remote Sensing of Environment* **101**: 283-298.

Chase J.M., Leibold M.A. (2003). Ecological niches: Linking classical and contemporary approaches. University of Chicago Press. 212p.

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

Coops, N.C. & Catling, P.C. (2002). Prediction of the spatial distribution and relative abundance of ground-dwelling mammals using remote sensing imagery and simulation models. *Landscape ecology*, **17**: 173-188.

Cranswick, P.A., Kirby, J.S., Salmon, D.G., Atkinson-Willes, G.L., Pollitt, M.S., Owen, M. (1997). A history of wildfowl counts by the Wildfowl and Wetlands Trust. *Wildfowl* **47**: 217–230.

Crozier G.E. & Niemi G.J. (2003). Using patch and landscape variables to model bird abundance in a naturally heterogeneous landscape. *Canadian Journal of Zoology* **81**: 441-452.

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A. and Hess, K.T. (2007). Random forests for classification in ecology. *Ecology*, **88**:2783-2792.

Dapporto L., Dennis R.L.H. (2007). Species richness, rarity and endemicity on Italian offshore islands: complementary signals from island-focused and species-focused analyses. *Journal of Biogeography* Published article online: 25-Oct-2007 doi: 10.1111/j.1365-2699.2007.01812.x

Davis A. J., Jenkinson L. S., Lawton J. H., Shorrocks B. & Wood S. (1998). Making mistakes when predicting shifts in species range in response to global warming. *Nature* **391**: 783-786.

Dayton, G.H., Fitzgerald, L.A. (2006). Habitat suitability models for desert amphibians. *Biological Conservation* **132**: 40-49.

De'ath G. (2007). Boosted trees for ecological modeling and prediction. *Ecology* **88** : 243-251.
De'ath, G. and Fabricius, K.E. (2000). Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology* **81**:3178-3192.

Donnelly M. A. (1998). Potential effects of climate change on two neotropical amphibian assemblages. *Climatic Change* **39**: 541-561.

Duncan, P., Hewison, A.J.M., Houte, S., Rosoux, R., Tournebize, T., Dubs, F., Burel, F. & Bretagnolle, V. (1999). Long-term changes in agricultural practices and wildfowling in an internationally important wetland, and their effects on the guild of wintering ducks. *Journal of Applied Ecology* **36**: 11-23.

Early, R., Thomas, C.D. (2007). Multispecies conservation planning: identifying landscapes for the conservation of viable populations using local and continental species priorities. *Journal of Applied Ecology* **44**: 253-262.

Edwards, T.C., Cutler D.R., Zimmermann N.E., Geiser L., and Moisen G.G. (2006). Effects of sample survey design on the accuracy of classification tree models in species distribution models. *Ecological modelling* **199**:132-141.

EEA (2000). CORINE land cover technical guide addendum 2000. Report no. 40. European Environment Agency, Copenhagen.

Elith J., Leathwick J. R., and Hastie T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology* online early doi:10.1111/j.1365-2656.2008.01390.x

Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J., 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., Soberón, J., Williams, S., Wisz, M.S. & Zimmermann, N.E. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29: 129–151.

Elmberg, J., Nummi, P., Poysa, H. & Sjoberg, K. (1994). Relationships between species number, lake size and resource diversity in assemblages of breeding waterfowl. *Journal of Biogeography* **21**: 75-84.

Elton, C. (1927). Animal Ecology. Sidgwick and Jackson, London.

Falcucci, A., Maiorano, L., Boitani, L. (2007). Changes in land-use/land-cover patterns in Italy and their implications for biodiversity conservation. *Landscape Ecology* **22**: 617-631.

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, A.H. (1999). Machine Learning Methods for Ecological Applications. Kluwer Academic Publishers, Norwell, MA.

Fornasari L. e de Carli E. (2002). A new project on breeding bird monitoring in Italy. *Bird Census News* **15**: 42-54.

Fox, A.D., Jones, T.A., Singleton, R. & Agnew, A.D.Q. (1994). Food-Supply And The Effects Of Recreational Disturbance On The Abundance And Distribution Of Wintering Pochard On A Gravel-Pit Complex in Southern Britain. *Hydrobiologia,* **280**: 253-61.

Franco, A.M.A., Hill, J.K., Kitschke, C., Collingham, Y.C., Roy, D.B., Fox, R., Huntley, B., Thomas, C.D. (2006). Impacts of climate warming and habitat loss on extinctions at species' low-latitude range boundaries. *Global Change Biology* **12**: 1545-1553.

Friedman, J.H. (2001). Greedy function approximation: the gradient boosting machine. *Annals of Statistics* **29**:1189–1232.

Friedman J.H., Meulman J.J. (2003). Multiple additive regression trees with application in epidemiology. *Stat Med* **22**:1365–1381.

Friedman, J.H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, **38**:367-378.

Fuller, R.M., Cox, R., Clarke, R.T., Rothery, P., Hill, R.A., Smith, G.M., Thomson, A.G., Brown, N.J., Howard, D.C., Stott, A.P. (2005). The UK Land Cover Map 2000: planning, construction and calibration of a user-oriented map of Broad Habitats from remotely sensed satellite images. *International Journal of Applied Earth Observation and Geoinformation* **7**: 202–216.

Garzon, M.B., Blazek, R., Neteler, M., de Dios, R.S., Ollero, H.S. and Furlanello, C. (2006). Predicting habitat suitability with machine learning models: The potential area of Pinus sylvestris L. in the Iberian Peninsula. *Ecological Modelling* **197**:383-393.

Gibbs, J.P. (2000). Wetland loss and biodiversity conservation. *Conservation Biology*, **14**: 314-17.

Govindasamy B., Duffy P. B. & Coquard J. (2003). High-resolution simulations of global climate, part 2: effects of increased greenhouse cases. *Climate Dynamics* **21**: 391-404.

Graham, C.H., Elith, J., Hijmans, R.J., Guisan, A., Peterson, A.T., Loiselle, B.A. and Nceas Predect Species Working, G. (2008). The influence of spatial errors in species occurrence data used in distribution models. *Journal of Applied Ecology* **45**:239-247.

Green, A.J. (1996). Analyses of globally threatened Anatidae in relation to threats, distribution, migration patterns and habitat use. *Conservation Biology* **10**: 1435–1445.

Grinnell, J. (1917). The niche relationship of the Californian Thrasher. *Auk* **34**: 427-433.

Guisan A, Graham CH, Elith J, Huettmann F. (2007a). Sensitivity of predictive species distribution models to change in grain size. *Diversity And Distributions* **13** : 332-340.

Guisan A. & Hofer U. (2003). Predicting reptile distributions at the mesoscale: relation to climate and topography. *Journal of Biogeography* **30**: 1233-1243.

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

Guisan, A., Graham, C.H., Elith, J., Huettmann, F. and Distri, N.S., (2007b). Sensitivity of predictive species distribution models to change in grain size. *Diversity and Distributions* **13**:332-340.

Guisan, A., Lehmann, A., Ferrier, S., Austin, M., Overton, J. M. C., Aspinall, R. & Hastie, T. (2006). Making better biogeographical predictions of species' distributions. *Journal of Applied Ecology*, 43: 386-392.

Guisan, A., Zimmermann, N.E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling* **135**: 147-186.

Guisan, A., Zimmermann, N.E., Elith, J., Graham, C.H., Phillips, S. and Peterson, A.T. (2007b). What matters for predicting the occurrences of trees: Techniques, data, or species' characteristics? *Ecological Monographs*, **77**:615-630.

Hannah L., Midgley G. F. & Millar D. (2002). Climate change-integrated conservation strategies. *Global Ecology and Biogeography* **11**: 485-495.

Harrison P. A., Berry P. M., Butt N. & New M. (2006). Modelling climate change impacts on species' distributions at the European scale: implications for conservation policy. *Environmental Science & Policy* **9**: 116-128.

Hastie, T., Tibshirani, R. and Friedman, J.H. (2001). The elements of statistical learning: data mining, inference, and prediction. Springer-Verlag, New York.

Hastie, T.J. and Tibshirani, R.J. (1990) Generalized Additive Models, Chapman & Hall, New York, 335 pp.

Heikkinen R. K., Luoto M., Araujo M. B., Virkkala R., Thuiller W. & Sykes M. T. (2006). Methods and uncertainties in bioclimatic envelope modelling under climate change. *Progress in Physical Geography* **30**: 751-777.

Hernandez, P.A., Graham, C.H., Master, L.L. and 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. & Graham C. H. (2006). The ability of climate envelope models to predict the effect of climate change on species distributions. *Global Change Biology* **12**: 2272-2281.

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.

Hilbert, D.W., Bradford, M., Parker, T. and Westcott, D.A. (2004). Golden bowerbird (Prionodura newtonia) habitat in past, present and future climates: predicted extinction of a vertebrate in tropical highlands due to global warming. *Biological Conservation*, **116**:367-377.

Hill, J.K., Thomas, C.D., Fox, R., Telfer, M.G., Willis, S.G., Asher, J., Huntley, B. (2002). Responses of butterflies to twentieth century climate warming: implications for future ranges. *Proceedings of the Royal Society of London Series B-Biological Sciences* **269**: 2163-2171.

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.

Holm TE, Clausen P (2006). Effects of water level management on autumn staging waterbird and macrophyte diversity in three Danish coastal lagoons *Biodiversity and Conservation* **15** : 4399-4423.

Hopkins W. A. (2007). Amphibians as models for studying environmental change. *Ilar Journal* **48**: 270-277.

Hortal, J., Garcia-Pereira, P., Garcia-Barros, E. (2004). Butterfly species richness in mainland Portugal: predictive models of geographic distribution patterns. *Ecography* **27**: 68-82.

Hutchinson, G.E. (1957). Concluding remarks. Cold Spring Harbor Symposium on Quantitative Biology, 22: 415-427.

Hutchinson, M.F. (1991). The application of thin plate smoothing splines to continentwide data assimilation. In: Data Assimilation Systems BMRC Research Report No.27 104–113.. Jasper J.D. (Ed.), Bureau of Meteorology, Melbourne.

Ihaka, R., Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal Computational and Graphical Statistics* **5**:299-314.

Johnson P. T. J., Chase J. M., Dosch K. L., Hartson R. B., Gross J. A., Larson D. J., Sutherland D. R. & Carpenter S. R. (2007). Aquatic eutrophication promotes pathogenic infection in amphibians. *Proceedings of the National Academy of Sciences of the United States of America* **104**: 15781-15786.

Kati, V., Devillers, P., Dufrene, M., Legakis, A., Vokou, D., Lebrun, P. (2004). Hotspots, complementarity or representativeness? designing optimal small-scale reserves for biodiversity conservation. *Biological Conservation* **120**: 471-480.

Kelley, C., Garson, J., Aggarwal, A., Sarkar, S.(2002). Place prioritization for biodiversity reserve network design: a comparison of the SITES and ResNet software packages for coverage and efficiency. *Diversity and Distributions* 8: 297-306.

Kershaw, M. & Cranswick, P.A. (2003). Numbers of wintering waterbirds in Great Britain, 1994/1995-1998/1999: I. Wildfowl and selected waterbirds. *Biological Conservation* **111**: 91-104.

Kirby, J.S. (1995). Winter Population Estimates For Selected Waterfowl Species In Britain. *Biological Conservation*, 73: 189-98.

Kirby, J.S., Salmon, D.G., Atkinsonwilles, G.L. & Cranswick, P.A. (1995). Index numbers for waterbird populations .3. long-term trends in the abundance of wintering wildfowl in Great-britain, 1966/67-1991/92. *Journal of Applied Ecology*, **32**: 536-51.

Klar, N., Fernandez, N., Kramer-Schadt, S., Herrmann, M., Trinzen, M., Buttner, I. & Niemitz, C. (2008). Habitat selection models for European wildcat conservation. *Biological Conservation* **141**: 308-319.

Lawrence, R., Bunn, A., Powell, S., Zambon, M. (2004). Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. *Remote Sensing of Environment* **90**: 331-336.

Leathwick JR, Elith J, Francis M.P., Hastie T., Taylor P. (2006). Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. *Marine Ecology-Progress Series* **321**: 267-281

Lehmann, A., Leathwick, J.R. and Overton, J.M., (2002). Assessing New Zealand fern diversity from spatial predictions of species assemblages. *Biodiversity and Conservation*, **11**:2217-2238.

Levinsky I., Skov F., Svenning J. C. & Rahbek C. (2007). Potential impacts of climate change on the distributions and diversity patterns of European mammals. *Biodiversity and Conservation* **16**: 3803-3816.

Liaw A, Wiener M. (2002). Classification and regression by Random Forests. *R News*, **2/3**:18-22..

Lobo, J.M., Jimenez-Valverde, A. and Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**:145-151.

Long PR, Szekely T, Kershaw M, O'Connell, M. (2007). Ecological factors and human threats both drive wildfowl population declines *Animal Conservation* **10**: 183-191.

Luoto, M., Heikkinen, R. K., Pöyry, J. & Saarinen, K. (2006). Determinants of the biogeographical distribution of butterflies in boreal regions. *Journal of Biogeography*, **33**: 1764-1778.

Luoto, M., Poyry, J., Heikkinen, R.K. and Saarinen, K., (2005). Uncertainty of bioclimate envelope models based on the geographical distribution of species. *Global Ecology and Biogeography*, **14**:575-584.

Maes D., Van Dyck H. (2001). Butterfly diversity loss in Flanders (north Belgium): Europe's worst case scenario? *Biological Conservation* **99**: 263-276.

Maiorano, L., Falcucci, A., Boitani, L. (2006). Gap analysis of terrestrial vertebrates in Italy: Priorities for conservation planning in a human dominated landscape. *Biological Conservation* **133**: 455-473.

Mallory, M.L., Venier, L.A. & McKenney, D. (2003). Winter weather and waterfowl surveys in north-western Ontario, Canada. *Journal of Biogeography* **30**: 441-48.

Manel S., Williams H. C. & Ormerod S. J. (2001). Evaluating presence-absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology* **38**: 921-931.

Margules, C.R., Nicholls, A.O., Pressey, R.L. (1988). Selecting Networks of reserves to maximize biological diversity. *Biological Conservation* **43**: 63-76.

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

Martinez, I., Carreno, F., Escudero, A. & Rubio, A. (2006). Are threatened lichen species well-protected in Spain? Effectiveness of a protected areas network. *Biological Conservation* **133**: 500-511.

McCullagh P. and Nelder J.A. (1989). Generalised Linear Models, Chapman & Hall, London, 532 pp.

McKinney, R.A., McWilliams, S.R. & Charpentier, M.A. (2006). Waterfowl-habitat associations during winter in an urban North Atlantic estuary. *Biological Conservation* **132**: 239-49.

McPherson, J.M. and Jetz, W. (2007). Effects of species' ecology on the accuracy of distribution models. *Ecography* **30**:135-151.

McPherson, J.M., Jetz, W. and Rogers, D.J. (2004). The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology* **41**:811-823.

Menendez, R., Gonzalez-Megias, A., Collingham, Y., Fox, R., Roy, D.B., Ohlemuller, R., Thomas, C.D. (2007). Direct and indirect effects of climate and habitat factors on butterfly diversity. *Ecology* **88**: 605-611.

Metzger, K. L., Sinclair, A. R. E., Campbell, K. L. I., Hilborn, R., Hopcraft, J. G. C., Mduma, S. A. R. & Reich, R. M. (2007). Using historical data to establish baselines for conservation: The black rhinoceros (Diceros bicornis) of the Serengeti as a case study. *Biological Conservation* **139**: 358-374.

Milne, D.J., Fisher A. & Pavey C.R. (2006). Models of the habitat associations and distributions of insectivorous bats of the Top End of the Northern Territory, Australia. *Biological Conservation* **130**: 370-385.

Moilanen A., Leathwick J., Elith J. (2008). A method for spatial freshwater conservation prioritization. *Freshwater Biology* **53**: 577-592 .

Moilanen, A. (2007). Landscape Zonation, benefit functions and target-based planning: Unifying reserve selection strategies. *Biological Conservation* **134**: 571-579.

Moilanen, A., Franco, A.M.A., Early, R.I., Fox, R., Wintle, B., Thomas, C.D. (2005). Prioritizing multiple-use landscapes for conservation: methods for large multi-species planning problems. *Proceedings of the Royal Society B-Biological Sciences* **272**: 1885-1891.

Moisen, G. G., Freeman, E. A., Blackard, J. A., Frescino, T. S., Zimmermann, N. E. & Edwards, T. C. (2006). Predicting tree species presence and basal area in Utah: A comparison of stochastic gradient boosting, generalized additive models, and tree-based methods. *Ecological Modelling* **199**: 176-187.

Morrison, M.L., Marcot, B.G. and Mannan, R.W. (1998). Wildlife-habitatrelationships: Concepts and applications. Madison, University of Wisconsin Press, 212 pp.

Myles, A.J., Feudale, R.N., Liu, Y., Woody, N.A., Brown, S.D. (2004). An introduction to decision tree modeling. *Journal of Chemometrics* **18**, 275-285.

Newton I (2007). Weather-related mass-mortality events in migrants *Ibis* **149**: 453-467

Nystrom P., Hansson J., Mansson J., Sundstedt M., Reslow C. & Brostrom A. (2007). A documented amphibian decline over 40 years: Possible causes and implications for species recovery. *Biological Conservation* **138**: 399-411.

Ortega-Huerta, M.A., Peterson, A.T. (2004). Modelling spatial patterns of biodiversity for conservation prioritization in North-eastern Mexico. *Diversity and Distributions* **10**: 39-54.

Paracuellos, M. & Telleria, J.L. (2004). Factors affecting the distribution of a waterbird community: The role of habitat configuration and bird abundance. *Waterbirds* **27**: 446-53.

Parra-Olea G., Martinez-Meyer E. & de Leon G. F. P. (2005). Forecasting climate change effects on salamander distribution in the highlands of central Mexico. *Biotropica* 37: 202-208.

Pawar, S., Koo, M.S., Kelley, C., Ahmed, M.F., Chaudhuri, S., Sarkay, S. (2007). Conservation assessment and prioritization of areas in Northeast India: Priorities for amphibians and reptiles. *Biological Conservation* **136**: 346-361.

Pearson R. G., Thuiller W., Araujo M. B., Martinez-Meyer E., Brotons L., McClean C., Miles L., Segurado P., Dawson T. P. & Lees D. C. (2006). Model-based uncertainty in species range prediction. *Journal of Biogeography* **33**: 1704-1711.

Pearson, R.G., Dawson, T.E. & Liu, C. (2004). Modelling species distributions in Britain: a hierarchical integration of climate and land-cover data. *Ecography*, **27**: 285–298.

Peters A, Hothorn T, Lausen B. (2002). ipred: Improved predictors. *R News* **2**:22–6.

Peterson A. T. (2003). Projected climate change effects on Rocky Mountain and Great Plains birds: generalities of biodiversity consequences. *Global Change Biology* **9**: 647-655.

Peterson A. T., Ortega-Huerta M. A., Bartley J., Sanchez-Cordero V., Soberon J., Buddemeier R. H. & Stockwell D. R. B. (2002). Future projections for Mexican faunas under global climate change scenarios. *Nature* **416**: 626-629.

Piha H., Luoto M., Piha M. & Merila J. (2007). Anuran abundance and persistence in agricultural landscapes during a climatic extreme. *Global Change Biology* **13**: 300-311.

Polus, E., Vandewoestijne, S., Choutt, J., Baguette, M. (2007). Tracking the effects of one century of habitat loss and fragmentation on calcareous grassland butterfly communities. *Biodiversity and Conservation* **16**: 3423-3436.

Pounds J. A., Bustamante M. R., Coloma L. A., Consuegra J. A., Fogden M. P. L., Foster P. N., La Marca E., Masters K. L., Merino-Viteri A., Puschendorf R., Ron S. R., Sanchez-Azofeifa G. A., Still C. J. & Young B. E. (2006). Widespread amphibian extinctions from epidemic disease driven by global warming. *Nature* **439**: 161-167.

Poyry, J., Luoto, M., Heikkinen, R.K. and Saarinen, K. (2008). Species traits are associated with the quality of bioclimatic models. *Global Ecology and Biogeography*, **17**:403-414.

Prasad, A.M., Iverson, L.R., Liaw, A. (2006). Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems* **9:** 181-199.

Pulliam, H. R. (2000). On the relationship between niche and distribution. *Ecology Letters* 3: 349-361.

Radivojac, P., Chawla, N.V., Dunker, A.K., Obradovic, Z. (2004). Classification and knowledge discovery in protein databases. *Journal of Biomedical Informatics* **37**: 224-239.

Raimo, V., Risto K. H., Niko L. and Miska L. (2008). Projected large-scale range reductions of northern-boreal land bird species due to climate change. *Biological Conservation* **141**: 1343-1353.

Ramos, M.A., Lobo, J.M., Esteban, M. (2001). Ten years inventorying the Iberian fauna: results and perspectives. *Biodiversity and Conservation* **10**: 19-28.

Ravenscroft, N.O.M. & Beardall, C.H. (2003). The importance of freshwater flows over estuarine mudflats for wintering waders and wildfowl. *Biological Conservation* **113**: 89-97.

Rayner, M. J., Clout, M. N., Stamp, R. K., Imber, M. J., Brunton, D. H. & Hauber, M. E. (2007). Predictive habitat modelling for the population census of a burrowing seabird: A study of the endangered Cook's petrel. *Biological Conservation* 138: 235-247.

Ridgill, S.C. & Fox, A.D. (1990). Cold weather movements of Waterfowl in Western Europe. IWEB Spec. Publ. 13, Slimbridge,UK, 87pp.

Rizzoli, A., Merler, S., Furlanello, C., Gench, C. (2002). Geographical information systems and bootstrap aggregation (Bagging) of tree-based classifiers for Lyme disease risk prediction in Trentino, Italian Alps. *Journal of Medical Entomology* **39**: 485-492.

Romero-Calcerrada, R., Luque, S. (2006). Habitat quality assessment using Weights-of-Evidence based GIS modelling: The case of Picoides tridactylus as species indicator of the biodiversity value of the Finnish forest. *Ecological Modelling* **196**: 62-76.

Rondinini, C., Stuart, S., Boitani, L. (2005). Habitat suitability models and the shortfall in conservation planning for African vertebrates. *Conservation Biology* **19**: 1488-1497.

Roy, D.B., Rothery, P., Moss, D., Pollard, E., Thomas, J.A. (2001). Butterfly numbers and weather: predicting historical trends in abundance and the future effects of climate change. *Journal of Animal Ecology* **70**: 201-217.

Ruffo, S. & Stoch F. (2005). Checklist and distribution of the Italian fauna. Memorie del Museo Civico di Storia Naturale del Museo di Verona. Italian Ministry of Environment Verona.

Rushton, S.P., Ormerod, S.J. & Kerby, G. (2004). New paradigms for modelling species distributions? *Journal of Applied Ecology* **41**: 193–200.

Santos, X., Brito, J. C., Sillero, N., Pleguezuelos, J. M., Llorente, G. A., Fahd, S. & Parellada, X. (2006). Inferring habitat-suitability areas with ecological modelling techniques and GIS: A contribution to assess the conservation status of Vipera latastei. *Biological Conservation* **130**: 416-425.

Santoul, F., Figuerola, J. & Green, A.J. (2004). Importance of gravel pits for the conservation of waterbirds in the Garonne river floodplain (southwest France). *Biodiversity and Conservation* **13**: 1231-43.

Scott, J.M., Heglund, P.J.,Morrison, M.L., Haufler, J.B., Raphael, M.G., Wall, W.A. and Samson, F.B. (2002). Predicting species occurrences: issues of accuracy and scale Island Press, Colelo, Washington, 868 pp.

Segurado, P. and Araujo, M.B., (2004). An evaluation of methods for modelling species distributions. *Journal of Biogeography* **31**:1555-1568.

Seiler, A. (2005). Predicting locations of moose-vehicle collisions in Sweden. *Journal of Applied Ecology* **42**: 371-382.

Seoane, J., Carrascal, L.M., Alonso, C.L. and Palomino, D. (2005). Species-specific traits associated to prediction errors in bird habitat suitability modelling. *Ecological Modelling* **185**:299-308.

Shah, S., Kusiak, A. (2007). Cancer gene search with data-mining and genetic algorithms. *Computers in Biology and Medicine* **37**: 251-261.

Sindaco, R., Doria, G., Mazzetti, E. & Bernini, F. (2006). Atlante degli Anfibi e dei Rettili d'Italia/Atlas of Italian Amphibians and Reptiles. Societas Herpetologica Italica, Edizioni Polistampa, Firenze.

Smith M. A. & Green D. M. (2005). Dispersal and the metapopulation paradigm in amphibian ecology and conservation: are all amphibian populations metapopulations? *Ecography* **28**: 110-128.

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

Stuart S. N., Chanson J. S., Cox N. A., Young B. E., Rodrigues A. S. L., Fischman D. L. & Waller R. W. (2004). Status and trends of amphibian declines and extinctions worldwide. *Science* **306**: 1783-1786.

Suarez-Seoane, S., Osborne, P.E. & Alonso, J.C., (2002). Large-scale habitat selection by agricultural steppe birds in Spain: identifying species habitat responses using generalized additive models. *Journal of Applied Ecology* **39**: 755-771.

Suter W. (1994). Overwintering waterfowl on Swiss lakes - how are abundance and species richness influenced by trophic status and lake morphology *Hydrobiologia* **280**: 1-14.

Swets J. A. (1988). Measuring the accuracy of diagnostic systems. *Science* **240**: 1285-1293.

Teixeira J. & Arntzen J. W. (2002). Potential impact of climate warming on the distribution of the Golden-striped salamander, Chioglossa lusitanica, on the Iberian Peninsula. *Biodiversity and Conservation* **11**: 2167-2176.

Thomaes, A., Kervyn, T. & Maes, D. (2008). Applying species distribution modelling for the conservation of the threatened saproxylic Stag Beetle (Lucanus cervus). *Biological Conservation* **141**: 1400-1410.

Thomas J.A., Telfer M.G., Roy D.B., Preston C.D., Greenwood J.J.D., Asher J., Fox R, Clarke R.T., Lawton J.H. (2004). *Science* **303**: 1879-1881.

Thomas, C.D., Franco, A.M.A., Hill, J.K. (2006). Range retractions and extinction in the face of climate warming. *Trends in Ecology and Evolution* **21**: 415-416.

Thomas, J.A. (1995). The conservation of declining butterfly populations in Britain and Europe: Priorities, problems and successes. *Biological Journal of the Linnean Society* **56**: 55-72.

Thomas, J.A. (2005). Monitoring change in the abundance and distribution of insects using butterflies and other indicator groups. *Philosophical Transactions of the Royal Society B-Biological Sciences* **360**: 339-357.

Thomas, J.A., Clarke, R.T. (2004). Extinction rates and butterflies – Response. *Science* **305**: 1563-1564.

Thuiller W. (2003). BIOMOD - optimizing predictions of species distributions and projecting potential future shifts under global change. *Global Change Biology* **9**: 1353-1362.

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

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.

Thuiller W., Araujo M.B., and Lavorel S. (2003). Generalized models vs. classification tree analysis: predicting spatial distributions of plant species at different scales. *Journal of Vegetation Science* **14**: 669–680.

Thuiller W., Midgley G.F., Rouget M., Cowling R.M. (2006). Predicting patterns of plant species richness in megadiverse *South Africa Ecography* **29**: 733-744

Thuiller, W., Lavorel, S., Araujo, M. B., Sykes, M. T. & Prentice, I. C. (2005) Climate change threats to plant diversity in Europe. *Proceedings of the National Academy of Sciences of the United States of America 102*: 8245-8250.

Tontini L., Castellano S., Bonelli S., Balletto E. (2003). Patterns of butterfly diversity and community ecology above the timber-line in the Italian Alps and Apennines. In: Alpine Biodiversity in Europe 297-306. G. Grabherr, C. Körner, L. Nagy and D. B. A. Thompson (Eds). Berlin and Heidelberg: Springer Verlag.

Tsoar, A., Allouche, O., Steinitz, O., Rotem, D. and Kadmon, R. (2007). A comparative evaluation of presence-only methods for modelling species distribution. *Diversity and Distributions* **13**:397-405.

Tuck, G. N., Polacheck, T., Croxall, J. P. & Weimerskirch, H. (2001). Modelling the impact of fishery by-catches on albatross populations. *Journal of Applied Ecology* **38**: 1182-1196.

Tuite, C.H., Hanson, P.R. & Owen, M. (1984). Some Ecological Factors Affecting Winter Wildfowl Distribution On Inland Waters In England And Wales, and the influence of water-based recreation. *Journal of Applied Ecology*, **21**: 41-61.

van Swaay, C., Warren, M., Lois, G. (2006a). Biotope use and trends of European butterflies. *Journal of Insect Conservation* **10**: 189-209.

van Swaay, C., Warren, M.S. (1999). Red Data book of European butterflies (Rhopalocera). Nature and Environment, No. 99, Council of Europe Publishing, Strasbourg.

van Swaay, C., Warren, M.S. (2006b). Prime Butterfly Areas of Europe: an initial selection of priority sites for conservation. *Journal of Insect Conservation* **10**: 5-11.

Virkkala, R., Luoto, M., Heikkinen, R.K. & Leikola N. (2005). Distribution patterns of boreal marshland birds: modelling the relationships to land cover and climate. *Journal of Biogeography* **32**: 1957-1970.

Wake D. B. (2007). Climate change implicated in amphibian and lizard declines. Proceedings of the National Academy of Sciences of the United States of America **104**: 8201-8202.

Wenzel, M., Schmitt, T., Weitzel, M., Seitz, A. (2006). The severe decline of butterflies on western German calcareous grasslands during the last 30 years: A conservation problem. *Biological Conservation* **128**: 542-552.

Whitfield S. M., Bell K. E., Philippi T., Sasa M., Bolanos F., Chaves G., Savage J. M. & Donnelly M. A. (2007). Amphibian and reptile declines over 35 years at La Selva, Costa Rica. Proceedings of the National Academy of Sciences of the United States of America **104**: 8352-8356.